# American Society of Human Genetics 64th Annual Meeting
# October 18–22, 2014 San Diego, CA

# PLATFORM ABSTRACTS

## 1

**The UK10K project: rare variants in health and disease.** *N. Soranzo, The UK10K Consortium.* Wellcome Trust Sanger Institute, Cambridge, United Kingdom.

Understanding of how genetic variation contributes to human traits and which genes are involved is still largely incomplete. Rare and low frequency genetic variants (defined as minor allele frequency [MAF] <1% and 1-5%, respectively) are thought to play a role in common and rare disease, but until recently it has not been possible to assess their contribution systematically. The UK10K project (www.uk10k.org) studies the contribution rare and low frequency variation to a wide spectrum of biomedically relevant quantitative traits and diseases with different predicted genetic architecture. Here we describe the data generated by the different arms of the UK10K project, and use it to address empirically important open questions. We show that the 24 million novel genetic variants identified provide an extensive reference genetic sample for the UK, with fewer than 10% rare variants shared with other large scale resources. The resulting haplotype panel boosts accuracy and coverage of imputation of rare variants in GWAS studies. We discuss the value of these data for discovery and interpretation of variants of clinical importance, including an evaluation of incidental findings that suggests that greater than 5% participants carry rare and potentially actionable genetic variants. We describe how demographic history influences the observed weak structuring of rare variation in the UK (MAF= 0.1-0.3%), and evaluate empirically the potential for confounding due to stratification in association studies of complex traits. We describe the contribution of common, low frequency and rare variants to 61 biomedically important quantitative traits, describing novel alleles associated with adiponectin (*ADPOQ*), lipids (*PCSK9, LPL, APOC3|APOA4|APOA1, PCSK7, CETP, LIPG, LDLR and APOE)* and several other traits. We further discuss general characteristics of rare variant associations from the comparative evaluation of the allelic architecture of the 61 traits. The data released to the scientific community includes individual-level genotypic and phenotypic data, a reference panel of haplotypes for imputation of rare variants into genome-wide SNP arrays and analysis protocols and summary results for complex trait associations based on single-marker and rare variant aggregation tests. Our results inform future whole-genome and whole-exome sequencing based studies seeking to characterize the role of rare genetic variation in predisposition to rare and common disease.

## 2

**Human-specific gene evolution and structural diversity of the chromosome 16p11.2 autism CNV.** *X. Nuttle[1], G. Giannuzzi[2], M.H. Duyzend[1], P.H. Sudmant[1], O. Penn[1], G. Chiatante[3], M. Malig[1], J. Huddleston[1,4], L. Denman[1], L. Harshman[1], C. Baker[1], A. Raja[1,4], K. Penewit[1], F. Antonacci[3], R. Bernier[5], A. Reymond[2], E.E. Eichler[1,4].* 1) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; 2) Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; 3) Department of Biology. University of Bari, Bari, Italy; 4) Howard Hughes Medical Institute, Seattle, WA, USA; 5) Department of Psychiatry, University of Washington, Seattle, WA, USA.

Recurrent deletions and duplications at 16p11.2 are a major contributor to autism and also associate with schizophrenia and extremes of body mass index and head circumference. These events occur via nonallelic homologous recombination (NAHR) between directly oriented segmental duplications (BP4 and BP5) ~600 kbp apart. Using whole genome sequencing (WGS) data from 2,551 humans, 86 great apes, a Neanderthal, and a Denisovan, we observed extensive copy number variation in BP4 and BP5 in human populations and identified *BOLA2* as a gene duplicated in *Homo sapiens* after our divergence from ancient hominins. Performing massively parallel and PacBio sequencing of large-insert clones from orangutan and chimpanzee, we generated complete sequence over the 16p11.2 locus in these apes and reconstructed its evolutionary history. We find three inversions occurred in the human lineage after divergence from orangutan, affecting > 1 Mbp of sequence, including 45 genes. In concert with these evolutionary inversions, > 950 kbp have been added to the region via segmental duplication. Comparative sequence analyses suggest that *BOLA2* was part of an ~110 kbp segment that duplicated from BP5 to BP4 ~250 kya at the time when *Homo sapiens* emerged as a species. Modern humans carry at least one additional copy of *BOLA2* (ranging from 3 to 14 diploid copies) in contrast to apes, Neanderthal, and Denisova, where the gene exists as two diploid copies. *BOLA2* is thought to be involved in cell proliferation or cell cycle regulation. RT-PCR and RNA-seq suggest *BOLA2* is widely expressed, and expression levels in lymphoblastoid cell lines correlate with copy number ($r = 0.29$). Using the same sequencing strategy as above, we completely sequenced four distinct human structural haplotypes at 16p11.2 (> 5 Mbp of sequence). We discover haplotypes where BP4 and BP5 differ by hundreds of kbp including tandem duplications of *BOLA2*, leading to likely differences in predisposition to NAHR. Leveraging our high quality human haplotype sequences, we are currently assaying *BOLA2* copy number and refining breakpoints in > 125 patients with a 16p11.2 deletion or duplication. Our preliminary data suggest that breakpoints cluster near the *Homo sapiens*-specific duplications involving *BOLA2*. These findings raise the exciting possibility that predisposition to recurrent rearrangements associated with autism is linked to the emergence of novel duplicated genes in the last 250,000 years of our species' evolution.

## 3

**Discovery and functional characterization of recurrent gene fusions from 7,470 primary tumor transcriptomes across 28 human cancers.** *C. Bandlamudi[1], P. Lin[1], J. Tian[2], R. Grossman[1], K. White[1].* 1) University of Chicago, Chicago, IL; 2) Duke University, NC.

Gene fusions are consequences of somatic rearrangements in cancer genomes. Many oncogenic fusions have been discovered in different tumor types that currently serve as diagnostic, prognostic and therapeutic markers. However, an emerging theme from recent sequencing studies is that many tumorigenic fusions appear at frequencies 2% or less within the respective tumor types, suggesting that many functional fusions with clinical significance remain to be discovered using large sample sizes and sensitive detection approaches. We have developed a novel algorithm, Minimum Overlap Junction Optimizer (MOJO), that uses a transcriptome guided approach to detect fusions. Using 20 tumor transcriptomes with experimentally validated fusions, we show that MOJO demonstrates the highest sensitivity and specificity compared to nine other published methods. We performed fusion discovery using MOJO on 7,470 transcriptomes in the Cancer Genome Atlas (TCGA). Using 1,800 normal tissue transcriptomes from Genotype Tissue Expression (GTEx) consortium, we developed filters to model and account for technical and biological noise in fusion discovery through RNAseq. We demonstrate our sensitivity by recovering all fusions that have so far been validated in these samples. We nominated 16,114 high confidence fusion calls including 430 known fusions and 1,039 events involving known cancer genes in COSMIC's Cancer Gene Census. We find that the frequency spectrum of fusion events ranges from a median of 8.3 events/tumor in Ovarian to 0.3 events/tumor in Thyroid. Our integrated analysis of copy number and fusion events in a subset of tumor types suggest that the rate of fusion events is correlated with the overall degree of genomic instability. Our analysis identified 201 fusion genes found in 5 or more samples across multiple tumor types. Using additional filtering criteria, we selected 29 in-frame fusion genes for gain-of-function validations. We generated stable MCF10A cell lines expressing these in-frame fusions, and so far, we have assayed 18 of them and found that 50% showed a statistically significant increase (p-value < 0.01) in proliferation, as compared to the GFP-only control. Intriguingly, we find that a majority of the validated fusions are identified in three or more tumor types, albeit, with low frequencies.

## 4

**Phase III Trial of Afamelanotide 16 mg Subcutaneous Bioresorbable Implants for the treatment of Erythropoietic Protoporphyria.** *R.J Desnick[1], K.E Anderson[2], D.M Bissell[3], J.R Bloomer[4], H.L Bonkovsky[5], M. Lebwohl[6], H. Lim[7], C. Parker[8], J. Phillips[8], H. Naik[1], M. Balwani[1].* 1) Dept Gen/Genomic Sci, Box 1498, Mount Sinai Sch Med, New York, NY; 2) Department of Preventive Medicine and Community Health, University of Texas Medical Branch, Galveston, TX; 3) Department of Medicine, University of California, San Francisco, CA; 4) Department of Medicine, University of Alabama at Birmingham, AL; 5) Department of Medicine, Carolinas Medical Center and Healthcare system, Charlotte, NC; 6) Department of Dermatology, Mount Sinai School of Medicine; 7) Department of Dermatology, Henry Ford Health System, Detroit, MI; 8) Department of Internal Medicine, University of Utah, Salt Lake City, UT.

Erythropoietic Protoporphyria (EPP) is a rare, autosomal recessive photodermatosis resulting from deficient activity of the heme biosynthetic enzyme, ferrochelatase, and the resultant erythroid accumulation of photoactive protoporphyrin IX. When EPP patients are exposed to sunlight, a phototoxic reaction is triggered resulting in incapacitating pain. There is no effective treatment and patients avoid sun exposure which significantly impacts their daily life activities and overall quality-of-life (QoL). Afamelanotide, in the form of a subcutaneously administered, bioresorbable implant is a potent analogue of alpha-melanocyte stimulating hormone ($\alpha$-MSH) and stimulates the production of eumelanin in the skin epidermis. Melanin, in the form of eumelanin, is a photoprotective agent and the postulated mechanism of afamelanotide photoprotection includes the absorption and scattering of UV light, free radical scavenging and quenching of UV light. To determine the safety and effectiveness of afamelanotide in increasing sun-exposure time and QoL, 93 North American patients were enrolled in a FDA-approved, multicenter, randomized, double-blind, placebo-controlled Phase III study. Afamelanotide or placebo implants were administered subcutaneously on days 0, 60 and 120, and the type and duration of sun exposure, number and severity of phototoxic reactions, and adverse events were recorded. A subset of patients underwent photoprovocation testing at baseline and during the study to determine tolerance to visible light exposure. The impact of treatment on QoL was evaluated using two validated questionnaires, the Dermatology Life Quality Index and an EPP-specific quality of life questionnaire (EPP-QoL). The primary endpoint analysis showed that treated patients were able to experience more pain free sun exposure, particularly between 10:00 to 18:00 hrs (69.4 v 40.8 hours over the entire study, p= 0.044). Photoprovocation testing showed increased tolerance to visible light on the dorsum of the hand and lower back (day 90 p=0.011, p<0.001, day 120 p=0.045, p=0.028) and the EPP-QoL showed significant treatment related improvement (p=<0.001, p=0.002, p=0.028 at day 60, 90 and 180). Adverse events were mainly mild and unrelated to the study drug. These results demonstrate that afamelanotide is safe, well tolerated, and improves pain-free sun exposure and QoL in patients with EPP.

## 5

**Re-engineering meiotic recombination in the mouse.** *E. Hatton[1], B. Davies*[*1], J. Hussin[1], F. Pratto[2], D. Biggs[1], N. Altemose[3], N. Hortin[1], C. Preece[1], D. Moralli[1], A. Gupta-Hinch[1], K. Brick[2], C. Green[1], D. Camerini-Otero[2], S. Myers*[*1,3], *P. Donnelly*[*1,3].* 1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 2) National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, USA; 3) Department of Statistics, University of Oxford, Oxford, United Kingdom.

PRDM9 is a zinc-finger DNA binding protein implicated in controlling the localisation of meiotic recombination events and displays a large diversity in its zinc-finger binding array within and between species. Here, we successfully replaced the zinc-finger array of Prdm9 in a C57BL/6 mouse strain by its human B allele counterpart, providing a precise assay to investigate PRDM9 binding properties and their consequences. We developed a new statistical approach for calling double-strand break (DSB) hotspots from ChipSeq data, which utilises asymmetries on the two DNA strands expected around real hotspots. Our results confirm that virtually all DSB hotspots are under PRDM9 control, and indicate that a change in its DNA recognition domain is sufficient to fully reset the hotspot landscape. Sequence motifs enriched within DSB hotspots in the humanized mouse closely match the consensus motif reported in humans, and more than 78% of the DSB hotspots contain such a motif. In the heterozygote mouse, the human allele dominates the mouse allele in specifying DSB hotspots, which suggests substantial differences in binding abilities between alleles and/or in efficiencies to access their target motifs: 69% of the DSB hotspots in the heterozygote are found in the humanized homozygote mouse, whereas only 26% are inherited from the wild type homozygote. We also compared the localisation of DSB hotspots with several genomic and epigenetic elements. In particular, we found that exons are enriched in DSB hotpots in both humanized and wild type mice. Overall, our results suggest that Prdm9 zinc-finger array plays a predominant role in the localisation at the fine genomic scale, but that broad scale properties of DSB hotspot maps are mainly independent of the zinc-finger array. Finally, we show that the replacement of the zinc-finger array of the C57BL/6 Prdm9 allele was sufficient to rescue fertility in a well-characterised M. m. musculus and M. m. domesticus cross, in which hybrid males are otherwise fully sterile.

## 6

**Examining Variation in Recombination Levels in the Human Female: A Test of the Production Line Hypothesis.** *R. Rowsey[1], J. Gruhn[1], K. Broman[2], P. Hunt[1], T. Hassold[1].* 1) School of Molecular Biosciences and Center for Reproductive Biology, Washington State University, Pullman, WA; 2) Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI.

The most important risk factor for human aneuploidy is increasing maternal age, but the basis of this association remains unknown. Indeed, one of the earliest models of the maternal age effect - the “Production Line Model” proposed by Henderson and Edwards in 1968 - remains one of the most-cited explanations. The model has two key components: that the first oocytes to enter meiosis are the first ovulated, and that the first to enter meiosis have more recombination events (crossovers) than those that enter meiosis later in fetal life. Studies in rodents demonstrate that the first oocytes to enter meiosis are, indeed, the first to be ovulated, but the association between timing of meiotic entry and recombination levels has not been tested. We recently initiated molecular cytogenetic studies of second trimester human fetal ovaries, allowing us to directly examine the number and distribution of crossover-associated proteins in prophase stage oocytes. Our observations on over 8,000 oocytes from 191 ovarian samples demonstrate extraordinary variation in recombination within and among individuals, but provide no evidence of a difference in recombination levels between oocytes entering meiosis early or late in fetal life. Thus, our data provide the first direct test of the second tenet of the Production Line model and suggest that it does not provide a plausible explanation for the human maternal age effect, meaning that - 45 years after its introduction - we can finally conclude that the Production Line Model is not the basis for the maternal age effect on trisomy.

## 7

**The fine-scale landscape of meiotic gene conversion.** *A.L. Williams[1], J. Blangero[2], M. Przeworski[1].* 1) Biological Sciences Department, Columbia University, New York, NY; 2) Texas Biomedical Research Institute, San Antonio, TX.

Meiotic recombination is essential to the proper alignment and segregation of chromosomes, and produces a haploid genome that is a mosaic of the two parental chromosomes. Among possible resolutions of recombination are crossovers, in which homologous chromosomes reciprocally exchange material, and non-crossover “gene conversion” events, in which short segments are copied from one homolog to the other across 50-1,000 bp. Despite dramatic progress in our understanding of crossover resolutions of recombination in mammals over the past decade, little is known about gene conversion events.

We present an analysis of meiotic gene conversion patterns identified using whole genome sequence data from 11 three-generation human pedigrees. We estimate the number of gene conversion events per meiosis and their tract lengths, and compare these to rates of crossing-over while accounting for differences in power. We then examine the location of crossover and non-crossover resolutions and their determinants. Specifically, we focus on an observation that we previously made on the basis of more limited data of complex recombination events in which multiple gene conversion tracts cluster near each other and near crossovers. This unexpected pattern, similar to recombination events reported previously in *S. cerevisiae*, is inconsistent with canonical models of double strand break repair, and if frequent, would be predicted to lead to a complex correlation structure among variants.

To characterize the impact of non-crossover resolutions on genomic base composition, we estimate the strength of GC-biased gene conversion—the over-transmission of G or C alleles at GC/AT heterozygous sites. This form of transmission distortion is hypothesized to have a major on impact base composition over evolutionary time, yet there is little direct evidence of its existence. Lastly, we leverage these pedigree sequence data to address the long-standing question of whether recombination events produce *de novo* mutations by analyzing whether these two classes of events co-localize more than expected from background levels.

## 8

**Recombination maps for Latino populations based on ancestry inference.** *S. Shringarpure[1], D. Wegmann[2], C. Gignoux[1], B. Maples[1], A. Ferrer-Admetlla[2], A. Moreno-Estrada[1], K. Sandoval[1], C. Eng[3], S. Huntsman[3], A. Ko[4,5], T. Tusie-Luna[6,7], C. Aguilar-Salinas[6], P. Pajukanta[4,5], D. Torgerson[3], E. Burchard[3], J. Below[8], B. Pasaniuc[4], S. Gravel[10], J. Novembre[9], C. Bustamante[1].* 1) Genetics Department, Stanford University, Stanford, CA, USA; 2) Department of Biology, University of Fribourg, Fribourg, Switzerland,; 3) Bioengineering, University of California, San Francisco, San Francisco, CA, USA; 4) Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, USA; 5) Molecular Biology Institute, UCLA, Los Angeles, USA; 6) Instituto Nacional de Ciencias Médicas y Nutrición, Salvador Zubiran,Mexico City, Mexico; 7) Instituto de Investigaciones Biomédicas de la UNAM, Mexico City, Mexico; 8) School of Public Health, University of Texas, Houston, TX, USA; 9) Department of Human Genetics, University of Chicago, Chicago, IL, USA; 10) Human Genetics, McGill University, Montreal, Canada.

Accurate estimation of recombination rates is important for studying recombination and its effect on genetic variation. We construct the largest high-resolution recombination map for Latino populations from more than 12,000 individuals of Mexican and Puerto Rican ancestry. Our recombination map inference leverages the recent admixture of African, European and Native American ancestries in the Americas to detect approximately 7.5 million recombination events. At coarse scales, the Latino recombination map correlates well with the European and African recombination maps (r2=0.95 with the HapMap CEU map and 0.85 with the HapMap YRI genetic map at the 1 Mb scale) but shows considerable differentiation at fine scales (r2=0.34 with the CEU map and r2=0.39 with the YRI map at the 10 kb scale). Using estimates of average admixture proportions from the source populations and the European and African recombination maps, we also infer a recombination map specific to Native American populations. In addition, we also construct population-specific recombination maps for Mexicans and Puerto Ricans to study the effect of different Native American ancestral contributions to these populations. Our results provide a useful resource for studying recombination and genetic variation in Latinos and Native Americans.

## 9

**The human X chromosome is the target of megabase wide selective sweeps associated with multi-copy genes expressed in male meiosis and involved in reproductive isolation.** *M.H. Schierup[1], K. Munch[1], K. Nam[1], T. Mailund[1], J.Y. Dutheil[2].* 1) Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark; 2) Max Planck Institute for terrestrial Microbiology, Marburg, Germany.

The X chromosome differs from the autosomes in its hemizogosity in males and in its intimate relationship with the very different Y chromosome. It has a different gene content than autosomes and undergo specific processes such as meiotic sex chromosome inactivation (MSCI) and XY body formation. Previous studies have shown that natural selection is more efficient against deleterious mutations and, in chimpanzee, that positive selection is prevalent. We show that in all great apes species, megabase wide regions of the X chromosome has severely reduced diversity (by more than 80%). These regions are partly shared among species and indicate a large number of strong selective sweeps that have occurred independently on the same set of targets in different great apes species. We use simulations and deterministic calculations to show that background selection or soft selective sweeps are unlikely to be responsible. The regions also bear all the hallmarks of selective sweeps such as an increased proportion of singletons and higher divergence among closely related populations. Human populations are differently affected, suggesting that a large fraction of sweeps are private to specific human populations. The regions of reduced diversity correlates strongly with the position of X-ampliconic regions, which are 100-500 kb regions containing multiple copies of genes that are solely expressed during male meiosis. We propose that the genes in these regions escape MSCI and participate in an intragenomic conflict with regions of similar function on the Y chromosome for transmission of sex chromosomes to the next generation, i.e. sex chromosome meiotic drive. Recent results from Neanderthal introgression into humans point to the same regions as showing no introgression, consistent with the above process leading to reproductive isolation. Strikingly, the same regions of the X also shows much reduced divergence between human and chimpanzee, suggesting either that this speciation process was indeed complex or that the same regions were under strong selection in the human chimpanzee ancestor.

## 10

**New insights on human *de novo* mutation rate and parental age.** *W.S.W. Wong, B. Solomon, D. Bodian, D. Thach, R. Iyer, J. Vockley, J. Niederhuber.* Inova Translational Medicine Institute, Falls Church, VA.

Germline mutations have a major role to play in evolution. Much attention has been given to studying the pattern and rate of human mutations using biochemical or phylogenetic methods based on closely related species. Massively parallel sequencing technologies have given scientists the opportunity to study directly measured *de novo* mutations (*DNM*s) at an unprecedented scale. Here we report the largest study (to our knowledge) of *de novo* point mutations in humans, in which we used whole genome deep sequencing (~60x) data from 605 family trios (father, mother and newborn). These trios represent the first group of approximately 2,700 trios who have undergone whole-genome sequencing (WGS) through our pediatric-based WGS research studies. The fathers ages range from 17 to 63 years and the mothers ages range from 17 to 43 years. We identified over 23000 *DNM*s (~40 per newborn) in the autosomal chromosomes using a customized pipeline and infer that the mutation rate per basepair is around $1.2\times10^{-8}$ per generation, well within the reported range in previous studies. We were also able to confirm that the total number of *DNM*s in the newborn was directly proportional to the paternal age ($P < 2\times10^{-16}$). Maternal age is shown to have a small but significant positive effect on the number of *DNM*s passed onto the offspring, ($P$=0.003) , even after accounting for the paternal age. This contradicts the prior dogma that maternal age only has an effect on chromosomal abnormalities related to nondisjunction events. Furthermore, 5% (22 total) of newborns in the analyzed group were conceived with assisted reproductive technologies (ARTs), and these infants have on average 5 more *DNM*s (Bias corrected and accelerated bootstrap 95% Confidence Interval, 1.24 to 8.00) than those conceived naturally, after controlling for both parents ages. Both parents ages remain significant as independently correlated with DNMs even after the families that used ARTs were removed from the analysis. Our study enhances current knowledge related to the human germline mutational rates.

## 11

**Cholera resistance in Bangladesh: combining signals of ancient, pathogen driven selection with genome wide association to understand immune response.** *E.K. Karlsson[1,2], I. Shylakhter[1,2], F. Qadri[3], J.B. Harris[4,5], S.B. Calderwood[4,6], E.T. Ryan[4,5,6], R.C. LaRocque[4,6], P.C. Sabeti[1,2,7].* 1) Broad Institute, Cambridge, MA; 2) Center for Systems Biology, Harvard Univ, Cambridge, MA; 3) iccdr,b, Dhaka, Bangladesh; 4) Div. of Infectious Diseases, MGH, Boston, MA; 5) Dept of Pediatrics, Harvard Medical School, Boston, MA; 6) Dept of Medicine, Harvard Medical School, Boston, MA; 7) Dept of Immunology and Infectious Disease, Harvard School of Public Health, Boston, MA.

Cholera is an ancient, deadly and common pathogen in the Ganges River Delta, and has likely exerted evolutionary pressure on human populations in the region. Cholera is caused by the bacterium *Vibrio cholerae*, which colonizes the small intestine and causes profound diarrhea, rapid dehydration, and, without medical intervention, mortality rates as high as 50%. In prior research, we combined a genome wide selection scan with a targeted association study and showed that cholera response immune pathways are enriched for selection in the Bengali population of Bangladeshi. We developed a model of the human innate immune response wherein inflammasome activation and the NF-κB signaling pathway play an integrated role in TLR4-mediated sensing of *V. cholerae*.

Here, we report the first genome wide association study of cholera susceptibility in Bangladesh. Our results suggest that positive selection drives protective variants of large effect to high frequency, increasing statistical power. We identified genome-wide significant associations to variants in major immune genes by comparing 94 Bengali patients hospitalized with severe cholera to just 80 unphenotyped population controls (543,832 SNPs after QC), results we replicated in a second cohort of 157 cases and 83 controls. We also completed a much denser scan for positive selection, applying our Composite of Multiple Signals approach to newly released 1000 Genomes Project (1000G) full sequence data for the Bengali population. In total, we have integrated association scores for 4.8 million polymorphic SNPs (after imputation) into the composite selection score and found specific immune signaling pathways targeted by cholera driven selection, including TLR4 mediated response to bacterial lipopolysaccharide (LPS). We are now investigating the top candidate variants using new high-throughput functional methods.

Our approach is broadly applicable to other historically prevalent infectious diseases, such as Lassa fever, tuberculosis, leishmaniasis, dengue fever and malaria, and to common diseases, such as inflammatory bowel disease, for which the associated genes may have been historically selected. Combining ancient history with modern genomics is a powerful approach for investigating genome function.

## 12

**Direct detection of genetic dominance from natural variation in human populations.** *D. Balick[1,2], R. Do[3], D. Reich[3], S. Sunyaev[1,2].* 1) Genetics Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 2) Broad Institute, Cambridge. MA; 3) Department of Genetics, Harvard Medical School, Boston, MA.

Despite ubiquitous evidence for genetic dominance in phenotypic and disease data, direct detection of the dominant or recessive action of natural selection in humans remains elusive. Unlike the analogous question in model organisms, experimental systems designed to differentiate between distinct types of selection in humans are infeasible. As such, there is a need for the development of a statistical test for the mode of selection from natural population samples. Here, we present such a test, derived from the differential action of recessive and dominant selection in populations with distinct demographic histories. We verify the efficacy of this test by showing that genes predicted to be recessive are enriched for genes responsible for recessive disease. Additionally, genes predicted to act dominantly do not show enrichment in this gene set. This pattern is consistent in both a large set of genes associated with autosomal recessive disease, and to a greater extent in a hand curated set of genes with no association with any disease phenotypes in heterozygous form. Aggregating genes in this high quality list into a unit, we find that the statistical test predicts recessivity for this sub-genome, in qualitative agreement with the enrichment analysis. This analysis was repeated for genes known to be responsible for congenital hearing loss, a well-known autosomal recessive disease. We again find agreement between our statistical prediction and the mode of inheritance of the disease phenotype. We confirm the detection of dominant selection by applying this test on the whole genome level to show that derived mutation classes thought to be under biased gene conversion are predicted to be under co-dominant selection. This is consistent with theoretical predictions for conversion of heterozygous sites. Together, these analyses allow us to quantify the joint distribution of selection and dominance effects, both on a per gene and on a whole genome level. This has important implications for medical and population genetics, particularly in helping us to understand the role and relevance of genetic dominance. Additionally, identifying candidate genes under recessive selection may aid in the discovery, understanding, and treatment of diseases with substantial genetic components.

## 13

**Exome analyses reveal new autism genes in synaptic, transcriptional, and chromatin networks.** *S. De Rubeis[1,2], K. Roeder[3,4], B. Devlin[5], M.J. Daly[6,7,8], J.D. Buxbaum[1,2,9,10,11,12], The Autism Sequencing Consortium.* 1) Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 2) Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 3) Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; 4) Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; 5) Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA; 6) The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 7) Harvard Medical School, Boston, Massachusetts, USA; 8) Center for Human Genetic Research, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA; 9) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 10) Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 11) Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 12) The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

The genetic architecture of autism spectrum disorder involves the interplay of common and rare variation and their impact on hundreds of genes. During the past two years, large-scale whole-exome sequencing (WES) has proved fruitful in uncovering risk-conferring variation, especially when considering *de novo* variation, which is sufficiently rare that recurrent *de novo* mutations in a gene provide strong causal evidence. Here, we conduct the largest ASD WES study to date, starting with 16,098 samples from seventeen distinct sample sources and ascertained by diverse designs. Unlike earlier WES studies, we do not rely solely on counting *de novo* loss-of-function (LoF) variants, rather we use novel statistical methods (Xin He for the Autism Sequencing Consortium) to assess association for autosomal genes by integrating *de novo*, inherited and case-control LoF counts, as well as *de novo* missense variants predicted to be damaging. Analyses of sequence data in a cleaned sample of 3,871 autism cases and 9,937 ancestry-matched or parent controls implicate 22 autosomal genes at a false discovery rate (FDR) < 0.05, and a much broader set of 107 autosomal genes strongly enriched for those likely to affect risk (FDR < 0.30). These 107 genes show unusual evolutionary constraint against mutations and map to modules in unbiased networks of early neocortical developmental. Our dataset is enriched with genes targeted by two autism-associated RNA-binding proteins (FMRP and RBFOX), genes found with *de novo* non-synonymous mutations in schizophrenia, and genes encoding synaptic components. Amongst critical synaptic genes found mutated in our study are voltage-gated ion channels, including those involved in propagation of action potentials (e.g., the $Na^+$ channel $Na_v1.2$ encoded by *SCN2A*), neuronal pacemaking, and excitability-transcription coupling (e.g., the $Ca^{2+}$ channel $Ca_v1.3$ encoded by *CACNA1D*). Our dataset is also enriched for chromatin remodeling genes, including enzymes involved in histone post-translational modifications, especially lysine methylation/demethylation, and regulators that recognize such marks and alter chromatin plasticity such as the emergent ASD gene *CHD8*. In conclusion, our study identifies a group of 107 high-confidence risk genes that incur *de novo* LoF mutations in over 5% of ASD subjects and expose two tightly intertwined pathways - chromatin remodeling and synaptic development - as major themes in ASD risk.

## 14

**Defining the contribution of different classes of de novo mutation to autism.** *I. Iossifov[1], B.J. O'Roak[2], S.J. Sanders[3,4], N. Krumm[5], M. Ronemus[1], D. Levy[1], J. Shendure[5], E.E. Eichler[5], M.W. State[3,4], M. Wigler[1].* 1) CSHL, Cold Spring Harbor, NY; 2) Molecular & Medical Genetics Department, Oregon Health & Science University, Portland, OR; 3) Department of Psychiatry, University of California, San Francisco, CA; 4) Department of Genetics, Yale School of Medicine, New Haven, CT; 5) Department of Genome Sciences, University of Washington, Seattle, WA.

Autism spectrum disorders (ASD) are the most prevalent form of neurodevelopmental disorders affecting ~1% of the human population, and are characterized by impairments in social interactions and communication as well as restricted interests and repetitive behavior. We have completed the exome sequencing of 2,519 simplex ASD families from the Simons Simplex Collection. For all families we have sequenced the affected child and both parents and, for 1,913 of the families, we have also sequenced an unaffected child.

We observe 392 'Likely Gene-Disrupting' (LGD) de novo mutations, which include nonsense, frame-shift and canonical splice-site mutations that are likely to result in a truncated gene product, and we estimate that 45% of them are contributory to the disease. There are 27 genes that are recurrently hit by de novo LGD mutations, 90% of which are likely to contribute to the disorder. *CHD8* is hit by 9 de novo LGDs (in nine affected children), *DYRK1A* by 4, and each of *ANK2, GRIN2B, DSCAM*, and *CHD2* is hit by 3 de novo LGDs. 147 genes are affected by more than one of the observed 1,688 missense mutations, and we expect that 30% of these genes are contributory. By measuring the recurrence against the 163 LGD mutations predicted to contribute to ASD risk and correcting for the distribution of gene sizes, we estimate a target size for LGD mutations between 400 and 1,200 genes, with a smaller range if we exclude affected males with higher IQ. Recurrence estimates of missense mutations show a similar target size and there is evidence of strong overlap between these two gene target classes. Genes affected by de novo mutation (LGD and missense) overlap strongly with sets of genes encoding transcripts bound by the Fragile X mental retardation protein (FMRP), genes encoding chromatin and transcription modulators, and genes expressed in early embryonic development. We find that affected children with de novo LGD mutations in recurrently hit genes or in FMRP associated genes have significantly lower non-verbal IQ.

We estimate that 9% of simplex autism can be attributed to de novo LGD mutations, and that de novo missense mutations contribute an additional 12%. Together with the established 6% contribution from de novo copy number variants, the total burden from observable de novo variants is nearly 30%. Moreover, de novo variants explain 45% of low IQ simplex autism.

## 15

**Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder.** *M. Uddin[1], K. Tammimies[1,2], G. Pellecchia[1], B. Alipanahi[3], P. Hu[1], Z. Wang[1], D. Pinto[4,5,6], L. Lau[1], T. Nalpathamkalam[1], C. Marshall[1,7], B. Blencowe[8,9], B. Frey[3], D. Merico[1], R. Yuen[1], S. Scherer[1,7,9].* 1) Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; 2) Neuropsychiatric Unit, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden.Center of Neurodevelopmental Disorders (KIND),; 3) Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada; 4) Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 5) Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 6) Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA; 7) McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada; 8) Donnelly Centre, University of Toronto, Toronto, Ontario, Canada; 9) Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.

A universal challenge in genetic studies of autism spectrum disorders (ASD) is determining whether a given DNA sequence alteration will manifest as disease. Among different population controls, we observed, for specific exons, an inverse correlation between exon expression level in brain and burden of rare missense mutations. For genes that harbor de novo mutations predicted to be deleterious, we found that specific critical exons were significantly enriched in individuals with ASD relative to their siblings without ASD ($P < 1.13 \times 10^{-38}$; odds ratio (OR) = 2.40). Spatio-temporal analyses reveal the most significant association is observed from prenatal samples and for prefrontal cortex region. The analysis of genes impacted by de novo CNVs also showed significant enrichment of critical exons with a similar spatio-temporal pattern. Furthermore, our analysis of genes with high exonic expression only in brain and low burden of rare mutations demonstrated enrichment for known ASD-associated genes ($P < 3.40 \times 10^{-11}$; OR = 6.08) and ASD-relevant fragile-X protein targets ($P < 2.91 \times 10^{-157}$; OR = 9.52). Gene set enrichment analysis for brain-critical exon genes showed enrichment in specific biological pathways involving synapse regulation, neuron differentiation, signaling complexes and synaptic vesicles. Our results suggest that brain-expressed exons under purifying selection should be prioritized in genotype-phenotype studies for ASD and related neurodevelopmental conditions.

## 16

**The landscape and clinical impact of cryptic structural variation in autism and related neuropsychiatric disorders.** *H. Brand[1,2,3,8], V. Pillalamarri[1,8], R. Collins[1], S. Eggert[1,4], M. Stone[1], I. Blumenthal[1], C. O'Doushlaine[3], E. Braaten[2], J. Rosenfeld[5], S. Mccarroll[3,4,6], J. Smoller[1,2,7], A. Doyle[1,2,3,6,7], M. Talkowski[1,2,3,7].* 1) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 2) Departments of Psychiatry, Neurology, Harvard Medical School; 3) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard; 4) Department of Genetics, Harvard Medical School; 5) Signature Genomic Laboratories, Perkin Elmer Inc., Spokane, WA; 6) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard; 7) Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital; 8) Co-First Author.

Rare, loss-of-function (LoF) structural variation (SV) represents a major component of the genetic risk in autism spectrum disorder (ASD) and other neuropsychiatric disorders (NPD), however much of the etiology remains unexplained. Most SV studies have focused on copy number variants (CNV) using microarray technology, which is blind to the presence of balanced chromosomal abnormalities (BCAs). Indeed, the diagnostic yield of BCAs in ASD is unknown as CNV analysis is the currently recommended first-tier screen, and there is no standardized clinical approach in other NPDs such as early onset schizophrenia or bipolar disorder. We have previously shown that cytogenetically visible BCAs represent a unique class of highly penetrant LoF variation, however the impact of cryptic BCAs is uncharacterized as they remain intractable to all conventional technology other than deep whole-genome sequencing (WGS). Here, we delineate the full mutational spectrum of SV in the initial 250 families in an SV sequencing study of ASD or other common NPDs (ADHD, bipolar disorder, major depressive disorder) using large-insert jumping library WGS with a median insert of 3.7 kb a genome-wide average insert coverage of 64X. We discovered a spectrum of SVs ranging from karyotypically visible and clinically significant BCAs to highly complex cryptic chromothripsis. We also find numerous 'duplications' as delineated by microarray that are duplicated insertions or complex rearrangements that can also cause LoF at the insertion site(s). On average, we find 61 cryptic CNVs and 48 cryptic BCAs per genome at this resolution. We also find that, on average, cryptic SVs detectable in our study rearrange 3.6 Mb of the genomes, disrupting 98 protein coding genes, and result in at least 19 additional LoF variants per individual. Using a convergent genomic approach of assessing CNV and exome burden from over 50,000 additional subjects, we estimate the clinically significant yield of cryptic SVs to be >5%, a potentially remarkable component of the disease variance that remains uncharacterized. Clinically significant SVs included de novo inversion of a known driver of a microdeletion syndrome (2q23.1), complex chromothripsis, and discovery of novel ASD genes such as *UBE2F*. These data suggest that the clinical yield from cryptic SVs in ASD and NPDs is substantial, and that SV detection warrants consideration in diagnostic testing for early onset NPDs. Analyses of larger cohorts is ongoing.

## 17

**Diagnostic utility of whole exome sequencing and chromosomal microarray in a clinically well-defined autism spectrum disorder cohort.** *K. Tammimies[1], B.A. Fernandez[2,3], S. Walker[1], B. Thiruvahindrapuram[1], G. Kaur[1], A.C. Lionel[1], W. Roberts[4], R. Weksberg[5], J.L. Howe[1], M. Uddin[1], R.K.C. Yuen[1], Z. Wang[1], L. Lau[1], P. Szatmari[6], K. Whitten[7,8], C. Vardy[7], V. Crosbie[8], B. Tsang[1], R. Liu[1], L. D'Abate[1], S. Luscombe[7], T. Doyle[7], S. Stuckless[2], D. Merico[1], D.J. Stavropoulos[9], C.R.M. Marshall[1,10], S.W. Scherer[1,11].* 1) The Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; 2) Disciplines of Genetics and Medicine, Memorial University of Newfoundland, St John's NL Canada; 3) Provincial Medical Genetic Program, Eastern Health, St. John's NL, Canada; 4) The Autism Research Unit, Hospital for Sick Children, Toronto, Ontario, Canada; 5) Department of Pediatrics, Division of Clinical and Metabolic Genetics, Hospital for Sick Children, Toronto, Ontario, Canada; 6) Centre for Addiction and Mental Health, University of Toronto, Toronto, Canada; 7) Child Health Program, Eastern Health, St. John's NL, Canada; 8) Discipline of Pediatrics, Memorial University of Newfoundland, St John's NL Canada; 9) Cytogenetics Laboratory, Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Canada; 10) Molecular Genetics, Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Canada; 11) Department of Molecular Genetics, McLaughlin Centre, University of Toronto, Toronto, Canada.

Autism Spectrum Disorder (ASD) is a group of neuropsychiatric conditions affecting 1 in 68 children. Twin and family studies have demonstrated a strong genetic basis for ASD; however both the genetic factors involved and the clinical presentation are highly heterogeneous. Etiological genetic variants have been identified in ~20% of ASD individuals, including chromosomal rearrangements, monogenic syndromes and rare de novo and inherited copy number variants (CNVs). Recent genetic studies of ASD using whole exome sequencing (WES) and whole genome sequencing have highlighted de novo and inherited sequence level variants, but their etiological contributions and diagnostic detection rate are still largely undefined. As the presentation of ASD is highly variable, it has been suggested that clinical morphology can aid in stratifying individuals into more genetically meaningful subtypes. Therefore, we performed comprehensive genomic analyses, combining high resolution chromosomal microarray (CMA) and WES, in a cohort of 284 ASD cases. Established detailed dysmorphology scoring and assessment for major congenital anomalies were used to stratify this cohort into 182 essential (non-dysmorphic), 41 equivocal dysmorphic and 61 complex (significantly dysmorphic) ASD subjects. We aimed to assess the rate of clinically relevant variants across the three subgroups and explored the contributions of different types and combinations of genetic variants to the variability of the clinical phenotype. Similar to previous reports of CMA diagnostic yield, we detected pathogenic CNVs in 7.2% of ASD individuals. The yield of such CNVs was significantly higher in the complex group 17.3% (9/52) compared with the essential group 3.1%, (5/163) (p=0.0003). These pathogenic CNVs affected known ASD loci, disrupting genes such as *SHANK3* and *NRXN1*. We also discovered novel CNVs with unknown clinical significance affecting genes for further follow up such as *CTNND2* and *GRK4*. The analysis of diagnostic yield of rare de novo and inherited mutations from WES is ongoing. We have identified de novo point mutations in new candidate genes such as *HTR5A* and *SCUBE2* as well as genes previously implicated in ASD such as *SCN2A* and *ASH1L*. Our data illustrates the value of detailed clinical information in the interpretation of genome-wide diagnostic test results of ASD. We also establish for the first time the combined diagnostic yield of WES and CMA genetic testing in an ASD cohort.

## 18

**Integrative functional genomics following suppression of CHD8 identifies transcriptional signatures that are enriched for autism genes and macrocephaly.** *M. Biagioli[1,4], A. Sugathan[1,2,4], C. Golzio[3], I. Blumenthal[1,2], S. Erdin[1,2], P. Manavalan[1], A. Ragavendran[1,2], D. Lucente[1], J. Miles[5], S.D. Sheridan[1], A. Stortchevoi[1,2], S.J. Haggarty[1,2,4,6], J.F. Gusella[1,4,6,7], N. Katsanis[3], M.E. Talkowski[1,2,4,6].* 1) Molecular Neurogenetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114; 2) Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114; 3) Center for Human Disease Modeling and Department of Cell biology, Duke University, Durham, NC 27710; 4) Department of Neurology, Harvard Medical School, Boston, MA 02114; 5) Departments of Pediatrics, Medical Genetics and Pathology, The Thompson Center for Autism and Neurodevelopmental Disorders, University of Missouri Hospitals and Clinics, Columbia, MO 65201; 6) Broad Institute of M.I.T. and Harvard, Boston, MA 02142; 7) Department of Genetics, Harvard Medical School, Boston, MA 02115.

Inactivating mutations of CHD8, and of a spectrum of genes with diverse functions, act as strong-effect risk factors for autism spectrum disorder (ASD), suggesting multiple mechanisms of pathogenesis. We perturbed the transcriptional networks that CHD8 regulates early in neurodevelopment by suppressing its expression in iPS-derived neural precursors using 5 independent shRNAs. We integrated transcriptome sequencing with genome-wide CHD8 binding from three independent antibodies. Suppressing CHD8 altered expression of 1,756 genes, most of which were up-regulated (~65%), consistent with its putative function as a transcriptional repressor. ChIP-seq revealed pervasive CHD8 binding, with 7,324 replicated sites from all three antibodies marking 5,658 genes, yet just 9% of these genes were differentially expressed. These data suggest that a limited array of direct regulatory effects of CHD8 produces a larger network of expression changes through secondary indirect regulatory mechanisms. Interestingly, the networks associated with direction of CHD8 regulation are functionally distinct. Genes indirectly down-regulated (i.e., without CHD8 binding sites) are strongly enriched for genes associated with ASD ($p = 1.0\times10^{-9}$) and reflect pathways involved in brain development. In contrast, genes with CHD8 binding sites are associated with cell maintenance and transcriptional regulation and are enriched for cancer related genes from three independent cancer datasets (enrichment $p < 1.0\times10^{-10}$ in all datasets). Among the most significant genes differentially expressed were known ASD genes (e.g., SCN2A, SHANK3) as well as a series of cell adhesion molecules. The most significant pathway and phenotypic enrichment observed was related to 'abnormality of skull size', prompting us to study *in vivo* effects. We observed macrocephaly in zebrafish models of *chd8* suppression initially using morpholino knockdown that was replicated by CRISPR, comparable to the phenotype reported in humans with inactivating mutations. These data indicate that heterozygous disruption of CHD8 precipitates gene expression changes that include indirect down-regulation of many other ASD risk genes. Perturbation of other genes within this ASD network are ongoing, and our results collectively suggest that some genes associated with ASD and neurodevelopmental disorders may converge on shared mechanism of pathogenesis.

## 19

**Transcriptional consequences of 16p11.2 microdeletion/microduplication syndrome in mouse cortex converges on genes and pathways associated with autism and known intellectual disability syndromes.** *I. Blumenthal[1], A. Ragavendran[1], S. Erdin[1], C. Golzio[2], A. Sugathan[1], J. Guide[1], V. Wheeler[1], A. Reymond[3], N. Katsanis[2], J.F. Gusella[1,4,5], M.E. Talkowski[1,4,5].* 1) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 2) Center for Human Disease Modeling and Department of Cell biology, Duke University, Durham, NC; 3) Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland; 4) Broad Institute of MIT and Harvard, Cambridge, MA; 5) Department of Neurology, Harvard Medical School, Boston, MA.

Reciprocal copy number variation (CNV) of a 593 kb region of 16p11.2 is a common genetic cause of autism spectrum disorder (ASD). However, it is not fully penetrant and confers risk across diverse phenotypic outcomes, including ASD, schizophrenia, obesity, and other neurological and anthropometric traits. To explore its molecular consequences, we performed RNA-sequencing in mouse models harboring CNV of the syntenic 7qf3 region. We initially sequenced the cerebral cortex of 16 mice (4 del, 4 dup, 8 control). Expression of all genes in the CNV region correlated with DNA copy number, with no evidence of dosage compensation. We observed positional effects in cis which were restricted to segments spanning a second genomic disorder locus: the 'distal' 16p11.2 syndrome in human. Overall, 16p11.2 CNV was associated with altered expression of genes and networks that converge on multiple hypotheses of ASD pathogenesis, including synaptic function (e.g., NRXN1, NRXN3), chromatin modification (e.g., CHD8, EHMT1, MECP2), and transcriptional regulation (e.g.,TCF4, SATB2). Notably, we observed differential expression of genes involved in classic forms of intellectual disability (e.g., FMR1, CEP290, BBS12), and among OMIM phenotypes we find strong statistical enrichment for genes related to Joubert and Bardet-Biedl syndromes ($p = 2.36e\text{-}4$ and 0.046, respectively), as well as pathways related to ciliary development ($p = 0.025$). Collectively, these data show perturbation of genes and networks involved in ASD and a spectrum of phenotypes associated with intellectual disability. We also performed RNA-seq analyses in lymphoblastoid cell lines (LCL) of multiplex families from AGRE, and compared our results to microarray studies of 81 CNV cases and 17 controls (see Reymond et al. abstract). These LCL datasets show striking consistency between the specificity of the cis positional effects, as well as the association with Bardet-Biedl and Joubert syndromes ($p = 0.026$ and 0.047, respectively), as well as obesity related pathways. Given the consistency of these findings and the pleiotropic effects of 16p11.2 CNV, we have expanded these analyses to 48 mice and 6 tissues (cortex, cerebellum, striatum, white fat, brown fat, liver), and these analyses are in progress. Our results suggest that 16p11.2 CNVs disrupts expression networks that involve other genes and pathways known to contribute to ASD and intellectual disability, suggesting an overlap in mechanisms of pathogenesis.

## 20

**Most genetic risk for autism resides with common variation.** *J.D. Buxbaum[1], B. Devlin[2], K. Roeder[3], the Population-based Autism Genetics and Environment Study (PAGES) Consortium.* 1) Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA; 3) Statistics, Carnegie Mellon University, Pittsburgh, PA.

Autism spectrum disorder (ASD) is a neurodevelopmental disorder typified by striking deficits in social communication. A key component of genetic architecture is the allelic spectrum influencing trait variability. For idiopathic ASD, the nature of its allelic spectrum is uncertain. Individual risk genes have been identified from rare variation, especially de novo mutations. From this evidence one might conclude that rare variation dominates its allelic spectrum, yet recent studies show that common variation, individually of small effect, has substantial impact en masse. At issue is how much of an impact relative to rare variation. Using a unique epidemiological sample from Sweden, novel methods that distinguish total narrow-sense heritability from that due to common variation, and by synthesizing results from other studies, we reach several conclusions about ASD genetic architecture: its narrow-sense heritability is ≈54% and most traces to common variation; rare de novo mutations contribute substantially to individuals' liability; still their contribution to variance in liability, 2.6%, is modest compared to heritable variation. Concurrent with our study, a comprehensive family study of autism in Sweden has been ongoing and it recently reported the largest study of familial risk for autism to date. The study analyzed additive and non-additive genetic effects and shared and non-shared environmental effects in a series of structured models to estimate which factors have a substantial effect on RR and thus heritability. The best model, consisting only of additive genetic and non-shared environmental effects, yielded quite precise estimates of the narrow-sense heritability of AD (h2=54%, SE=5). That these studies, with very different design, converge on similar estimates of heritability lends strong support for our conclusion that the bulk of risk for autism arises from genetic variation. This study was supported by National Institute of Mental Health (NIMH) Grants MH057881 and MH097849.

## 21

**Utilizing rare variants for phasing and imputation in pedigrees.** *A. Blackburn, J. Blangero, H. Göring.* Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX.

Whole genome sequencing of multigenerational pedigrees is poised to be the new paradigm of statistical genetics studies in the post-GWAS era. Efficient methods to phase and impute whole genome sequencing data between sequenced and non-sequenced pedigree members are essential to realizing the potential of this paradigm. Methods to phase whole genome sequencing data in extended pedigrees will enable haplotype specific genetic analyses, improve identification of compound heterozygosity, improve genotype error checking, and can be used to reduce the multiple testing burden inherent to the whole genome sequencing approach by applying segment based tests. Current imputation methods in related individuals are limited by pedigree size, by the distance of relationships, or by computation time. Here we explore the potential to utilize rare variants to identify DNA segments shared between pedigree members that are identical by descent using simulated and real whole genome sequencing data. We apply this information to phase and impute whole genome sequencing genotypes, and to reduce multiple testing burden. To fully explore the robustness of this approach across pedigree structures, we randomly simulated pedigree structures ranging from small nuclear families to large pedigrees consisting of up to ~2500 individuals from 10 generations. Whole genome sequencing data was then simulated for these pedigrees based on Kimura's infinite alleles model under the assumption of neutrality. Additionally, real whole genome sequencing data was used to simulate founder genotypes and transmission was simulated through the pedigree. Imputation accuracy was estimated for diallelic variants across levels of available sequencing data and sequencing accuracy by masking genotypes and using the IQS statistic to correct for random concordance of genotypes. Imputation accuracy varies based on levels of available data and pedigree size, while being generally robust to low genotyping error rates. We conclude that rare variants, especially those that are specific to a single founder, are of increased utility compared to common variants for the purpose of phasing and imputation of whole genome sequencing genotypes in multigenerational pedigrees.

## 22

**The "Jackpot" Effect - When Do Family Samples Provide More Power To Detect Trait-associated Rare Variants?** *S. Feng[1], G. Pistis[1,2], A. Mulas[2], M. Zoledziewska[2], F. Busonero[2], S. Sanna[2], D. Liu[3], F. Cucca[2], G.R. Abecasis[1].* 1) Biostatistics, University of Michigan, Ann Arbor, MI 48109; 2) Istituto di Ricerca Genetica e Biomedica-CNR, Monserrato (CA), 09042, Italy; 3) Department of Public Health Sciences, College of Medicine, Penn State University, Hershey, PA, 17033.

Population samples have been used to discover many trait associated common variants using array-based GWAS, but may not be the optimal choice for detecting rare variants with moderate to large effects. Here, we set out to compare the power of population and family samples to study trait associated rare variants. We describe settings where family samples can provide more power than population samples for rare variant association studies. In particular, we show that in population samples observing enough trait associated alleles in each causal gene can be extremely challenging and require very large samples. In contrast, in family samples that include large numbers of related individuals, there will more often be one or more genes where many observations of a trait associated allele are made.

For example, by simulation, we show that when there are 100 loci where trait associated variants segregate at a frequency of ~0.1% and modify a quantitative trait by one standard deviation, a study of 5,000 unrelated individuals provides ~60% power to detect at least one of these loci at genomewide significance. In contrast, a study of 5,000 individuals distributed across 100 large families, each with ~50 individuals, provides ~90% power to detect at least one of the trait associated loci in the same setting. For gene-level association tests similar results are observed and the contrast is even more dramatic when association signal is dominated by singleton variants, defined as variants present in a single founder or unrelated individual.

The advantages of family samples are due to a "Jackpot" effect, where multiple copies of some trait-associated rare alleles are shared among individuals with a common ancestor, greatly increasing power. We show that the advantages of family samples for rare variant studies increase as the number of individuals with a shared recent ancestor increases. Using simulations based on an isolated population sample from the island of Sardinia, we show often dramatic differences in power. Our results provide guidance to investigators hoping to identify trait associated rare variants and deciding between family and population based designs.

## 23

**Sequence Kernel Association Test for Multivariate Quantitative Phenotype in Family Samples.** *Q. Yan[1], B. Li[2], W. Chen[1], N. Liu[3].* 1) Division of Pulmonary Medicine, Allergy and Immunology, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA, 15224, USA; 2) Department of Molecular Physiology & Biophysics, and Neurology, Vanderbilt University Medical Center, Nashville, TN 37232, USA; 3) Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, 35294, USA.

The recent development of sequencing technology allows identification of the association between rare genetic variants and complex diseases. Over the past few years, a number of rare variant association test approaches have been developed. Among these methods, the kernel machine test as a set-based method has been shown to perform well in different scenarios. Many genetic studies have been conducted to collect multiple correlated phenotypes for one complex disease. Because of the relatedness between phenotypes, jointly testing the association between phenotypes and genetic variants may increase the power to detect causal genes. In addition, family based designs have been widely used to study the association between diseases and genetic factors. Thus, familial correlation needs to be appropriately handled to avoid inflated type I error rate. In this work, we aim to conduct the association test of rare variants in family samples, which uses multiple phenotype measurements for each subject. Our proposed method uses kernel machine regression and denoted as MF-SKAT. It is based on linear mixed model framework and can be applied to a larger range of studies with different types of traits. In our simulation studies, the results show that the kernel machine test (M-SKAT considering the correlation between multiple phenotypes) has inflated Type I error rate when applying to familial data directly. By contrast, our proposed MF-SKAT has correct Type I error rate. Furthermore, MF-SKAT jointly analyzing phenotypes has increased power comparing to the methods separately analyzing phenotypes (F-SKAT considering the family structure) in all the scenarios we considered. Finally, we illustrate our proposed methodology by analyzing whole-genome genotyping data from a study of lung function and exome sequencing data from a study of Anorexia Nervosa.

## 24

**Genetic network inference in studies of multiple phenotypes from related individuals.** *J. Marchini[1,2], A. Dahl[2], V. Hore[1].* 1) Dept Statistics, Oxford Univ, Oxford, United Kingdom; 2) Wellcome Trust Centre for Human Genetics, Oxford Univ, Oxford, United Kingdom.

As the field of human genetics moves beyond simple GWAS, we are observing two clear trends in the types of datasets being collected, and the types of analysis used, to help elucidate disease etiology. Firstly, cohorts are collecting multiple disease and molecular phenotypes on the same samples. Secondly, heritability analysis in related samples is becoming more popular. In such datasets, it maybe of interest to infer the underlying network of dependence between phenotypes that has a genetic basis. Such networks summarize graphically the joint heritability of the measured phenotypes. Inference of such a network is complicated, as can we expect significant correlations between phenotypes, due to both the underlying network and observational noise. We model observations as a sum of two matrix normal variates, such that the joint covariance function is a sum of Kronecker products. This model, which generalizes the Graphical Lasso, assumes observations are correlated due to known genetic relationships and corrupted with non-independent noise. We have developed a computationally efficient EM algorithm to fit this model. On simulated datasets we illustrate substantially improved performance in network reconstruction by allowing for a general noise distribution.

## 25

**G-STRATEGY: Optimal Selection of Individuals to Genotype in Genetic Association Studies with Related Individuals.** *M. Wang[1], J. Jakobsdottir[2], A.V. Smith[2,3], M.S. McPeek[1,4].* 1) Department of Statistics, University of Chicago, Chicago, IL, USA; 2) Icelandic Heart Association, Holtasmári 1, IS-201 Kópavogur, Iceland; 3) University of Iceland, Reykjavík, Iceland; 4) Departments of Human Genetics, University of Chicago, Chicago, IL, USA.

A common problem in genetic association studies is choosing a fixed-size subset of individuals to genotype. The problem arises naturally in studies in which the genotyping budget is limited. Suppose a cohort of phenotyped individuals is available, with some subset of them possibly already genotyped, and one wants to choose an additional fixed-size subset of individuals to genotype in such a way that the power to detect association is maximized. When the phenotyped sample includes related individuals, power can be gained by including partial information, such as phenotype data of ungenotyped relatives, in the association analysis. It is important to take this into account when assessing whom to genotype. We propose G-STRATEGY, a method for selection of individuals for genotyping, conditional on phenotypes and kinship. G-STRATEGY uses simulated annealing to maximize the noncentrality parameter of either the MQLS or MASTOR statistic, both of which increase power in this context by incorporating phenotype information on ungenotyped relatives. In simulations, G-STRATEGY performs well for a range of complex disease models and outperforms other strategies (selection of maximally unrelated individuals, extreme phenotype enrichment, and GIGI-pick, a previously proposed method) with, in many cases, relative power increases of 20-40% over the next best strategy, while maintaining correct type 1 error. Importantly, G-STRATEGY is computationally feasible even for large datasets. When we applied G-STRATEGY to data on high-density lipoprotein (HDL) from the AGES-Reykjavik and REFINE-Reykjavik studies, with over 8000 phenotyped and 3000 genotyped individuals, it took G-STRATEGY <5 minutes to choose 380 additional individuals for genotyping, from among those not already genotyped. To further evaluate performance, we masked the available genotypes during the selection process, and selected either 1000 or 2000 individuals for genotyping from among those with masked genotypes. For the resulting samples, we then ran targeted association analysis among known HDL genes. For association with SNPs in CETP, based on 1000 individuals chosen by G-STRATEGY, a p-value of 8×10-13 is obtained, while the smallest p-value for the maximally-unrelated strategy is 2×10-10. With 2,000 individuals, the corresponding p-values are 2×10-19 for G-STRATEGY and 9×10-13 for the maximally-unrelated strategy, demonstrating the power advantage G-STRATEGY can provide over other methods.

## 26

**Using a population-based linkage analysis approach to identify transcript QTL in skeletal muscle tissues in a founder population.** *W.-C. Hsueh, S. Kobes, R.L. Hanson.* PECRB, NIDDK, NIH, Phoenix, AZ.

Linkage analysis offers an unbiased way to identify QTLs and it may identify loci with multiple and/or rare variants that are not amenable to association studies. However, its power is limited by the need to calculate identical-by-descent allele sharing (%IBD) in known relatives. Recent development in statistical methods to calculate %IBD for all possible pairs of subjects using actual genotypes (e.g. GWAS SNP data) could boost the power for linkage studies in founder populations greatly by providing more precise %IBD estimates, including for pairs of subjects with no known relationship. We assessed its utility to identify eQTLs in muscle tissue. Skeletal muscle biopsies were obtained from 149 healthy Pima Indians. Transcript levels were measured on the Affymetrix Human Exon 1.0 ST Array. Both genome-wide and gene-specific %IBD for all pairs of subjects were estimated based on ~400,000 SNPs with Beagle. We defined the gene-specific region to include SNPs within a given transcript and ±200kb of its coding region. The mean %IBD for "unrelated" pairs was 0.021, comparable to that for 2nd cousins, consistent with a founder population. We conducted linkage analyses to identify *cis*-eQTLs, adjusting for age, sex and the 1st principal component from our GWAS. Permutation analyses showed that a LOD score ≥2.07 was equal to an empirical false discovery rate <0.05. In *cis*-linkage analyses of 16,840 core autosomal transcripts with SNP data, we identified 188 loci with LOD ≥2.07 (mean: 3.74±2.04, range: 2.08-13.81). The mean effect size of the *cis*-elements at these loci was 30%±10% (range: 21%-100%). To assess power, we selected a SNP at the midpoint of each transcript region and simulated a hypothetical QTL with specified effect size at each of 15,451 unique SNPs. We conducted power analyses using the %IBD estimates based on both all SNP data (n=11,026 pairs) and self-reported relationships (n=234 pairs). We have 80% power to detect QTLs with effect size ≥48% for linkage analysis using empirically estimated %IBD, compared to 3.4% power using pedigree-based %IBD estimates. In summary, using one of the largest samples for such studies in muscle, we identified 188 *cis*-eQTLs, which can be prioritized for follow-up association studies with transcripts and related clinical outcomes. Our data also suggest that population-based linkage studies in founder populations using empirically derived %IBD may provide much greater power than using pedigree-based estimates.

## 27

**Testing for disease association with rare compound heterozygous and recessive mutations in case-parent sequencing studies.** *A. Allen[1,2], Y. Jiang[1], J. McCarthy[1].* 1) Biostatistics & Bioinformatics, DukeUniversity, Durham, NC; 2) Center for Human Genome Variation, Duke University School of Medicine, Durham, NC.

Compound heterozygous mutations occur on different copies of genes and may completely "knock-out" gene function. Compound heterozygous mutations have been implicated in a large number of diseases, but there has been very little work developing statistical methods for testing for their role in disease. A major barrier is that phase information is required to determine that both gene copies are affected and phasing rare variants is difficult. Here, we propose a method to test compound heterozygous and recessive disease models in case-parent trios. We propose a simple algorithm for phasing and show via simulations that tests based on phased trios have almost the same power as tests using true phase information. A further complication is that only families where both parents carrying mutations are informative. Thus, the informative sample size will be quite small even when the overall sample size is not, making asymptotic approximations of the null distribution of the statistic inappropriate. In order to deal with this issue we develop an exact test that will give accurate p-values regardless of sample size. Using the simulation, we show that our method is robust to sequencing errors, population stratification, and significantly outperforms other methods when the causal model is recessive.

## 28

**Statistical Approaches for Rare-Variant Association Testing in Affected Sibships.** *M.P. Epstein[1], E. Ware[2], M.A. Jhun[2], L.F. Bielak[2], W. Zhao[2], J. Smith[2], P.A. Peyser[2], S.L.R. Kardia[2], G.A. Satten[3].* 1) Dept Human Gen, Emory Univ, Atlanta, GA; 2) Dept Epidemiology, University of Michigan, Ann Arbor, MI; 3) Centers for Disease Control and Prevention, Atlanta, GA.

The emergence of sequencing and exome-chip technologies has propelled the development of novel statistical tests to identify rare genetic variation that influence complex diseases. While many rare-variant association tests exist for case-control or cross-sectional studies, far fewer methods exist for association testing in families. This is unfortunate, as families possess many valuable features for rare-variant mapping that population-based studies lack. Many projects have begun sequencing or exome-chipping relatives from families; either recently sampled or previously collected as part of linkage studies. As many past linkage studies employed an affected-sibship design, we propose a novel test of rare-variant association for use in such studies. The logic behind our approach is that, for relative pairs concordant for phenotype, rare susceptibility variation should be found more often on regions shared identical by descent. Our approach is applicable to affected sibships of arbitrary size and does not require genotype information from either unaffected siblings (although this information is helpful, if available) or independent controls. The method is robust to population stratification, can adjust for other covariates, and produces analytic p-values, thereby enabling the approach to scale to genome wide studies of rare variation. We illustrate the approach using exome chip data from sibships ascertained for hypertension collected as part of the GENOA study.

## 29

**Functional partitioning of Prostate Cancer heritability in European Americans and African Americans from AAPC and BPC3 consortia reveals tissue specific regulation.** *B. Pasaniuc[1,2], A. Gusev[3], F.R. Schumacher[4], S. Lindstrom[3], M. Pomerantz[5], F. Li[5], H. Long[5], P. Kraft[3,6], A.L. Price[3,6], M. Freedman[5,7], C.A. Haiman[4], The BPC3 Consortium, The AAPC Consortium.* 1) Pathology and Laboratory Medicine, Geffen School of Medicine at UCLA, Los Angeles, CA; 2) Human Genetics, Geffen School of Medicine at UCLA, Los Angeles, CA; 3) Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health. Boston MA; 4) Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; 5) Medical Oncology, Dana-Farber Cancer Institute, Boston MA; 6) Dept. Epidemiology, Harvard School of Public Health. Boston MA; 7) Department of Medicine, Harvard Medical School, Boston, MA.

Although GWAS have identified over 90 genetic loci associated with prostate cancer, they jointly explain a fraction of the overall genetic component of the risk, leaving much of heritability missing. Here we use variance components models to dissect the genetic contribution to risk of prostate cancer across various functional categories identified in the ENCODE project using genome-wide array SNP data across ~20,000 prostate cancer cases and controls of European and African-American ancestry from the BPC3 and AAPC consortia. We find that 1000 Genomes imputed SNPs explain a heritability of h2g=0.19 (s.e. 0.04) in Europeans, a significant increase from heritability explained by the known risk variants in this data (0.08, s.e. 0.02). Interestingly, we find that typed and imputed SNPs explain similar amount of variance in prostate cancer in African Americans (h2g=0.23, s.e. 0.04) suggesting a similar genetic architecture to the risk of prostate cancer across the two ethnicities (meta h2g=0.21, s.e. 0.03, 36% of total total h2=0.58 [Hjelmborg et al 2014]). We partition heritability explained by SNPs across ENCODE functional elements and find that variants in DNaseI Hypersensitivity Sites (DHS) specific to Prostate Adenocarcinoma LNCaP cell lines that span only 1.1% of the genome explain 29% of h2g (s.e. 13%, 25-fold enrichment). We also explored the role of Androgen Receptor (AR) binding sites in prostate tissue, (which overlap by 33% with the LNCaP DHS) and found that variants in these regions (1.0% of genome) explain 37% of h2g (s.e. 11%, 39-fold enrichment). In comparison, coding variants that account for 0.7% of the genome only explain 5% of h2g (s.e. 8%, 8-fold enrichment). Neither LNCaP DHS nor AR-binding variants were significantly enriched in a meta-analysis of 11 common non-cancer traits from the WTCCC, serving as a negative control and supporting the cancer specificity of the tested functional annotations. Analyses of admixed populations present complexities and we show by simulation that our estimates of enrichment are robust to population structure and population-specific selection. Overall, our results demonstrate a significant contribution of phenotype-specific regulatory variants to the genetic risk for prostate cancer across different ethnic groups (no significant differences were observed between the two ethnicities in any analysis above).

## 30

**Prostate Cancer Risk Locus at 8q24 as a Regulatory Hub by Physical Interactions with Multiple Genomic Loci across the Genome.** *M. Du[1], T. Yuan[1], K. Schilter[1], R. Dittmar[1], A. Mackinnon[1], X. Huang[1], M. Tschannen[2], E. Worthey[2], H. Jacob[2], S. Xia[1,3], J. Gao[4], L. Tillmans[5], Y. Lu[6], P. Liu[6], S. Thibodeau[5], L. wang[1].* 1) Pathology, Medical college of wisconsin, Milwaukee, WI; 2) Human Molecular Genetics Center, Medical College of Wisconsin, WI 53226; 3) Department of Oncology, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, 430030; 4) Beijing 3H Medical Technology Co. Ltd., Beijing, China, 100176; 5) Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905; 6) Department of Physiology, Medical College of Wisconsin, Milwaukee, WI 53226.

Chromosome 8q24 locus contains regulatory variants that modulate genetic risk to various cancers including prostate cancer (PC). However, the biological mechanism underlying this regulation is not well understood. Due to lack of annotated genes, it is hypothesized that the 8q24 risk locus may affect other genes through long range chromatin interaction. To thoroughly survey the possible chromatin interactions, we developed a chromosome conformation capture (3C)-based multiple target sequencing (3C-MTS) technology to examine the 8q24 risk locus and its target loci in five prostate-derived cell lines and one lymphoblastoid cell line. By targeted enrichment of DNA fragments defined by 77 EcoRI sites covering all three PC risk regions (~302kb), we successfully identified multiple genomic regions that showed frequent intra- or inter-chromosomal interactions with the risk locus. We observed the most frequent inter-chromosomal interaction between CD96 intron 2 at 3q13 and 8q24 risk region 1. Subsequent 3C-qPCR assays and Fluorescence In Situ Hybridization (FISH) analysis confirmed the inter-chromosomal interaction. The second most common interaction occurred at MYC locus. We identified at least five interaction hot spots within the predicted functional regulatory elements at the 8q24 risk locus. We also found intra-chromosomal interaction genes PVT1, FAM84B, GSDMC and inter-chromosomal interaction gene CXorf36 in all six cell lines. Other gene regions appeared to be cell line-specific such as RRP12 in LNCaP, USP14 in DU-145 and SMIN3 in LCL. To predict potential functional consequences of these frequent contacts, we applied GREAT (Genomic Regions Enrichment of Annotations Tool) in the suggestive interaction regions and found that the 8q24 functional domains more likely interacted with genomic regions containing genes enriched in critical pathways such as Wnt signaling and promoter motifs such as E2F1 and TCF3. This result suggests that the risk locus may function as a regulatory hub by physical interactions with multiple genes important for prostate carcinogenesis. Using the 3C-MTS technology, for the first time, we are able to systematically examine the complete 8q24 risk locus and its potential target genes at genome-wide level. Further understanding genetic effect and biological mechanism of these chromatin interactions will shed light on the newly discovered regulatory role of the risk locus in PC etiology and progression.

## 31

**Genome-wide association study of 35K men with 300K prostate specific antigen measures identifies numerous novel loci: potential for personalized screening for prostate cancer.** *J.S. Witte[1,2], T.J. Hoffmann[1,2], L. Sakoda[3], E. Jorgenson[3], D.S. Aaronson[3], J. Shan[3], L.A. Habel[3], J.C. Presti[3], C. Schaefer[3], N. Risch[1,2,3], S.K Van Den Eeden[3].* 1) Institute for Human Genetics, UC San Francisco, San Francisco, CA; 2) Department of Epidemiology and Biostatistics, UC San Francisco, San Francisco, CA; 3) Kaiser Permanente, Division of Research, Oakland CA.

Using the prostate-specific antigen (PSA) test to screen for prostate cancer is controversial: it has modest predictive value and the potential to lead to over-diagnosis and over-treatment. One can improve the traditional single-cutpoint PSA screening test with more personalized thresholds to determine whether a man should be further evaluated for prostate cancer. Since there is a clear genetic component to PSA, integrating information about genetic factors that influence PSA independently of prostate cancer may improve test performance. To this end, we have undertaken a very large genome-wide association study of PSA concentrations among 35,520 men free of prostate cancer in the Kaiser Permanente Genetic Epidemiology Resource in Adult Health and Aging (GERA) cohort. These men had a total of 295,398 PSA measures from electronic health records. We evaluated the potential association between genome-wide variants and log(PSA) levels using a longitudinal generalized estimating equations model, adjusting for age, body mass index, and genetic ancestry. We detected over twenty independent genome-wide significant loci for PSA levels. Four of these have previously been reported as associated with PSA, and five with prostate cancer. The PSA replications were at *KLK3* ($p = 3.6\times10^{-151}$), *MSMB* ($p = 5.7\times10^{-33}$), *HNF1B* ($p = 1.1\times10^{-16}$), and *TBX5* ($p = 3.4\times10^{-12}$). The hits we detected for PSA that have previously been reported as associated with prostate cancer were at *JAZF1* ($p = 7.3\times10^{-18}$), 8p21.2 ($p = 4.6\times10^{-23}$), 8q24 ($p = 7.5\times10^{-15}$), 10q26.12 ($p = 1.6\times10^{-36}$), and 11q22.2 ($p = 8.5\times10^{-9}$). The strongest novel locus was on chromosome 13 ($p = 9.8\times10^{-19}$), and we detected numerous other significant novel loci for PSA concentrations, including on chromosomes 1, 2, 6, 9, 10, 14 and X. In addition, we detected over 30 'suggestive' hits that merit further follow up ($p < 10^{-5}$). The variants detected here may explain some of the inherent variability in—and the biological basis of—serum PSA concentrations. We are presently evaluating how much incorporating these variants into the assessment of serum PSA concentrations improves the decision of whether to perform a prostate needle biopsy and the resulting prostate cancer outcome. The large number of loci we have detected substantially helps our efforts here. This is an important step toward incorporating genetic markers into PSA screening, with the ultimate goal of devising personalized PSA tests for use in the clinic.

## 32

**Germline sequencing for genetic markers of aggressive prostate carcinoma susceptibility.** *D. Koboldt[1], K. Kanchi[1], D. Larson[1], R. Fulton[1], E. Mardis[1], A. Kibel[2].* 1) The Genome Institute, Washington University, 4444 Forest Park Pkwy., Saint Louis, MO; 2) Dana Farber Cancer Center, 45 Francis Street, Boston, MA.

Prostate cancer (PCa) is the second most common cancer among men in the United States and a growing medical problem worldwide. Most men will develop histologic prostate cancer, but few will be diagnosed with aggressive disease. Unfortunately, the most common screening tool (PSA) is not effective for making this prognosis. Several groups have sought genetic markers for risk of aggressive PCa. While GWAS approaches have identified and validated several PCa risk loci, none of the variants have been reproducibly linked to aggressive disease. One possible explanation for this is that genetic predisposition for aggressive PCa is driven by rare, functionally disruptive variants not interrogated by SNP arrays. To search for such variants, we performed exome sequencing in two populations: European-Americans (150 cases, 150 controls) and African Americans (122 cases, 150 controls). The latter are an important population to study for PCa; African-American men have a higher risk of disease, more advanced tumors at presentation, and a worse prognosis compared to European-Americans. Exome sequencing uncovered ~30,000 and ~35,000 rare, potentially-deleterious variants in the European-American and African-American cohorts, respectively. Gene-based burden tests revealed ~400 genes with significant case-control differences. Targeted sequencing of these genes in an independent cohort of 835 samples (223 cases, 612 controls) confirmed associations for 35 genes. In European-Americans, associated genes included *PCSK6*, encoding a proprotein convertase thought to play a role in tumor progression; and *DMBT1*, a candidate tumor suppressor gene for brain, lung, esophageal, gastric, and colorectal cancers. In African-Americans, among the associated genes were *DCLRE1A*, a regulator of mitotic cell checkpoint; and *TRPM2*, encoding a Ca(2+)-permeable channel that is activated by oxidative stress and confers susceptibility to cell death. Importantly, few genes showed evidence of association in both European-Americans and African-Americans, suggesting that population genetic differences may underlie the contrast in disease susceptibility between these groups.

## 33

**Identification of candidate target genes for prostate cancer risk-SNPs utilizing a normal prostate tissue eQTL dataset.** *S.N. Thibodeau[1], A.J. French[1], S.K. McDonnell[2], J.C. Cheville[1], S. Middha[2], S.M. Riska[2], S. Baheti[2], Z.C. Fogarty[2], L.S. Tillmans[1], M.C. Larson[2], N.B. Larson[2], A.A. Nair[2], D.R. O'Brien[2], J.I. Davila[2], Y. Zhang[2], L. Wang[3], J.M. Cunninghman[1], D.J. Schaid[2].* 1) Dept Lab Med & Pathology, Mayo Clinic, Rochester, MN; 2) Health Sciences Research, Mayo Clinic, Rochester, MN; 3) Medical College of Wisconsin, Milwaukee, WI.

Prostate Cancer (PC) is the most frequently diagnosed solid tumor in men in the U.S. For PC, multiple genome-wide association studies have now been performed yielding a substantial number of well-validated SNPs that are associated with an increased risk of PC. A significant problem for many of the PC risk-SNPs identified, however, is that they do not lie within or near a known gene and they have no obvious functional properties. These findings suggest that many of these risk-SNPs will be located in regulatory domains that control gene expression. However, in order to define the functional role of these non-coding risk-SNPs, the target genes must first be identified. A frequently used strategy to address this problem involves the use of expression quantitative trait loci (eQTL) analysis. In this study, we created a tissue-specific eQTL dataset and then applied this dataset to 123 well-established PC risk-SNPs in an effort to identify candidate target genes. To construct this dataset, normal prostate tissue from over 4000 men having a radical prostatectomy for PC was histologically screened in order to identify approximately 500 samples meeting a strict selection criteria: tissue localized to the posterior region of the prostate, no tumor, no high-grade PIN, no BPH, ≤2% lymphocytes, and ≥40% epithelial cells. Genome-wide genotypes and genome-wide mRNA expression levels were obtained with the use of the Illumina Human Omni 2.5M SNP array and by RNA sequencing, respectively. Of 500 processed samples, 471 samples passed stringent QC and were available for further analysis. Our primary analysis focused on identifying eQTLs for 123 PC risk-SNPs, including all SNPs in linkage disequilibrium with each risk-SNP ($r^2 > 0.5$), resulting in 78 unique risk-intervals. Furthermore, we focused on cis-acting associations only where the transcript was located within a 2Mb region (±1Mb) of the risk-SNP interval. Of all SNPs located in the 78 risk-intervals (N=5116 SNPs), 1002 demonstrated a significant eQTL signal after adjustment for sample histology (% lymphocytes and % epithelial cells) and meeting a Bonferroni-adjusted p-value threshold of 1.96e-7 (ranged from 1.52e-91). Of the 78 PC risk-intervals, 22 (28%) demonstrated a significant eQTL signal and these were associated with 43 genes, supporting a number of novel candidate susceptibility genes for PC. Mapping of the causative risk-SNPs and their corresponding affected regulatory elements is currently in progress.

## 34

**Association of Prostate Cancer Risk Variants with Gene Expression in Normal Prostate and Tumor Tissue.** *K.L. Penney[1,2], J.A. Sinnott[1,2,4], S. Tyekucheva[2,6], T. Gerke[2], I. Shui[2], P. Kraft[2,4], H.D. Sesso[2,5], M.L. Freedman[7,9], M. Loda[7,8,9], L.A. Mucci[1,2], M.J. Stampfer[1,2,3].* 1) Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA; 2) Department of Epidemiology, Harvard School of Public Health, Boston, MA; 3) Department of Nutrition, Harvard School of Public Health, Boston, MA; 4) Department of Biostatistics, Harvard School of Public Health, Boston, MA; 5) Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA; 6) Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA; 7) Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA; 8) Department of Pathology, Brigham and Women's Hospital, Boston, MA; 9) The Broad Institute, Cambridge, MA.

**Introduction:** Numerous germline genetic variants are associated with prostate cancer risk, but their biological role is only beginning to be understood. One possibility is that these variants influence gene expression in prostate tissue. We therefore examined the association of prostate cancer risk variants with the expression of genes nearby and genome-wide. **Methods:** We generated mRNA expression data for 20,254 genes with the Affymetrix GeneChip Human Gene 1.0 ST microarray from both normal prostate and prostate tumor tissue from 264 participants of the Physicians' Health Study and Health Professionals Follow-up Study Tumor Cohort. With linear models, we tested for the association of 39 risk variants with nearby genes (1 megabase window) and all genes. We also tested the association of each variant with canonical pathways using a global test. **Results:** In addition to confirming previously reported associations, we detected several new significant ($p<0.05$) associations of variants with the expression of nearby genes including *C2orf43*, *ITGA6*, *MLPH*, *CHMP2B*, *BMPR1B*, and *MTL5*. Genome-wide, four genes were statistically significantly associated after accounting for multiple comparisons ($p<2.5\times10^{-6}$), 367 genes in tumor and 16 genes in normal (including *SRD5A1* and *PSCA*) had a false discovery rate <10%, and we observed significant associations with several pathways in tumor tissue. **Conclusions:** Several genes were associated with the risk variants, including promising prostate cancer candidates and lipid metabolism pathways that should be further explored in biological and epidemiological studies. Determining the biological role of these variants can lead to a mechanistic understanding of prostate cancer etiology and possibly identify new targets for chemoprevention.

## 35

**Discovery and functional characterization of an oncogenic PTEN mutation: Implications for personalized cancer genome therapy.** *H.A. Costa[1], M.G. Leitner[2], M.L. Sos[3], A. Mavrantoni[2], A. Rychkova[1], M.C. Yee[1], F.M. De La Vega[1], J.M. Ford[1], K.M. Shokat[3], D. Oliver[2], C.R. Halaszovich[2], C.D. Bustamante[1].* 1) Stanford University School of Medicine, Department of Genetics, Stanford, CA, 94305, USA; 2) Philipps-Universität Marburg, Department of Neurophysiology, 35037, Marburg, Germany; 3) Univeristy of California, San Francisco, Department of Cellular and Molecular Pharmacology, Howard Hughes Medical Institute, San Francisco, CA, 94158, USA.

While a variety of genetic alterations have been found across cancer types, the functional characterization of genetic lesions in an individual patient and their translation into clinically actionable strategies remain major hurdles. Here, we use whole genome sequencing of a prostate cancer tumor (and matched germline genome), computational analysis, and experimental validation to identify a novel (A126G) oncogenic mutation in the PTEN tumor suppressor protein. We demonstrate that this mutation produces an enzymatic gain-of-function in PTEN, shifting its function from a phosphosinositide (PI) 3-phosphatase to a phosphosinositide (PI) 5-phosphatase and resulting in an increase in PI(3,4)P2 levels. Using cellular assays we show that PTEN A126G hyperactivates the PI3K/Akt cell proliferation pathway and exhibits increased cell invasion and we conclude that these effects can be substantially mitigated through chemical PI3K-β inhibitors. These results suggest a new dysfunction paradigm for PTEN cancer biology, highlight the difficulty of inferring the impact of mutations solely through variant annotation and bioinformatic analysis, and display the power of experimental validation to uncover new biology and guide individualized therapy and diagnostics.

## 36

**Convergent Genomics Validates *C2orf43* Role in Prostate Cancer.** *B.B. Currall[1,2], K.E. Wong[1], N.G. Robertson[1], A. Lunardi[2,3], M. Reschke[2,3], P.P. Pandolfi[2,3], C.C. Morton[1,2,4].* 1) Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women's Hospital, Boston, MA 02115, USA; 2) Harvard Medical School, Boston, MA 20115; 3) Cancer Research Institute, Beth Israel Deaconess Cancer Center, Department of Medicine and Pathology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA; 4) Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA.

Genome-wide association studies (GWAS) have implicated thousands of genes in multiple disease states, but it is often difficult to assess the impact of any given GWAS "hit". Several recent GWAS and linkage studies have associated the SNP, rs13385191, with prostate cancer (PCa). Moreover, this SNP is suggested to be a cis-acting expressed quantitative locus (eQTL), which down-regulates the expression of a poorly annotated gene known as *C2orf43*. These data parallel findings in a human subject, enrolled in the Developmental Genome Anatomy Project (DGAP, www.dgap.harvard.edu) and referred to as DGAP056, who has a chromosomal translocation that disrupts *C2orf43*, resulting in decreased expression of this gene, and who developed early-onset PCa (age 38). Analysis of RNA expression in human primary and metastatic tumors, from the Memorial Sloan Kettering Cancer Center (MSKCC) prostate repository, shows that *C2orf43* is one of the most frequently down-regulated genes in PCa and is even further reduced in metastatic prostate tumors as compared to primary prostate tumors. Importantly, *C2orf43* knockout (KO) mice show both more severe and an increased rate (>3-fold) of prostate tumors than their wild-type (WT) littermates, substantiating that *C2orf43* down-regulation is causative for prostate tumorigenesis. Accompanied with recent findings that C2orf43 is involved in intracellular cholesterol ester (iCE) metabolism and that elevated iCE is a common finding in PCa, these convergent data suggest that *C2orf43* is central to a relatively unknown PCa pathway. This type of convergent genomic analysis demonstrates a useful and successful approach for both prioritization and validation of the pathological consequences of GWAS "hits" and helps further characterize their underlying biology.

## 37

**Alignment to an Ancestry Specific Reference Genome Discovers Additional Variants Among 1000 Genomes ASW Cohort.** *R.A. Neff, J. Vargas, G.H. Gibbons, A.R. Davis.* Cardiovascular Disease Section, GMCID, National Human Genome Research Institute, Bethesda, MD.

Whole genome sequencing studies across certain populations, such as those with African ancestry, are often underpowered due to a larger divergence between the common reference genome and the true genetic sequence of the population. However, a common reference genome is not designed to account for this divergence in population-specific studies. Strong signals from common (MAF>50%) single nucleotide polymorphisms (SNPs), insertion-deletions (indels), and structural variants (SVs) can make alignment and variant calling difficult by masking nearby variants with weaker genetic signals. We present the results generated from alignment to an African descent population-specific reference genome by applying variants present in a majority of individuals with African descent from all phases of the 1000 Genomes Project and the International HapMap Consortium. We identified 882,826 single nucleotide polymorphisms, short insertion-deletion events, and large structural variations present at MAF>50%; in the population, representing 2.39 MB of genetic variation changed from hg19. We demonstrate that utilization of a population-specific reference improves variant call quality, coverage level, and imputation accuracy. We compared alignment of 27 African-American SW population (ASW) samples from the 1000 Genomes Phase 1 project between the population-specific and the hg19 reference. We discovered an additional 443,036 SNPs by alignment to the population specific reference in union across all samples, including thousands of exonic variants that are non-synonymous and are clinically relevant to the study of disease.

## 38

**Detecting novel sequence insertions in 3000 individuals from short read sequencing data.** *B. Kehr[1], P. Melsted[1,2], A. Jónasdóttir[1], A. Jónasdóttir[1], A. Sigurðsson[1], A. Gylfason[1], D. Guðbjartsson[1,3], B.V. Halldórsson[1,4], K. Stefánsson[1,5].* 1) deCODE genetics/Amgen Inc., Reykjavík, Iceland; 2) Mechanical Engineering and Computer Science, University of Iceland, Reykjavík, Iceland; 3) School of Engineering and Natural Sciences, University of Iceland, Reykjavík, Iceland; 4) Institute of Biomedical and Neural Engineering, Reykjavík University, Reykjavík, Iceland; 5) Faculty of Medicine, University of Iceland, Reykjavík, Iceland.

The detection of genomic structural variation (SV) has advanced tremendously in recent years due to progress in high-throughput sequencing technologies. Novel sequence insertions, sequences without similarity to a human reference genome, have received less attention than other types of SVs due to the computational challenges in their detection from short read sequencing data. It inherently involves de novo assembly, which is not only computationally challenging, but also requires high-quality data. While the reads from a single individual may not always meet this requirement, using reads from multiple individuals can increase power to detect novel insertions. We have developed a method to accurately characterize non-reference insertions of 100 base pairs or longer on a population scale. Our input is a mapping of all read sequences and we use a standard assembly tool to generate contigs from unmapped reads. Instead of directly anchoring these contigs into the reference genome, we merge the contigs of different individuals into high-confidence sequences, improving on quality and reliability. Subsequently, we anchor the merged sequences into the reference genome using read-pair information and LD mapping, and identify insertion positions at base-pair resolution using split-reads. Finally, we genotype these variants on 3000 sequenced individuals and impute using pedigree information into 104.000 microarray genotyped individuals, with the goal of associating the presence of an insertion with a disease phenotype. By considering simultaneously the sequence reads of multiple individuals we are able to more accurately determine both the sequences of the insertions and their location. We identify 20% more insertions when considering multiple individuals simultaneously instead of considering each individual separately. We find a large number of novel insertions, varying in frequency from 0.017% to 100%. Insertions of higher frequency commonly have a close homology to a sequence present in other primate genomes, suggesting that the inserted sequence is ancestral to humans. Novel insertions are skewed towards lower frequencies with no homology to primate sequence. Our experimental validation confirms that predicted insertions have a high probability of truly being inserted.

## 39

**SNAP: fast, accurate sequence alignment enabling biological applications.** *R. Pandya[1], W. Bolosky[1], M. Zaharia[3], T. Sittler[2,5], K. Curtis[2], C. Hartl[4], A. Fox[2], S. Schenker[2], I. Stoica[2], D. Patterson[2].* 1) eScience Research Group, Microsoft Research, Redmond, WA; 2) University of California, Berkeley, CA; 3) CSAIL, Massachusetts Institute of Technology, Cambridge, MA; 4) Broad Institute of MIT and Harvard, Cambridge, MA; 5) University of California, San Francisco, CA.

We present the Scalable Nucleotide Alignment Program (SNAP), a novel and efficient alignment algorithm and software package that enables new applications in sequence analysis. We present important examples in 1) outbreak detection, 2) sample quality control, and 3) genome remapping, and describe how SNAP has been designed to make them possible. SNAP provides accuracy equivalent to the current state-of-the-art aligners (substantiated by comparing variant calls) in 1/2 to 1/30 the time. SNAP is ready for upcoming developments in sequencing technology, with improved accuracy and increasing speed on longer paired end read lengths in contrast to some other popular algorithms. SNAP accepts multiple file formats (e.g., FASTQ, SAM, and BAM). It can sort, mark duplicates, and generate an indexed BAM file, and can align a typical paired-end human genome dataset in approximately four hours on a single commodity server, or at double that speed on longer read lengths.

## 40

**Precise identification of copy number variants in whole-genome data using Median Coverage Profiles.** *G. Glusman[1], T. Farrah[1], D.E. Mauldin[1], A.B. Stittrich[1], S. Ament[1], L. Rowen[1], J.C. Roach[1], M. Brunkow[1], M. Robinson[1], A.F.A. Smit[1], R. Hubley[1], D. Bodian[2], J. Vockley[2], I. Shmulevich[1], J. Niederhuber[2], L. Hood[1].* 1) Institute for Systems Biology, Seattle, WA; 2) Inova Translational Medicine Institute, Inova Health System, Falls Church, VA.

The identification of DNA copy numbers from short-read sequencing data remains a challenge. Depth of sequencing coverage has strong sequence-specific fluctuations only partially explained by global parameters like %GC. Current analysis methods frequently misidentify structural variants, particularly hemizygous deletions in the 1-100 kb range. We developed a method that enables precise identification of copy number variants (CNVs) and rare deletions in single individual genomes, based on comparison to joint profiles derived from a large cohort of genomes. The Family Genomics group (family genomics.systemsbiology.net) at the Institute for Systems Biology and the Inova Translational Medicine Institute (www.inova.org/clinical-education-and-research/research/inova-translational-medicine-institute) are undertaking multiple collaborative projects related to understanding the genetic basis of disease. We have produced high quality (>40x) whole-genome sequence (WGS) data from over 6000 individuals, including family trios and larger pedigrees. Our collective WGS dataset serves as a superb resource for modeling systematic failures and biases in sequencing technology, deriving population statistics, and developing and testing genome analysis software. We analyzed coverage in thousands of genomes sequenced using diverse technologies and processed using many versions of analysis pipelines. We scaled each genome to its total autosomal coverage, stratified by %GC. We then constructed joint profiles characterized by the median scaled value at each position along the genome. These Median Coverage Profiles (MCPs) take into account the diverse technologies and pipeline versions. MCPs can also help identify and correct batch effects. Normalization to the MCP followed by hidden Markov model (HMM) segmentation enables very efficient and precise detection of CNVs and large deletions in individual genomes. Use of multi-genome models improves our ability to analyze each individual genome, leading to fewer false positive and false negative findings. Several of the rare deletions we identified are prime disease-causing candidates in a variety of studies. We make available MCPs, HMM parameters, population frequencies for all CNVs and tools for improving the quality of personal genome analyses, individually and in the context of family pedigrees. The increased sensitivity and specificity for individual genome analysis are crucial for achieving clinical-grade genome interpretation.

## 41

**Accurate read mapping using a graph-based human pan-genome.** *W. Lee[1,2], E. Garrison[2], D. Kural[1,2], G. Marth[2,3].* 1) Seven Bridges Genomics Inc., Cambridge, MA; 2) Department of Biology, Boston College, Chestnut Hill, MA; 3) Department of Human Genetics, the University of Utah, Salt Lake City, UT.

Current short-read mapping algorithms utilize species-specific genome reference sequences to align reads from a newly sequenced individual. Many reads fail to map or are incorrectly mapped because each new genome typically contains many genetic variations not captured by the reference sequence. As a result, while it is possible to detect SNPs and short INDEL variants using such mappings, longer/structural variant alleles and more complex variations are often missed. Furthermore, undetected structural variants in a newly genome often cause mismappings that lead to false positive variant predictions.

As lots of novel variants are discovered by high profile projects, accounting for those novel variants when aligning newly reads becomes imperative and vastly improves sensitivity. This is based on the fact that most of variants found in a single individual are shared in that species. We thus develop a novel whole-genome read mapper that can take into account known variations, in addition to the genome reference, for mapping reads more accurately. Our approach is to construct a directed acyclic graph (DAG) representing the reference sequence and the allelic alternates. Our mapper works in two phases. In a first read localization step, we identify regions where a read is likely to map in the DAG. In a second local alignment step, we align the read against the DAG, using a graph-aware extension of the Smith-Waterman optimal alignment algorithm.

We demonstrate the power of this new read mapper for the detection of mobile element insertions (MEIs) in a human sample. When constructing a DAG using known MEI sites in YRI population in the 1000 Genomes Project, we are able to detect >95% of such sites present in a simulated genome. Similar results are achieved when detecting MEIs in NA12878. Moreover, using our mappings considering known MEIs, we are able to eliminate >95% of falsely called SNPs and INDELs at or near the MEI insertion sites in traditionally mapped sequence alignments. These false positives are almost always caused by mismapping reads containing the MEI sequences in the sample but that are not present in the reference genome. Our mapper, accounting for these insertions within the DAG, is able to correctly align the reads. These initial results indicate that read mapping that accounts for known variations can substantially improve read placement and supports vast improvements in variant calling accuracy.

## 42

**The Impact of GRCh38 on Clinical Sequencing.** *D.M. Church, J. Harris, G. Bartha, M. Pratt, A. Patwardhan, S. Chervitz, S. Kirk, M. Clark, S. Garcia, J. West, R. Chen.* Personalis, Inc, Menlo Park, CA.

Our ability to analyze and interpret individual human genomes relies upon comparison to a high quality reference assembly, the guidepost upon which we identify variants and interpret sequence data. GRCh37 has been the reference assembly for over four years. During this time a number of large-scale projects such as 1000 genomes, ENCODE and GO-ESP have placed detailed annotation in the context of GRCh37. Additionally, this assembly has been a workhorse in the clinical sequencing arena and is still the standard in most clinical testing labs doing genome wide testing. In December of 2013, the Genome Reference Consortium (GRC) released an updated version of the reference assembly called GRCh38. The new assembly adds several megabases of sequence not present in GRCh37, corrects numerous misassembled regions, provides better representation of several hundred medically relevant genes and adds previously unrepresented genes. Despite the marked improvement of this assembly, clinical labs face numerous challenges transitioning to GRCh38. Annotation content needs to be added to GRCh38 before a lab can consider re-validating clinical protocols using the new assembly. Additionally, new characterization methods along the entire analysis pipeline that take advantage of the full assembly, which now includes over 170 regions with alternate sequence representations, need to be developed. Addressing these challenges, we have continued work on our variant calling and annotation pipeline. Employing a stepwise approach, we are initially developing a pipeline to take advantage of the FIX patches that the GRC releases on a quarterly basis. This allows us to start investigating some GRCh38 sequence in the larger context of GRCh37 annotation and positions us to take advantage of GRCh38 FIX patches when they are released. Preliminary data shows that these sequences improve alignments both within the patch regions and outside of the patch regions as off-target alignments are reduced globally. We are also investigating approaches that will allow us to use the alternate loci that are released as part of the full assembly. Lastly we are transitioning annotation content from GRCh37 to both the FIX patches as well to GRCh38. This process has identified regions of the new assembly that are completely devoid of biological information, and has uncovered sets of data that need to be re-evaluated in light of the new assembly.

## 43

**Optimized Exome Sequencing for Discovery Research: Improved Metrics and Methods to Enhance Variant Discovery Across the Biomedical Footprint of the Genome.** *M. Pratt, S. Luo, G. Bartha, J. Harris, N. Leng, C. Haudenschild, R. Chen, J. West.* Personalis, Menlo Park, CA., USA.

Whole exome sequencing (WES) is a broadly used technique for variant detection and discovery in a wide range of research study types, yet metrics to measure system-level sensitivity of these assays (vs. metrics such as average depth of coverage) are rarely used. We have developed alternative methods for assessing variant detection sensitivity, and have optimized an augmented exome protocol to efficiently detect variants across a large biomedical footprint by leveling coverage and targeting minimum rather than average read depth. We have compared multiple capture and sequencing protocols to determine a preferred assay configuration to efficiently discover variants in research studies across the biomedically relevant content of the genome. Using a standard sample (NA12878) run at high depth on two different exome platforms, a titration of data sets of decreasing sequencing volume were created by randomly downsampling reads. Variant detection was performed utilizing the Personalis pipeline for all assays. Variant calls were evaluated over a prioritized content set including clinical genes, genes with phenotype associations, UTRs, splice and non-coding clinical variants, GWAS variants and highly conserved loci near previously identified biomedical content. Variant detection sensitivity was determined by comparing to an internal "gold" set of variants previously characterized within this sample by repeated high-depth whole genome sequencing on multiple platforms. In addition, we utilized the NIST genome-in-a-bottle call set on a reduced footprint for confirmation. Using an empirically determined sensitivity by depth function, we assess the effective sensitivity and discovery footprint of each configuration and find an optimum. We found that the augmented exome was the most efficient assay for variant discovery at almost all levels of sequence data analyzed. We also derived the nature of the variant discovery curves for a standard v. augmented exome, and utilized this to estimate the optimal use of sequencing data across a large sample set for cost-effective variant discovery.

## 44

**SpeedSeq: A 24-hour alignment, variant calling, and genome interpretation pipeline.** *C. Chiang[1], R.M. Layer[2], G.G. Faust[1], M.R. Lindberg[1], A.R. Quinlan[1,3,4], I.M. Hall[1,3,4].* 1) Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908; 2) Computer Science, University of Virginia, Charlottesville, VA 22903; 3) Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22904; 4) Department of Public Health Sciences, University of Virginia, Charlottesville VA 22908.

Bioinformatic turn-around time is currently a major obstacle for clinical adoption of genome sequencing technologies. For many whole genome sequencing applications such as cancer genotyping or newborn diagnosis, the clinically actionable timeframe is days or weeks. While the time required to generate whole genome sequencing reads has been reduced from ~2 weeks to ~3 days, bioinformatic analysis remains a major challenge, typically requiring weeks or months to go from raw DNA sequence data to causal variants, with extensive hands-on involvement. We set out to systematically reduce bioinformatic turn-around time and simplify variant interpretation without sacrificing accuracy. To this end, we present SpeedSeq, a rapid and comprehensive pipeline for characterizing and prioritizing genetic variation in human genomes. We show that our ultra-fast pipeline produces high-quality SNV and indel calls with specificity and sensitivity on par with current standards. In a paired tumor/normal analysis, SpeedSeq achieves high recall and precision rates even for subclonal variants, and near perfect recall of orthogonally validated mutations in tumors from The Cancer Genome Atlas (TCGA). We further show that our structural variation detection approach (LUMPY) significantly outperforms other available tools, and we have validated variant detection power against available gold standards from the 1000 Genomes Project and Genome in a Bottle Consortium. Additionally, updates to the GEMINI framework can identify actionable mutations in a clinically relevant timeframe with minimal human involvement. In under 24-hours, our approach moves raw sequence data to fully processed variant calls with genetic implications. Our pipeline is composed entirely of free open source software tools including BWA, Samblaster, Sambamba, BEDTools, Freebayes, LUMPY, SnpEff, GEMINI, and DGIdb, as well as several new data processing tools designed to greatly increase speed. SpeedSeq is available on Github at https://github.com/cc2qe/speedseq.

## 45

**Genomic Sequencing Approaches Identifies Novel Rare Variants in Patients with Mendelian Neurologic Diseases.** *E. KARACA[1], D. PEHLIVAN[1], T. HAREL[1], S. Weitzer[2], H. Shiraishi[2], T. GAMBIN[1], Y. BAYRAM[1], W. Wiszniewski[1], S.N. JHANGIANI[3], G. YESIL[4], S. ISIKAY[5], O. OZALP YUREGIR[6], S. BOZDOGAN[6], H. ASLAN[6], T. TOS[7], D. GUL[8], B. YILMAZ[9], O. COGULU[9], K. KARAER[10], H. ULUCAN[11], D. Muzny[3], M. SEVEN[11], A. YUKSEL[11], T. CLAUSEN[12], T. Tuschl[13], A. HESS[14], R.A. GIBBS[1,3], J. MARTINEZ[2], J.M. PENNINGER[2], J.R. LUPSKI[1,15,16].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) IMBA, Institute of Molecular Biotechnology of the Austrian Academy of Sciences, 1030 Vienna, Austria; 3) Human Genome Sequencing Center, Baylor College of Medicine, Houston TX, USA; 4) Department of Medical Genetics, Bezmialem University, Istanbul, Turkey; 5) Gaziantep Childreǹs Hospital, Gaziantep, Turkey; 6) Department of Medical Genetics, Numune Training and Research Hospital, Adana, Turkey; 7) Department of Medical Genetics, Sami Ulus Children's Hospital, Ankara, Turkey; 8) Department of Medical Genetics, Gülhane Military Medical School, Ankara, Turkey; 9) Department of Medical Genetics, Ege University, Faculty of Medicine, 35100, Izmir Turkey; 10) Intergen Genetic Center, Ankara, Turkey; 11) Department of Medical Genetics, Cerrahpasa Medical School of Istanbul University, Istanbul, Turkey; 12) IMP, Institute of Molecular Pathology, 1030 Vienna; 13) Howard Hughes Medical Institute, Laboratory of RNA Molecular Biology, Rockefeller University, New York, NY 10065, USA; 14) Institute for Experimental Pharmacology Friedrich-Alexander University Erlangen-Nuremberg, Germany, and Campus Support Facility (CSF), Vienna BioCentre, Vienna; 15) Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA; 16) Texas Children's Hospital, Houston, TX, USA.

Development of the human nervous system involves a variety of complex interactions between fundamental cellular processes including proliferation, differentiation, migration and apoptosis. The advent of next generation sequencing has enabled rapid identification of numerous genes and mechanisms that contribute to human neurogenesis. However, considering the magnitude and complexity of the mechanisms involved, the search for genes underlying Mendelian neurological disease is far from complete. We applied whole exome sequencing (WES) to a cohort of 114 Turkish patients with congenital brain malformations from 82 mostly consanguineous families. In 40 patients (35%), WES revealed deleterious mutations in known genes. An additional 14 patients (12%) harbored deleterious mutations in known genes but were noted to have distinctive clinical features when compared to the available literature, thus representing phenotypic expansion. Potential disease causing variants were identified in novel candidate genes in the remaining 60 patients (52%) and segregated with the phenotype in available family members. Among the genes identified, we found a homozygous CLP1 rare single variant (c.G419A; pR240H) in 7 unrelated families. CLP1 is a RNA kinase involved in tRNA splicing. We recently showed that this homozygous missense mutation leads to a loss of CLP1 interaction with the tRNA splicing endonuclease (TSEN) complex, largely reducing pre-tRNA cleavage activity and resulting in accumulation of linear tRNA introns. Furthermore, we showed that mice carrying kinase-dead CLP1 displayed microcephaly and reduced cortical brain volume due to the enhanced cell death of neuronal progenitors that is associated with reduced numbers of cortical neurons. The use of genomic sequencing approaches, along with other genome-wide interrogation technologies, is invaluable in the identification of causative rare variants in families segregating Mendelian disease traits. We expect that these will provide novel insights into human nervous system development and human biology, as well as human neurologic disease.

## 46

**Individualized iterative phenotyping for genome-wide analysis of loss of function mutations.** *J.J. Johnston[1], K. Lewis[1], D. Ng[1], L.N. Singh[1], J. Wynter[1], C. Brewer[2], B.P. Brooks[3], I. Brownell[4], F. Candotti[1], S.G. Gonsalves[1], P.S. Hart[1], H.H. Kong[4], K.I. Rother[5], R. Sokolic[1], B.D. Solomon[1], W.M. Zein[3], D.N. Cooper[6], P.D. Stenson[6], J.C. Mullikin[1,7], L.G. Biesecker[1,7].* 1) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; 2) National Institute on Deafness and other Communicative Disorders, National Institutes of Health, Bethesda, MD, USA; 3) National Eye Institute, National Institutes of Health, Bethesda, MD, USA; 4) National Cancer Institute, National Institutes of Health, Bethesda, MD, USA; 5) National Institute of Diabetes and Digestive Diseases, National Institutes of Health, Bethesda, MD, USA; 6) Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, United Kingdom; 7) NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, MD, USA.

Putative loss of function (pLOF) variants are common in genomes and it is critical for understanding gene function and predictive medicine to assess the consequences of pLOF variants. Towards this end, we characterized the consequences of pLOF variants in a largely healthy adult exome cohort by iterative phenotyping. Exome data were generated on 951 participants from the ClinSeq® cohort and filtered for pLOF variants likely to cause a phenotype in heterozygotes. 106 of 951 exomes had such a pLOF variant and 79 participants were available for evaluation. Of those 79, 38 had positive findings or family histories (31 variants in 19 genes), two had indeterminate findings (two variants, one each in two genes) and 39 had negative findings or a negative family history (32 variants in 27 genes) for that trait. We correlated these findings with the haploinsufficiency score of the affected gene when available, the Combined Annotation-Dependent Depletion (CADD) score of the variant, and a novel variant composite pathogenicity score. The composite pathogenicity score was calculated based on the following seven attributes of each variant: whether the variant predicted a frameshift with >10 aberrant amino acids, whether MutationTaster predicted nonsense-mediated decay, whether ≥5 pLOF mutations were present for the gene in the Human Gene Mutation Database (HGMD), whether there was a mutation in that exon in the public version of HGMD, whether that specific mutation was in HGMD as a disease causing (DM) mutation, whether the pLOF was in the middle 90% of the gene (i.e., not in the first or last 5%), and whether the exon with the pLOF variant was an exon in the dominant gene model as defined by presence in >75% of overlapping spliced ESTs. Using these measures we developed an algorithm to classify variants. We conclude that 1/30 unselected individuals had a manifest phenotype attributable to rare LOF variants, which is more common than may be assumed.

## 47

**Genomic approach identifies novel proteins necessary for inner ear function and development across multiple species.** *O. Diaz-Horta[1], M. Grati[2], C. Abad[1], A. DeSmidt[3], G. Bademci[1], A. Subasioglu-Uzak[4], J. Foster II[1], S. Tokgoz-Yilmaz[4], D. Duman[4], F.B. Cengiz[4], S.H. Blanton[1], X.Z. Liu[2], A. Farooq[5], Z. Lu[3], K. Walz[1], M. Tekin[1].* 1) University of Miami Miller School of Medicine John P. Hussman Institute for Human Genomics, Miami, FL; 2) Department of Otolaryngology, Miller School of Medicine, University of Miami, Miami, FL 33136, USA; 3) Department of Biology, University of Miami, Miami, FL, 33146, USA; 4) Division of Pediatric Genetics, Ankara University School of Medicine, Ankara, 06100, Turkey; 5) Department of Biochemistry and Molecular Biology, Miller School of Medicine, University of Miami, Miami, FL 33136, USA.

The concerted action of thousands of proteins is required for inner ear development and function. Many of these proteins are currently unknown. In this study, we used a genomic approach to identify novel components of inner ear development and function. We present two such proteins that have not been previously recognized to be involved in hearing. The first protein was identified through studying a large Turkish kindred with autosomal recessive non-syndromic deafness. Whole exome sequencing (WES) in this family detected a splice site mutation (c.102-1G>A) in FAM65B (MIM611410) that co-segregates with the phenotype and is absent in 330 ethnicity-matched controls. The mutation leads to skipping of an exon and deletion of 52 amino acid residues in a membrane localization domain. We show that wild type Fam65b is expressed during embryonic and postnatal developmental stages in murine cochlea, and that the protein localizes to the plasma membranes of the stereocilia of inner and outer hair cells. The wild type protein targets the plasma membrane, whereas the mutant protein accumulates in cytoplasmic inclusion bodies and does not reach the membrane. In zebrafish, knockdown of fam65b leads to significant reduction in the number of saccular hair cells and neuromasts, and to hearing loss. We conclude that FAM65B is a plasma membrane-associated protein of hair cell stereocilia that is essential for hearing. The second protein was identified through studying another multiplex Turkish family with autosomal recessive non-syndromic deafness. Affected individuals had common cavity inner ear anomaly. Autozygosity mapping and WES identified a missense mutation in a gene encoding a tyrosine kinase-like receptor. In vitro studies showed that the mutant protein is unable to reach the plasma membrane, the natural localization of the wild type receptor. In zebrafish, knockdown of the orthologous gene with splice-blocking morpholinos led to anatomic changes in the ear and to hearing loss. Investigations of knock-out mouse demonstrate profound deafness along with cochlear anomalies. We will disclose the name of the second gene at the meeting. We conclude that genomic studies along with animal models are effective strategies to characterize the remaining proteins that are necessary for hearing.

## 48

**A *Drosophila* genetic resource to study human disease genes and its use for gene discovery in human exome data.** *M.F. Wangler[1,2], S. Yamamoto[1,3,4], M. Jaiswal[1,5], W.L. Charng[1,4], T. Gambin[1,6], E. Karaca[1], G. Mirzaa[7,8], W. Wiszniewski[1,2], H. Sandoval[1], N. Haelterman[4], V. Bayat[4], D. Pehlivan[1], S. Penney[1,2], L. Vissers[10], S. Jhangiani[11], S. Tsang[12,13], Y. Xie[12], Y. Parman[14], E. Battaloglu[15], D. Muzny[1,11], Z. Liu[3,16], R. Clark[17], C. Curry[18], E. Boerwinkle[11,19], W. Dobyns[7,8,20], R. Allikmets[12,13], R. Gibbs[1,11], R. Chen[1,4,11], J.R. Lupski[1,2,11,16], H. Bellen[1,2,3,4,9,21].* 1) Department of Molecular and Human Genetics, BCM, Houston, TX, 77030; 2) Texas Children's Hospital, Houston, TX, 77030; 3) Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital (TCH), Houston, TX, 77030; 4) Program in Developmental Biology, Baylor College of Medicine (BCM), Houston, TX, 77030; 5) Howard Hughes Medical Institute, Houston, TX, 77030; 6) Institute of Computer Science, Warsaw University of Technology, 00-661 Warsaw, Poland; 7) Department of Pediatrics, University of Washington, Seattle, WA, 98195; 8) Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, 98101; 9) Program in Structural and Computational Biology and Molecular Biophysics, BCM, Houston, TX, 77030; 10) Department of Human Genetics, Radboudumc, PO Box 9101, 6500 HB, Nijmegen, The Netherlands; 11) Human Genome Sequencing Center, BCM, Houston, TX, 77030; 12) Department of Ophthalmology, Columbia University College of Physicians and Surgeons, New York, NY, 10032; 13) Department of Pathology & Cell Biology, Columbia University College of Physicians and Surgeons, New York, NY, 10032; 14) Neurology Department and Neuropathology Laboratory, Istanbul University Medical School, Istanbul, Turkey; 15) Department of Molecular Biology and Genetics, Bogazici University, Istanbul, Turkey; 16) Department of Pediatrics, Baylor College of Medicine, Houston, TX, 77030; 17) Division of Medical Genetics, Department of Pediatrics, Loma Linda University Medical Center, Loma Linda, CA, 92354; 18) Department of Pediatrics, University of California San Francisco, San Francisco, CA, 94143, and Genetic Medicine Central California, Fresno, CA, 93701Division of Medical Genetics, Department of Pediatrics, Loma Linda University Medical Center, Loma Linda,; 19) Human Genetics Center, University of Texas, Health Science Center at Houston, Houston, TX, 77030; 20) Department of Neurology, University of Washington, 98195; 21) Department of Neuroscience, BCM, Houston, TX, 77030Department of Neurology, University of Washington, 98195.

Exploring the genetic mechanisms of disease in model organisms such as *Drosophila* has traditionally relied primarily upon gene discovery in humans followed by reverse genetic studies of the disease gene in the model system. While forward genetic screens in *Drosophila* have provided numerous fundamental contributions to basic biology, technical limitations have made it more difficult to quickly use these screens for disease discovery. The widespread use of exome sequencing now allows for more rapid application of *Drosophila* screens to personal genomes of patients with rare disease. In order to provide functional information about many genes, we conducted a large forward-genetic mutagenesis screen of the *Drosophila X*-chromosome and selected lethal mutations. Numerous phenotypes were tested in these lethal stocks to capture genes required for development, function, and maintenance of the nervous system. We mapped, and rescued the phenotypes associated with mutations in 165 fly genes. We then undertook a systematic study of the 250 human homologues. We observed that the human homologues of lethal fly genes were enriched for association with Mendelian disease compared to the whole genome. In addition we made a surprising observation, that genes that are 1) essential in flies and 2) have multiple human homologs are the most likely to be associated with Mendelian disease. We then undertook a systematic study of these human homologues within 1,929 human exomes from families with unsolved rare disease from the Baylor Hopkins Center for Mendelian Genomics study. We extracted all the variants in these genes under dominant and recessive models, and by sequencing family members we identified disease-associated, co-segregating mutations in six independent families. These families provided some examples of findings consistent with what has been previously reported (*DNM2* and Charcot-Marie Tooth disease), examples of phenotypic expansion for a known disease gene (*CRX* and Bull's Eye Maculopathy), and potential novel disease gene discoveries, namely mutations in *ANKLE2* associated with severe microcephaly. This latter family and the resulting *Drosophila* and human studies are described in additional abstracts from this project (see Charng et al, and Clark et al). Our approach provides an example of the use of forward genetic screens in *Drosophila* and human genomic data in order to facilitate gene discovery and unbiased functional studies of Mendelian disease.

## 49

**Genome sequencing identifies major causes of severe intellectual disability.** *C. Gilissen[1], J.Y. Hehir-Kwa[1], D. Thung[1], M. van de Vorst[1], B.W.M. van Bon[1], M.H. Willemsen[1], M. Kwint[1], I.M. Janssen[1], A. Hoischen[1], A. Schenck[1], R. Leach[2], R. Klein[2], R. Tearle[2], T. Bo[1,3], R. Pfundt[1], H.G. Yntema[1], B.B.A. de Vries[1], T. Kleefstra[1], H.G. Brunner[1,4], L.E.L.M. Vissers[1], J.A. Veltman[1,4].* 1) Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands; 2) Complete Genomics Inc., Mountain View, CA, USA; 3) State Key Laboratory of Medical Genetics, Central South University, Changsha, Hunan, China; 4) Department of Clinical Genetics, Maastricht University Medical Centre, Maastricht, The Netherlands.

Severe intellectual disability (ID) occurs in 0.5% of newborns and is thought to be largely genetic in origin. The extensive genetic heterogeneity of the disorder requires a genome wide detection of all types of genetic variation. Microarray studies and more recently exome sequencing have demonstrated the importance of *de novo* copy number variations (CNVs) and single nucleotide variations (SNVs) in ID, but the majority of cases remains undiagnosed. Here we applied whole genome sequencing (WGS) to 50 patients with severe ID and their unaffected parents. All patients were negative after extensive genetic prescreening, including microarray-based CNV studies and exome sequencing. Notwithstanding this prescreening, *de novo* SNVs affecting the coding region provided a conclusive genetic cause in 13 patients and a possible cause for another 8 patients. In addition, we identified 7 clinically relevant *de novo* CNVs as well as one recessively inherited compound heterozygous CNV. These CNVs included single exon and intraexonic deletions of known ID genes as well as interchromosomal duplications. Local realignment of sequence reads allowed the mapping of most of these CNVs at single nucleotide resolution level and provided positional information for duplicated sequences. These results show that *de novo* mutations and CNVs affecting the coding region are the major cause of severe ID. Genome sequencing can be applied as a single genetic test to reliably identify and characterize the comprehensive spectrum of genetic variation, providing a genetic diagnosis in the majority of patients with severe ID.

## 50

**Assessment of the success rate of two years of large-scale exome sequencing efforts to identify genes for Mendelian conditions at the University of Washington Center for Mendelian Genomics.** *J.X. Chong[1], J. Shendure[2], D.A. Nickerson[2], M.J. Bamshad[1,2,3], University of Washington Center for Mendelian Genomics.* 1) Department of Pediatrics, University of Washington, Seattle, WA; 2) Department of Genome Sciences, University of Washington, Seattle, WA; 3) Seattle Children's Hospital, Seattle, WA.

To evaluate clinical services and large-scale gene discovery using exome sequencing (ES), it is important to apply objective metrics to assess the success of finding causal genes for Mendelian conditions (i.e. solve rates). To date, reported solve rates are hard to interpret and use for comparisons across different contexts (i.e. clinical service vs. research). Because of its simplicity, "solve rate" has been reported as the proportion of families in which a causal variant for a Mendelian condition is identified or alternatively as the proportion of phenotypes for which the gene is identified. The former definition isn't particularly useful on its own as one could have a high solve rate by only sequencing families diagnosed with disorders with a single known gene, while the latter has multiple interpretations because the same phenotype can be caused by variants in multiple genes, variants in a single gene can cause multiple phenotypes, and causal variants often cannot be identified in all families. We developed three complementary metrics for "solve rate" and applied them to >600 families and >200 phenotypes studied at the University of Washington Center for Mendelian Genomics using strict criteria for variant causality and clear definitions of phenotype and gene novelty. The overall diagnostic rate, defined as the proportion of families for which a causal variant was identified, is 40%, while the diagnostic rate for the subset of families with causal variants in a gene previously known to cause their condition is 23%. This is comparable to diagnostic rates achieved by clinical ES even though the conditions we studied are biased against phenotypes expected to be explained by a known Mendelian disease gene. The gene identification rate, defined as the ratio of causal gene identifications to phenotypes, ranged from 0.56 (dominant) to 1 (X-linked) depending on inheritance model; for comparison, if a causal gene was identified for every phenotype, this ratio would approach its maximum value of one. Gene identifications within consanguineous families were frequently complicated by locus heterogeneity consistent with a lower than anticipated solve rate (0.81). Lastly, the novel discovery rate, or proportion of gene identifications in which the gene was newly discovered to underlie a Mendelian condition or the condition itself was novel/unexplained was 49%. Solve rates were further used to determine aspects of experimental design shared by successful projects.

## 51

**A comparative analysis of allele frequencies for incidental findings among five populations based on the analyses of 11K whole exome sequences.** *T. Gambin[1], S. Jhangiani[2], J.E. Below[3], J. Staples[6], A. Morrison[3], A. Li[3], I. Campbell[1], W. Wiszniewski[1], D.M. Muzny[2], M.N. Bainbridge[2], R.A. Gibbs[2], J.R. Lupski[1,2,4,5], E. Boerwinkle[2,3].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) The Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX; 4) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 5) Texas Children's Hospital, Houston, TX; 6) Department of genome sciences, University of Washington, Seattle,WA.

Whole exome sequencing is growing to become a routine diagnostic test in many settings, and some predict that an individual's sequence will be a routine element in the electronic medical record. Incidental findings are potential medically actionable variants detected in the results of the exome or genome sequencing that are not etiologically related to a diagnostic indication for which the genome sequencing test was ordered. Such sequence information also provides an easily accessible carrier test for autosomal recessive diseases and can reveal potential pharmacogenetic risk variant alleles. The American College of Medical Genetics and Genomics (ACMG) has identified a list of 56 genes for which pathogenic or likely pathogenic variants should be identified as incidental findings. We identified non-synonymous rare variants within the ACMG genes in 2,347 individual whole exomes sequenced in Houston at the Baylor-Johns Hopkins Center for Mendelian Genomics and in 8,602 exomes from the Atherosclerosis Risk In Communities (ARIC) study. Common variants were excluded if MAF was >1% in 1KG or EVS databases. We identified a total of 23,092 mutations (6,135 distinct) in 56 genes. The number of mutations in an individual ranged from 0 to 11 with an average of 2.1 for any given individual. Approximately 22% of all mutations were previously reported in HGMD as disease-causing mutations (DM;4702) or disease-associated polymorphisms (DP/DFP;461). Moreover, we determined the average number of HGMD non-synonymous (NS) and stop-gains (SG) among the following ethnic groups: Europeans (NS=0.46;SG=0.018), Africans (NS=0.51;SG=0.027), Mexicans (NS= 0.45;SG=0.003), Turkish (NS=0.42;SG=0.021) and Asians (NS=0.45;SG= 0). In addition to the analysis of incidental findings, we used the ARIC cohort to evaluate the frequency of 8 pharmacogenetic risk alleles that are being currently screened at the Whole Genome Laboratory at Baylor as a part of routine WES data analysis. We found an average of 0.5 mutations per individual in Europeans vs. 0.38 in Africans. Last but not least, we investigated the carrier frequency for 1602 known recessive genes. We determined that every individual is a carrier of 0-14 (mean=4.38) likely pathogenic variants. Our results reaffirms that analysis of WES data not only helps in the clinical diagnosis and facilitates discoveries of novel genes but also provides multifaceted insight into clinically relevant genetic information for individuals and populations.

## 52

**Why Next-Generation Sequencing Studies May Fail: Challenges and Solutions for Gene Identification in the Presence of Familial Locus Heterogeneity.** *R.L.P. Santos-Cortez[1], A.U. Rehman[2], M.C. Drummond[2], M. Shahzad[3,4], K. Lee[1], R.J. Morell[2], M. Ansar[1,5], A. Jan[5], X. Wang[1], A. Aziz[5], S. Riazuddin[3,6], J.D. Smith[7], G.T. Wang[1], Z.M. Ahmed[4,6], K. Gul[3], A.E. Shearer[8], R.J.H. Smith[8], J. Shendure[7], M.J. Bamshad[7], D.A. Nickerson[7], J. Hinnant[9], S.N. Khan[3], R.A. Fisher[10], W. Ahmad[5], K.H. Friderici[10,11], S. Riazuddin[3,12,13], T.B. Friedman[2], E.S. Wilch[11], S.M. Leal[1], University of Washington Center for Mendelian Genomics.* 1) Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA; 2) Laboratory of Molecular Genetics, National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Rockville, Maryland 20850, USA; 3) National Center of Excellence in Molecular Biology, University of the Punjab, Lahore 54590, Pakistan; 4) Division of Pediatric Ophthalmology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio 45229, USA; 5) Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan; 6) Laboratory of Molecular Genetics, Division of Pediatric Otolaryngology Head and Neck Surgery, Cincinnati Children's Hospital Medical Center, Cincinnati 45229, Ohio, USA; 7) Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; 8) Molecular Otolaryngology & Renal Research Labs, Department of Otolaryngology—Head & Neck Surgery, University of Iowa, Iowa City, Iowa 52242, USA; 9) Department of Religious Studies, Michigan State University, East Lansing, Michigan 48824, USA; 10) Department of Pediatrics and Human Development, Michigan State University, East Lansing, Michigan 48824, USA; 11) Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA; 12) Allama Iqbal Medical College-Jinnah Hospital Complex, University of Health Sciences, Lahore 54550, Pakistan; 13) University of Lahore, Lahore 54000, Pakistan.

Next-generation sequencing (NGS) of exomes and genomes has accelerated the identification of genes involved in Mendelian phenotypes. However, many NGS studies fail to identify causal variants, with estimates for success rates of uncovering the pathologic variant underlying disease etiology being as low as 25 percent (Yang et al. 2013). An important reason for such failures is familial locus heterogeneity, where causal variants in two or more genes within a single pedigree underlie Mendelian trait etiology. As examples of intra- and inter-sibship familial locus heterogeneity, we present 10 consanguineous Pakistani families segregating hearing impairment due to homozygous mutations in two different hearing impairment genes and a large European-American pedigree in which hearing impairment is caused by pathogenic variants in three different genes. We have identified 41 additional pedigrees with syndromic and nonsyndromic hearing impairment for which a single known hearing impairment gene has been identified but only segregates with the phenotype in a subset of affected pedigree members. We estimate that locus heterogeneity occurs in 15.3 percent (95 percent confidence interval: 11.9 to 19.9 percent) of the families in our collection where we have identified at least one variant in a previously published hearing impairment gene which only segregates with hearing impairment phenotype in a subset of affected pedigree members. These families have been evaluated by screening genes which commonly underlie nonsyndromic hearing impairment, whole-genome linkage analysis and NGS. We demonstrate novel approaches to apply linkage analysis and homozygosity mapping (for autosomal recessive consanguineous pedigrees) which can be used to detect locus heterogeneity using either NGS or SNP array data. Results from the linkage analysis and homozygosity mapping can also be used to group sibships or individuals most likely to be segregating the same causal variants and thereby aid in gene identification. If the analysis is performed using SNP genotyping arrays, before sequencing the results can be used to aid in the selection of pedigree members for NGS. It is demonstrated how the novel applications of linkage analysis and homozygosity mapping can increase the success rate of gene identification for families with locus heterogeneity.

## 53

**Genome-wide association study imputed to 1000 Genomes reveals 18 novel associations with type 2 diabetes.** *R.A. Scott[1], R. Magi[2], A.P. Morris[3], L. Marullo[4], K. Gaulton[5], M. Boehnke[6], J. Dupuis[7], M.I. McCarthy[5], L.J. Scott[6], I. Prokopenko[8], DIAGRAM consortium.* 1) MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK; 2) Estonian Genome Center, University of Tartu, Tartu, Estonia; 3) Department of Biostatistics, University of Liverpool, Liverpool, UK; 4) Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy; 5) University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK; 6) Center for Statistical Genetics and Dept of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA; 7) Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, 02118, USA; 8) Genomics of Common Disease, Imperial College London, London, UK.

Genome-wide association studies (GWAS) imputed to Hapmap reference panels have identified over 70 loci showing associations with type 2 diabetes (T2D). Associated variants identified by GWAS to date are exclusively common (MAF>5%). The 1000 Genomes (1000G) reference panel is a comprehensive catalogue of variation down to 1% MAF and allows imputation of a larger number of variants, including many at lower frequency. Here, we aimed to identify new low frequency and common variant associations with T2D by 1000G imputation. We performed an inverse-variance weighted meta-analysis of genome-wide data from up to 26,676 individuals with T2D and 132,532 controls of European ancestry from 18 studies. Genetic data were imputed using the March 2012 1000G multi-ethnic reference panel. Following discovery analyses, we sought follow-up for single nucleotide variants (SNVs) reaching 5×10-5 in up to 14,545 cases and 38,994 controls of European ancestry genotyped on the Metabochip. We also defined credible sets of SNVs in a 1Mb window that were 99% likely to contain the causal variant. Analysis of ~12M SNPs revealed 18 loci showing new associations with T2D (p<5×10-8). All lead SNVs were common (MAF>10%). Seven loci contained genome-wide significant associations in discovery analyses, including variants in or near *CENPW* (chr6:126792095; OR [95% CI]= 1.10[1.06,1.13]), *HSF1* (chr8:145536056; 1.08[1.05,1.11]), *PLEKHA1* (chr10:124186714; 1.09[1.06,1.11]), *HSD17B12* (chr11:43877934; 1.08[1.05,1.11]), *CMIP* (chr16:81534790; 1.08[1.05,1.10]), *APOE* (chr19:45411941; 1.13[1.09,1.17]) and *HORMAD2* (chr22:30599562; 1.13[1.09,1.18]). Eleven further signals were seen after follow-up in or near *GLP2R*, *ATP5G1*, *NHEG1*, *MAP3K11*, *ACSL1*, *ABO*, *TCF19*, *HLA-DQ1*, *UBE3C*, *NRXN3*, and *ZZEF1*. In credible-set mapping of 48 known loci, 17 contained fewer than 20 SNVs in the 99% credible set. We identified seven loci for which the previous Hapmap lead SNV was not included in the 99% credible set, including *BCAR1*, where the 1000G-lead SNV (OR=1.16, p= 3.7×10-11) had a stronger association than the previous HapMap lead (OR= 1.1, p=1.9×10-6). Imputation using 1000G reference panel has allowed identification of 18 novel associations with T2D and identified new lead SNVs at known T2D loci, including some for which the previous Hapmap-based lead SNV is no longer in the credible set. Further annotation analyses in these regions will inform the extent to which 1000G imputation can aid in fine-mapping T2D associations.

## 54

**Genome wide association and exome sequence data analysis for more than 100 traits in Mexican Americans.** *J.E. Below[1], B.E. Cade[3], D. Aguilar[2], E. Brown[1], H.M. Highland[1], S. Redline[3,4], G.I. Bell[5,6], N.J. Cox[5,6], C.L. Hanis[1].* 1) Epidemiology, Human Genetics & Environmental Sciences, University of Texas Health Science Center, Houston, TX; 2) Department of Medicine, Baylor College of Medicine, Houston, TX; 3) Division of Sleep Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 4) Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA; 5) Department of Medicine, The University of Chicago, Chicago, IL; 6) Department of Human Genetics, The University of Chicago, Chicago, IL.

As incidence of type 2 diabetes continues to rise in the United States unabated, the genetic epidemiology at play in Starr County Texas, where trends in the rates of type 2 diabetes and obesity have stayed approximately 30 years ahead of rest of the country, is of unprecedented relevance. We have now applied the fruits of decades of phenotypic measurements in a massive data analysis of more than 1400 Mexican Americans. In addition to more than 37 million tests of variation imputed from the latest release of 1000 genomes, we examined the complete exome sequence data, generated through our participation in the T2D-GENES consortium (U01 DK085501). We have identified more than 40 independent genome wide significant findings in traits ranging from uncorrelated to highly intertwined (Pearson's correlation: 0-0.98, median = .11). Together, these capture a much more complete picture of type 2 diabetes: from subclinical measures to a broad spectrum of untoward effects of impaired glucose control and comorbidities. Top findings ranged from measures of blood glucose control (rs190455070: p-value $2\times10^{-9}$, rs181520960: $3\times10^{-14}$) to biomarkers of cardiovascular health (rs74769851: $1\times10^{-9}$ , *LGAL3*: $<10^{-20}$) and measures of quality and duration of sleep (rs80125356: $2\times10^{-16}$, rs78640598: $5\times10^{-8}$, rs12424863: $1\times10^{-9}$). We present findings from a number of traits never before genetically analyzed, including more than a dozen characteristics of infectious disease. Together these findings represent a concert of specific genetic and pleiotropic effects influencing a myriad set of immunological, inflammatory, cardiovascular, ocular, metabolic, and sleep related traits in a population undergoing an epidemic of obesity and diabetes.

## 55

**Three common recessive variations explain more than 20% of all cases of type 2 diabetes in Greenland.** *A. Albrechtsen[1], I. Moltke[2], M.E. Jørgensen[3], P. Bjerregaard[4], E.V.R. Appel[5], R. Nielsen[6], O. Pedersen[5], N. Grarup[5], T. Hansen[5].* 1) The Bioinformatics Centre, University, Copenhagen, Copenhagen, Denmark; 2) Department of Human Genetics, University of Chicago, Chicago, IL, USA; 3) Steno Diabetes Center, Gentofte, Denmark; 4) National Institute of Public Health, University of Southern Denmark, Copenhagen, Denmark; 5) The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; 6) Department of Integrative Biology, University of California, Berkeley, CA, USA.

Background and aims: We recently identified a common genetic variant in TBC1D4, which strongly influence plasma glucose levels. This variant also explained more than 10% of all cases of type 2 diabetes in Greenland and showed a recessive inheritance pattern. Motivated by this finding we analyzed Illumina metaboChip data for association to type 2 diabetes in Greenland applying a recessive model. Methods and Results: We performed GWAS and subsequent meta-analysis of two Greenlandic cohorts consisting in total of 3000 individuals that were either normal glucose tolerant or had type 2 diabetes. Association testing was done using a linear mixed model to control for admixture and relatedness. Besides the TBC1D4 locus, we identified two additional variations that under a recessive model are associated to type 2 diabetes ($P < 5\times10\text{-}8$ ). They are located on 22q12.3 and 5q1.2 far from known metabolic loci. Both variations are so common in the Inuit ancestry (MAF: 19.2% & 33.4%) that they have a strong impact under a recessive model. In contrast, they are not as common in Europeans (MAF: 7.3% & 3.3%) and therefore do not show a large recessive impact on the population. The type 2 diabetes prevalence in the Greenlandic cohorts is 10.1%, however homozygous carriers of the two identified variants have 30.1% and 24.0% risk of type 2 diabetes, respectively. Homozygous carriers of the previously identified TBC1D4 variant have a risk of 54.3%. A little more than 10% of the Greenlandic population is homozygous carriers of at least one of the 3 variants, which increases their probability of having type 2 diabetes to 25.8% compared to a risk of 8.5%. This corresponds to a population attributable risk of 24.1%. Conclusion: We have identified two additional recessive variants, each with a large impact on type 2 diabetes risk in Greenland that would not have been found applying an additive model. Together, the two identified variants and the previously identified TBC1D4 variant explains more than 20% of all cases of type 2 diabetes in the Greenlandic population. These findings illustrate that it can be valuable to also apply a recessive model when performing GWAS, especially in historically small and isolated founder populations, like the Greenlandic.

## 56

**A low frequency *AKT2* coding variant enriched in the Finnish population is associated with fasting insulin levels.** *A.K. Manning[1,2,3], H.H. Highland[4], X. Sim[5], N. Grarup[6], T. Tukiainen[1,7,8], J. Gasser[1], A. Mahajan[9], M.A. Rivas[9], A.E. Locke[5], J. Tuomilehto[10, 11, 12,13, 14], M. Laakso[15, 16], S. Ripatti[7,17,18], J.B. Meigs[19, 20], D. Altshuler[1,3,21,20,22,2], M. Boehnke[5], M.I. McCarthy[9,23,24], A.L. Gloyn[23, 24], C.M. Lindgren[9, 1], T2D Genes, GoT2D.* 1) Medical and Population Genetics Program, Broad Institute, Cambridge, Massachusetts, USA; 2) Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA; 3) Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA; 4) Human Genetics Center, The University of Texas Graduate School of Biomedical Sciences at Houston, The University of Texas Health Science Center at Houston, Houston, Texas, USA; 5) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA; 6) The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; 7) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; 8) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA; 9) Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK; 10) Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia; 11) Instituto de Investigacion Sanitaria del Hospital Universario LaPaz (IdiPAZ), University; 12) Hospital LaPaz, Autonomous University of Madrid, Madrid, Spain; 13) Center for Vascular Prevention, Danube University Krems, Krems, Austria; 14) Diabetes Prevention Unit, National Institute for Health and Welfare, Helsinki, Finland; 15) Department of Medicine, University of Eastern Finland; 16) Kuopio University Hospital, Kuopio, Finland; 17) Department of Public Health, Hjelt Institute, University of Helsinki, Helsinki, Finland; 18) Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; 19) General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts, USA; 20) Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA; 21) Diabetes Research Center (Diabetes Unit), Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA; 22) Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; 23) Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK; 24) Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, UK.

Undertaking genetic studies in multiple ancestries can identify disease-related alleles that are common in one population but rare in others. Defects in fasting glucose (FG) and insulin (FI) regulation are hallmarks of type 2 diabetes (T2D). To increase our understanding of the role of low-frequency (minor allele frequency, MAF<5%) and rare (MAF<0.5%) coding variants in influencing these traits, we performed exome-array genotyping (N=33,392) and whole exome sequencing (N=5,108) in a multi-ethnic sample from five ancestries. We aimed to identify novel coding loci and to evaluate coding variants at known loci, thereby highlighting causal transcripts and facilitating characterization of molecular mechanisms that influence glycemic traits and T2D susceptibility. FG and log(FI) levels were adjusted for age, sex and BMI and an inverse normalized transformation was applied to cohort-specific residuals. Regression based association statistics from both the exome-array and the exome sequencing analyses were first meta-analyzed within ancestry and then across ancestries allowing for heterogeneity in genetic effects using MANTRA. Gene-based tests (SKAT and burden) were applied to multiple variant sets including: protein truncating variants (PTV); and PTV and missense variants predicted to be deleterious by 5 algorithms. In addition to replicating reported GWAS signals, we identified novel gene based associations in *AKT2* (SKAT $P=9\times10^{-7}$) and *NUDFAF1* (burden $P=1.1\times10^{-6}$) with FI; *GIMAP8* (burden $P=2.3\times10^{-6}$) with FG; and a single variant association in *AKT2* with FI (p.P50T, rs184042322, MAF from 1.2% in Finland, 0.1% in Sweden and <0.01% in other ancestries). The additive effect of p.P50T on the inverse normalized FI was 0.21 (0.14-0.29 95% CI), meta-analyzed over the GoT2D and T2D-GENES studies in which it is observed ($P=3.7\times10^{-7}$, N=19,259). We replicated this finding in 4 independent Finnish cohorts (replication $P=5.4\times10^{-4}$, N=5,833, combined $P=8.6\times10^{-9}$). Rare penetrant *AKT2* mutations cause monogenic disorders of insulin signaling: kinetically inactivating mutations cause hyperinsulinemia and lipodystrophy while kinetically activating mutations lead to hypoglycemia. Our identification of a missense variant in *AKT2* associated with FI levels extends the allelic spectrum for coding variants in *AKT2* associated with glucose homeostasis and demonstrates the value of studying different populations where "enrichment" of rare alleles may occur.

## 57

**Association of genetic variants with metabolic traits and multiple disease outcomes to inform therapeutic target validation: strengths and limitations of a *GLP1R* variant.** *D.F. Freitag[1,2], R.A. Scott[3], L. Li[4], J.L. Aponte[5], S.M. Willems[6], J. Wessel[7,8], A.Y. Chu[9], S. Wang[10], P. Munroe[11], M. den Hoed[12], I.B. Borecki[13], C. Liu[14], G.M. Peloso[15,16,17], J.M.M. Howson[1], A.S. Butterworth[1], J. Danesh[1,2], J. Dupuis[10], J.I. Rotter[18], J.B. Meigs[19,20], M.O. Goodarzi[21], S. O'Rahilly[22], M.G. Ehm[4], N.J. Wareham[3], D. Waterworth[23], CVD50 consortium, CHARGE Consortium,The CHD Exome+ Consortium, CARDIOGRAM Exome.* 1) Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 2) The Wellcome Trust Sanger Institute, Hinxton, UK; 3) MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK; 4) Statistical Genetics, PCPS, GlaxoSmithKline, RTP, NC, USA; 5) Genetics, PCPS, GlaxoSmithKline, RTP, NC, USA; 6) Genetic Epidemiology Unit, Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands; 7) Fairbanks School of Public Health, Department of Epidemiology, Indianapolis, IN, USA; 8) Indiana University School of Medicine, Department of Medicine, Indianapolis, IN, USA; 9) Division of Preventive Medicine, Brigham and Women's Hospital, Boston MA, USA; 10) Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; 11) Clinical Pharmacology, William Harvey Research Institute, Barts and The London, Queen Mary University of London, London, UK; 12) Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden; 13) Department of Genetics, Division of Statistical Genomics, Washington University School of Medicine. St. Louis, MO, USA; 14) Framingham Heart Study, Population Sciences Branch, NHLBI/NIH, Bethesda, MD, USA; 15) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA; 16) Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA; 17) Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA; 18) Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA; 19) Division of General Internal Medicine, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA; 20) Department of Medicine, Harvard Medical School, Boston, MA, USA; 21) Division of Endocrinology, Diabetes and Metabolism, Cedars-Sinai Medical Center, Los Angeles, CA, USA; 22) University of Cambridge Metabolic Research Laboratories, MRC Metabolic Diseases Unit and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK; 23) Genetics, PCPS, GlaxoSmithKline, Philadelphia, PA, USA.

We investigated the association of genetic variants in 202 genes encoding drug targets with a range of type 2 diabetes (T2D) and obesity-related traits in three studies comprising up to 11,806 individuals with exome sequencing. We sought to identify genetic variants showing promising "on-target" associations with these traits, which may then inform the safety profile or alternative indications for drugs being developed or marketed for obesity and/or T2D. We identified associations in seven genes encoding relevant targets which we took forward for replication by targeted genotyping in five additional studies comprising up to 39,979 participants of European ancestry, and by *in silico* replication for SNPs where available. Following replication, we identified a low frequency (~1% MAF) missense variant (Ala316Thr; rs10305492) in the *GLP1R* gene (encoding the target of the GLP1R-agonist class of T2D-therapies) associated with fasting glucose (β in SDs per (minor) A allele [95% CI]=-0.14 [-0.19, -0.09]; $p=1.1\times10^{-8}$; n=39,753). The minor allele at this variant was also associated with lower risk of T2D (Odds Ratio (OR) =0.83 [0.76, 0.91]; $p=9.4\times10^{-5}$; $n_{cases}$=25,868, $n_{controls}$=122,393). We then performed a comprehensive assessment of the association of the GLP1R variant with a range of phenotypes, comparing them to results observed in clinical trials of GLP1R agonists. While we observed that, similar to GLP1R-agonist therapy, the minor allele was associated with lower fasting glucose and lower risk of T2D, it was not associated with 2-h glucose (β in SDs =0.07 [-0.02, 0.16]; p=0.15; n=39,600), with an opposite direction of effect to that observed in clinical trials. We investigated the association of the variant with other disease outcomes and observed an association of the minor allele with lower risk of coronary heart disease (CHD) (OR)=0.93 [0.87-0.98]; p=0.009; $n_{cases}$=61,846, $n_{controls}$=163,728). However, we found no evidence of a significant association of this variant with pancreatic cancer (OR = 1.23 [0.77, 1.97]; p=0.39; $n_{cases}$=2,307, $n_{controls}$=2,333), with the wide confidence interval indicating the need for far larger sample sizes to inform the safety profile. We demonstrate the potential in using genetic variants with "on-target" associations to predict downstream or "off-target" effects, whilst highlighting some of the challenges of this approach, including the scale of international collaboration required to realise this potential.

## 58

**Dense fine-mapping reveals FOXA2-bound sites as a genomic marker of type 2 diabetes risk.** *K.J. Gaulton[1], T.M. Teslovich[2], T. Ferreira[1], M. Reschen[3], A. Mahajan[1], Y. Lee[2], M. van de Bunt[1], N.W. Rayner[1], A. Raimondo[4], C. O'Callaghan[3], A.L. Gloyn[4,6], A.P. Morris[5], M.I. McCarthy[1,4,6], DIAGRAM Consortium.* 1) Wellcome Trust Centre for Human Genetics, Oxford, UK; 2) Department of Biostatistics, University of Michigan, Ann Arbor MI, USA; 3) Oxford Centre for Cellular and Molecular Physiology, Oxford, UK; 4) Oxford Centre for Diabetes, Endocrinology and Metabolism, Oxford, UK; 5) Department of Biostatistics, University of Liverpool, Liverpool, UK; 6) Oxford NIHR, Churchill Hospital, Oxford, UK.

Genome-wide association studies (GWAS) for type 2 diabetes (T2D) have identified over 80 risk loci. These loci are broadly enriched for regulatory enhancers but the specific causal variants and factors driving a regulatory contribution to disease risk are mostly unknown. In this study we aimed to improve resolution of causal variant(s), and identify regulatory factors through which causal variants may influence T2D. We used Metabochip data from 27,206 T2D cases and 57,574 controls of European ancestry, supplemented by imputation up to 1000 Genomes Project data (March 2012 release) and performed fine-mapping of the 39 established T2D-loci captured on the chip. Conditional analyses revealed a total of 48 independent association signals at these loci: at each, we defined "credible sets" of SNPs that have >99% probability of including the causal variant. Credible sets included no more than twenty variants for 22 signals, with greatest refinement at MTNR1B (only rs10830963, 99.8% causal probability). Credible sets were then integrated with published ChIP-seq sites for 141 DNA-binding proteins (data from ENCODE). We compared the average probability of credible set variants in sites for each protein with the average probability of variants in site locations permuted within 1Mb. We found marked enrichment for FOXA2 binding sites ($P=6\times10^{-5}$), for which ChIP data was available from HepG2 cells and primary pancreatic islets. This enrichment was primarily driven by FOXA2 sites shared with at least one other factor ($P_{shared}=8\times10^{-5}$, $P_{unique}=.007$), and sites identified in islets ($P_{islets}=3\times10^{-4}$, $P_{HepG2}=.008$). To identify specific loci though which the FOXA2 signal might act, we considered FOXA2 binding site overlap at individual loci, finding nominally significant enrichment ($P<.05$) at 13 loci. We then identified 21 candidate regulatory variants in FOXA2 sites at these 13 loci predicted to disrupt a sequence motif (from JASPAR, ENCODE, and HOMER), including rs10830963 (MTNR1B). Electrophoretic mobility shift assays (EMSA) of rs10830963 confirmed allele-specific binding of NEUROD1 to the risk allele in rodent pancreatic beta-cell lines (MIN6, INS1). These results confirm the utility of fine-mapping experiments to resolve causal variants at established risk loci and provide potential molecular mechanisms. Further, FOXA2 binding appears to be a genomic marker of causal T2D variation, through which we identify a putative mechanism for the causal variant at MTNR1B.

## 59

**Integrated analysis of pancreatic islet eQTLs and regulatory state maps identifies putative causal mechanisms at T2D associated loci.** *M. van de Bunt[1,2], J.E. Manning Fox[3], K.J. Gaulton[2], A. Barrett[1], X.Q. Dai[3], M. Ferdaoussi[3], P.E. MacDonald[3], M.I. McCarthy[1,2,4], A.L. Gloyn[1,4].* 1) Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford, United Kingdom; 2) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 3) Alberta Diabetes Institute, Li Ka Shing Centre, University of Alberta, Edmonton, Canada; 4) Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom.

To date around 80 loci have been found associated with type 2 diabetes (T2D). There is compelling evidence that many of these signals act through islet regulatory mechanisms. To define underlying causal mechanisms we therefore need to connect risk-associated regulatory variants to their effector transcripts. We aimed to further mechanistic insight by integrating maps of islet regulatory elements and expression quantitative trait loci (eQTLs). We profiled transcript levels in 142 human islet samples through RNA sequencing on the Illumina HiSeq2000 platform, and obtained genotypes using the Illumina HumanOmni2.5-Exome array and imputation into the 1000 Genomes Phase 1v3 cosmopolitan panel. Reads were aligned to the genome (GRCh37) with TopHat2 and raw read counts were used for exon quantification. To account for potential noise introduced by, for example, differences in donor characteristics and islet purity, exon counts were normalized using PEER. We used data from the 120 individuals of European ancestry in this study to derive islet exon eQTLs. For 74 known common T2D association loci, genetic data were integrated with islet eQTLs and islet ChromHMM regulatory state maps. To be considered, at a given locus the most significant eQTL variant (of all variants with $P<0.01$) had to be in strong LD ($r^2>0.8$) with the lead T2D variant, and either the GWAS or eQTL lead variant had to overlap an islet-active regulatory state annotation. In total, our analysis uncovered a single potential effector transcript at 14 loci. We have started to functionally validate these genes, beginning with the most significant coinciding islet cis-eQTL at *ZMIZ1* ($P=8.1\times10^{-6}$). Immunohistochemistry showed that within the pancreas ZMIZ1 localizes to the islet. We then determined gene expression levels in non-diabetic and T2D islets (n=6 each), which showed significantly higher *ZMIZ1* levels in T2D islets ($P=0.01$) in line with the observed higher *ZMIZ1* expression level for the T2D risk allele ($\beta=0.55$). Overexpression of *ZMIZ1* in human islets (n=6) resulted in a significant reduction in insulin exocytosis ($P<0.0001$). Conversely, knockdown of *ZMIZ1* in T2D islets (n=3) increased insulin exocytosis ($P<0.0001$), and could rescue the blunted exocytotic response normally seen in T2D. Our results demonstrate the power of integrating multiple islet genomic annotations to deliver effector transcripts and molecular mechanisms underlying T2D association signals.

## 60

**T2D-associated ARAP1 regulates GTPase activity, insulin processing and secretion in the pancreatic beta cell.** *J.R. Kulzer, R.C. McMullan, M.P. Fogarty, K.L. Mohlke.* Department of Genetics, UNC Chapel Hill, Chapel Hill, NC.

We previously proposed *ARAP1* as a functional gene at a GWAS locus for type 2 diabetes (T2D) and fasting proinsulin, however, the biological mechanisms connecting *ARAP1*, proinsulin, and T2D remain unknown. We showed that the T2D-risk and proinsulin-decreasing rs11603334-C allele, located in an *ARAP1* promoter, is associated with higher *ARAP1* mRNA expression in human islets. ARAP1 protein contains both an ArfGAP and a RhoGAP domain, which catalyze inactivation of Arf and Rho family GTPases by facilitating GTP to GDP hydrolysis. These GTPases regulate Golgi transport, membrane trafficking, and actin cytoskeleton dynamics, processes involved in insulin granule packaging and trafficking. We hypothesized that ARAP1 may act through one or more Arf or Rho GTPases to exert a regulatory role on insulin processing and/or secretion. To identify which GTPases are regulated by ARAP1 in the beta cell, we measured GTPase activity following transient ARAP1 overexpression. Preliminary results show that exogenous overexpression of wild-type (wt) human ARAP1 in the mouse insulinoma cell line MIN6 led to a 58% decrease in Arf5-GTP levels and a 48% decrease in Arf1-GTP levels. Exogenous overexpression of an ARAP1 mutant with catalytically inactive ArfGAP and RhoGAP domains rescued levels of Arf1-GTP, but not Arf5-GTP. Using confocal fluorescence microscopy, we observed moderate colocalization of ARAP1 with Arf1 near the Golgi in MIN6 and in 832/13 rat insulinoma cells, and with Arf6 in dispersed human islet beta cells. To examine the effect of ARAP1 on proinsulin and insulin secretion, we exogenously overexpressed wt ARAP1 in MIN6 cells and measured supernatant insulin levels following KCl stimulation. We observed a 20% decrease in proinsulin secretion ($P < .001$), consistent in direction with GWAS findings that rs11603334-C is strongly associated with decreased plasma proinsulin levels. In the same samples, insulin secretion was found to be increased 70% ($P < .03$), suggesting a role for ARAP1 in proinsulin-to-insulin processing. ARAP1 ArfGAP/RhoGAP mutants rescued both proinsulin and insulin levels. Given the potential for species differences, we are extending these studies to human islets; we are also evaluating additional candidate GTPases. Our results suggest that upon acute stimulation, overexpressed ARAP1 may lead to decreased levels of secreted proinsulin through its regulation of Arf1 and/or Arf5 GTPase activity in pancreatic beta cells.

## 61

**Discordant Non-invasive Prenatal Testing and Cytogenetic Results: Is There a Cause for Concern?** *J. Wang[1], T. Sahoo[1, 2], S. Schonberg[3], K. Kopita[1], L. Ross[1], K. Patek[3], C. Strom[1].* 1) Cytogenetics Laboratory, Nichols Institute, Quest Diagnostics, San Juan Capistrano, CA; 2) CombiMatrix, 300 Goddard Suite 100, Irvine, CA 92618, USA; 3) Cytogenetics Laboratory, Quest Diagnostics Nichols Institute, 14225 Newbrook Drive, Chantilly, VA, 20151, USA.

Recent published studies have demonstrated the incremental value of the use of cfDNA for non-invasive prenatal testing (NIPT), with 100% sensitivity for trisomies 21 and 18, and specificity of ≥99.7% for both. Data presented by two independent groups suggested positive results by NIPT were not conformed by cytogenetic studies. Concordance of results between cases with positive NIPT results referred for cytogenetic prenatal and/or postnatal studies by karyotyping, FISH, and/or oligo-SNP microarray were evaluated for 98 consecutive specimens. Cytogenetic results were positive for trisomy 21 in 38 of the 41 NIPT-positive cases (true positive rate: 93%) and for trisomy 18 in 16 of the 25 NIPT-positive cases (true positive rate: 64%). The true positive rate was only 44% (7/16 cases) for trisomy 13 and 38% (6/16 cases) for sex chromosome aneuploidy (SCA). Confined placental mosaicism was confirmed in 2 of 98 cases (2%). Our data (N = 98) and that presented by the above mentioned two groups (N = 80 and N = 46) show that the positive predictive values (PPV) is highest for cases positive for trisomy 21 (119/126, 94.4%). A significantly lower PPV ($p < 0.001$) is displayed for trisomy 18 (25/42, 59.5%), trisomy 13 (12/27, 44.4%), and SCA (11/29, 37.9%). These findings raise concerns about the limitations of NIPT and the need for analysis of a larger number of false positive cases to provide true PPVs for such non-invasive testing and search for potential biological or technical causes. It is very important to understand that PPV is not intrinsic to the test; it depends also on the prevalence of the condition in the tested population. The difference of PPV for trisomy 21 by NIPT from other aneuploidies could be due to either a more prevalence or a higher specificity of NIPT for Down syndrome than other aneuploidies, or both. These findings suggest the need for a careful interpretation of NIPT results and cautious transmission of the same to providers and patients. It is vital to educate ordering physicians regarding the differences between specificity and PPV. To an average clinician, the claim that a test is >99% specific leads him or her to expect that the false positive rate will be less than 1%. As can been seen from the data here, the performance of NIPT for correctly predicting positive for trisomy18 and trisomy 13 is less than 60%, necessitating a cautious evaluation of the causes and consequences before NIPT is made available to the general population.

## 62

**Implementation of microarray analysis for oncology samples: Effectiveness for detection of both copy number changes and copy-neutral loss of heterozygosity.** *S. Schwartz[1], R. Burnside[1], I. Gadi[1], V. Jaswaney[1], E. Keitges[2], A. Penton[1], K. Phillips[1], H. Risheg[2], J. Schleede[1], J. Tepperberg[1], P. Papenhausen[1].* 1) Laboratory Corporation of America, Research Triangle, NC; 2) DynaCare/Laboratory Corporation of America, Seattle, WA.

Over the past six years cytogenetic analysis has undergone a renaissance with microarray analysis utilized for the delineation of constitutional abnormalities. However, in oncology the acceptance of microarray analysis has been slower. We have studied over 4,000 oncology patients (including MDS, CLL, AML, ALL and MM), utilizing reporting cut-offs of 1 Mb for deletions and 2 Mb for duplications. However, changes as low as 50 kb were reported when significant pathogenic genes were involved. This work has not only proven the importance and efficacy of utilizing a SNP/oligo platform, but has yielded some illuminating results including: (a) The importance of establishing clear reporting cut-offs based on 70,000 previous pediatric studies; (b) The consideration that each hematologic disorder needs to be approached individually, some as an adjunct procedure (AML, ALL) and some as a primary tool (MDS, CLL); (c) The effectiveness in delineating abnormalities in all hematologic groups when the karyotype was normal, ranging from ~18.1% of MDS patients to ~78.6% of ALL patients; (d) The usefulness in delineating additional important abnormalities, even after FISH/cytogenetics revealed an abnormality [~50% (MDS and CLL) to ~88% (myeloma and ALL)]; (e) The importance and high incidence of copy-neutral loss of heterozygosity (CN-LOH) in all hematologic disorders [~10.2% of patients (MDS) to ~40% in CMML]. The CN-LOH was seen most often as a single finding, but multiple CN-LOH findings as well as temporal expansions were also seen; (f) The efficacy of its use as a heme-diagnostic tool in MDS, in which array abnormalities were detected in ~18% of patients in which Flow/hematopathology were not diagnostic and in ~36% of patients where Flow/hematopathology findings were only suspicious; (g) The prognostic utility for all groups, but specifically in ALL, where the genotyping array differentiates good prognosis hyperdiploidy from a poor prognostic doubled hyperhaploid karyotype; (h) The frequency of variants of unknown significance, due to detection cut-offs was only ~3.6%; (i) The ability to resolve structurally abnormal chromosomes in all cases, in the ability to detect chromothripsis (associated with a poor prognosis) and for the discovery of oncogenic fusion genes; and (h) Lastly, this work showed that a firm knowledge of pediatric array findings is necessary as constitutional pathogenic syndromes and consanguinity have been detected at a higher frequency than expected.

## 63

**A cost-effective screen for identifying novel transposable element insertions in human genomes.** *E.M. Kvikstad, G. Lunter.* Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.

Transposable elements (TEs) are mobile genetic elements that randomly insert copies of themselves into their host's genome. As such, they are major drivers of genome evolution and responsible for substantial genomic variation between individuals. The potential mutational target for disease-causing TE insertions is large, as mutagenesis can occur by disrupting genes or modifying their expression through interference with proper splicing, disruption of exons, premature transcript termination or donating alternative transcription start sites. Since existing screens for mutations underlying disease typically ignore TE insertions, we hypothesize that the role of TEs in disease has thus far been underestimated.

Here, we present a cost-effective high-throughput and genome-wide screen for TE insertions to enable investigating the impact of TEs on disease. Briefly, the strategy consists of standard Illumina library preparation, followed by polymerase chain reaction (PCR) to enrich for genomic fragments containing sequence signatures of known active TEs subfamilies. Unlike existing approaches, our strategy simultaneously targets three of the most active and prolific TE subfamilies (AluYb8/b9, AluYa5, and human-specific L1HS elements) that together constitute only ~0.36% of the human genome, but the majority of polymorphic insertions. Preliminary results indicate that our approach yields up to 20,000-fold enrichment for targeted sequence fragments, each containing on average 62 nucleotides of unique sequence flanking the poly-A signature, allowing accurate mapping and detection of polymorphic and novel insertion sites.

Our protocol is compatible with exome sequencing protocols, and is particularly cost-effective in that setting by sharing the library preparation stage as well as allowing high levels of multiplexing. Therefore, applications of this protocol in parallel to whole-exome screens of sporadic disease cases will result in improved estimates of the contribution of TEs to sporadic disease, and potentially reduce the number of unresolved cases in such screens.

## 64

**NUC-Seq: Single-Cell Exome Sequencing Using G2/M Nuclei.** *M.L. Leung[1,3], Y. Wang[1], J. Waters[1], N.E. Navin[1,2,3].* 1) Department of Genetics, University of Texas MD Anderson Cancer Center, Houston, TX; 2) Department of Bioinformatics and Computational biology, University of Texas MD Anderson Cancer Center, Houston, TX; 3) Graduate Program in Genes and Development, Graduate School of biomedical Sciences, University of Texas Health Science Center at Houston, Houston, TX.

Single-cell sequencing (SCS) methods have the potential to provide great insight into the genomics of rare cells and diverse populations, but are currently challenged by extensive technical errors. These errors include poor physical coverage and high FP and FN error rates, making it difficult to distinguish real biological variants. To address this problem, we have developed an SCS method called NUC-Seq that combines flow-sorting of single nuclei, limited multiple-displacement-amplification, low-input library preparation and exome capture to generate high coverage (>90%) data on single mammalian cells. To mitigate error rates, we collected nuclei that are in the G2/M stage of the cell cycle, providing 4 copies of the genome as input material for whole genome amplification (instead of 2 copies). We developed our method using a normal fibroblast cell line (SKN2) in which we can assume that variants in single cells will be highly similar to the population sample. We sequenced the reference population sample and the exomes of 9 single cells in G1/0 stage and 10 single cells in G2/M stage of the cell cycle. Our data suggest that G2/M cells provide several major technical improvements over using G1/0 cells, including decreased allelic dropout rates (21.52%), improved coverage breadth (95.94%), improved coverage uniformity and better detection efficiencies for SNVs (92.53%). On average, we detect 24 false positive errors per mega-bases in each single cell genome. We show that these errors occur randomly in each cell, allowing accurate variant detection using two or more single cells. Our data suggest that, regardless of whether G1/0 or G2/M cells are used as input material, major technical improvement are achieved compared to existing single-cell sequencing methods. In summary, we expect that NUC-Seq will have broad applications in fields as diverse as cancer research, microbiology, neurobiology, development and in vitro prenatal genetic diagnosis, and will greatly improve our fundamental understanding of human diseases.

## 65

**Simple and robust NGS RNA-based assay to assess impact of VUS on splicing.** *E. Girard[1], J. Tarabeux[2], E. Bernard[1], A. Collet[3], A. Legrand[3], V. Moncoutier[3], C. Dehainault[3], J.P. Vert[1], D. Stoppa-Lyonnet[2,4], N. Servant[1], C. Houdayer[2,4].* 1) Institut Curie, INSERM U900, Mines ParisTech, Paris, France; 2) Institut Curie, Service de Génétique, INSERM U830, Paris, France; 3) Institut Curie, Service de Génétique, Paris, France; 4) Université Paris Descartes, Sorbonne Paris Cité, France.

One of the key issues raised in molecular diagnosis is the correct interpretation of variants of unknown significance (VUS). Each VUS can potentially affect pre-mRNA splicing and be deleterious via disruption / creation of consensus sequences. Assessing the impact of VUSs on splicing is a central issue in order to determine their pathogenicity. Complete splicing defects can be easily demonstrated e.g. for exonic variants and for intronic variants when an exonic SNP is used as an indirect marker of instability. Conversely, defects affecting a fraction of the transcripts produced are more difficult to interpret as quantitative evaluation of partial splicing defects cannot be based on inspection of Sanger traces. Suitable quantitative methods such as pyrosequencing are available but not used in routine diagnosis. We therefore developed a targeted-single gene RNAseq strategy which lends itself as a simple and robust method providing qualitative as well as quantitative information on VUS impact at the RNA level. Patients with distinct *RB1* splice mutations previously ascertained by Sanger sequencing were tested (resulting in out-of-frame/in-frame exon skipping and/or intronic retention). The complete cDNA sequence was amplified in a single PCR experiment and sequenced on a PGM (Life Technologies) following enzymatic shearing, barcoding and adaptor ligation. Reads were trimmed to a fixed length of 200bp, and aligned on the Human reference genome (hg19) with the splice mapper Tophat2. The different isoforms were then reconstructed from the aligned data and quantified using two different strategies (Cufflinks and R BioConductor package flipflop). All expected splicing events evidenced by Sanger were detected and the relative proportion of these anomalies was quantified. A number of alternative transcripts previously undetected by Sanger were also identified at a low frequency. This first result demonstrates the capabilities of RNA sequencing technics to assess the impact of VUS on splicing. The design of the PCR amplicons is crucial in order to extend this approach to more complex genes. This cost effective and simple approach can be easily implemented and has the potential to replace existing Sanger studies in a diagnostic context as the complex mixture of unstable, truncated, in frame and wild-type transcripts generated by splice mutations can be deciphered and quantified, hence allowing robust VUS interpretation and in turn reliable genetic counseling.

## 66

**Making sense of sequence variation in PPARG: a comprehensive experimental approach.** *A. Majithia, J. Flannick, T. Mikkelsen, D. Altshuler.* Broad Institute, Cambridge, MA.

Discriminating benign and functional genetic variants is a key challenge in clinical interpretation of exome sequencing. The current gold standard for functional evaluation requires that variants identified in sequencing be recreated and tested in vitro, a resource intensive process that is not feasible on clinical time scales. We present a prospective experimental approach and demonstrate its application to interpret non-synonymous variants in PPARG, a master regulator of adipocyte differentiation. To generate a complete functional catalog, we synthesized a comprehensive library of 9576 PPARG variants consisting of every single amino acid substitution at every position and performed pooled screens for protein function in human pre-adipocytes. We validate this functional catalog by re-identifying previously known loss-of-function mutations in PPARG that cause human lipodystrophy and clinical insulin resistance. Subsequently, we apply it to prospectively evaluate the consequence of 50 novel non-synonymous PPARG variants of unknown function identified from exome sequencing of 20,000 type 2 diabetes (T2D) cases and controls. As a control, we also generated these newly discovered PPARG variants by conventional means, tested their function individually in an adipocyte differentiation assay, and compare results with our comprehensive functional catalog. When taken together, these variants showed no association with T2D (OR=1.35 p=0.17). However, of the 50 variants, 9 experimentally demonstrated reduced function, ranging from mild to complete inactivation of PPARG in the adipocyte differentiation assay. These functional variants substantially increased risk of T2D (OR= 7.22 p=0.005). Our findings establish that rare, loss-of-function variants in PPARG increase T2D risk in the general population, and that they must be distinguished from a majority of benign variants. Our comprehensive functional catalog serves as a resource for interpreting the clinical consequence of novel PPARG variants as they are discovered in future exome sequencing.

## 67

**Molecular Combing for Fascioscapulohumeral Dystrophy Type 1: Benefits of Direct Visualization of DNA Fibers.** *C.M. Strom, J.C. Wang, X.J. Yang, B.H. Nguyen, V. Sulcova, P. Chan, Y. Liu, A. Anguiano, F.Z. Boyar.* Department of Cytogenetics, Nichols Institute, Quest Diagnostics, 33608 Ortega Highway, San Juan Capistrano, CA 92690.

Fascioscapulohumeral dystrophy (FSHD) is the third most common muscular dystrophy. Type-1 FSHD is due to a contraction of the D4Z4 microsatellite repeat motif at chromosome 4q subtelomeric region. Chromosome 10q also contains copies of this repeat motif, of which contractions are not associated with FSHD. Two common haplotypes exist at 4q locus: 4qA and 4qB. Differentiation is important for diagnosis, as contractions of the 4qA but not the 4qB allele are associated with FSHD. Prior to the advent of DNA combing, molecular diagnosis relied on a series of pulsed-field Southern blots. These analyses can lead to false-positive results, because they cannot differentiate the A allele from the B allele on chromosome 4q nor can they differentiate chromosome 4q from chromosome 10q. Molecular combing is a technique in which DNA is uniformly stretched and then hybridized with gene-specific probes of various colors to create a molecular bar code. We developed and validated a molecular combing test to identify and measure contractions of the 4qA allele. Here we report the first 44 suspected FSHD cases tested with the molecular combing assay in our laboratory. In all 44, we were able to unambiguously identify all 4q and 10q alleles and determine the size of the D4Z4 repeat. Of the 44 samples, 13 (30%) were clearly affected, with 4qA repeat sizes between 3 and 8; 28 (63%) were clearly normal, with 4qA repeat sizes between 12 and 68; and 3 (7%) had a 4qA allele of 10-11 repeats, which is a borderline result. Two patients had a 4qB contraction and 15 had either 10qA or 10qB contraction, which could have led to a false-positive result if Southern blot had been used. We conclude that a molecular combing assay for FSHD is capable of determining the 4qA repeat size in clinical samples and can prevent false-positive results by differentiating 4qA from 4qB, 10qA, 10qB alleles. Additionally, more accurate measurement of repeat number by DNA combing methods may help correlate the size of the contracted 4qA allele with the timing of FSHD1 onset.

## 68

**An Augmented Exome Providing Accurate Structural Variant Detection.** *A. Patwardhan, S. Chervitz, M. Li, J. Harris, G. Bartha, D. Newburger, M. Pratt, S. Garcia, J. Tirch, N. Leng, C. Haudenschild, S. Luo, D. Church, J. West, R. Chen.* Personalis, Inc., Menlo Park, Ca.

Whole-Exome Sequencing (WES) has proven to be an efficient tool in identifying common and rare disease-associated variants in the protein-coding region of the genome. This has had a direct impact on the clinical setting, where WES has provided definitive diagnoses to patients affected by Mendelian Disorders, even in cases where other diagnostic tests have failed. While WES theoretically provides coverage of all potential disease-associated variants within the exome, the majority of published exome studies report only the detection of small variants (SNVs, InDels). The false-negative rate for Structural Variations (SVs), a class of large (>50bp) variants that include copy number variations, is thought to remain high given their established or expected importance in disease phenotype. The size and complexity of SVs increases the likelihood that they will encompass a region of variation not well-captured by standard WES, partially due to the targeted nature of the capture methods, incomplete gene coverage, and the limited size of coding sequence.

We describe a novel approach that combines an augmented exome assay with novel informatics to address many of these technical challenges and improve SV calling accuracy. The augmented exome improves coverage over all biomedically interpretable genes compared to standard WES assays and extends coverage to detect large SVs genome wide. Using read-depth information, SV detection is performed concurrently in exome-targeted regions and genome-wide, with corrections for non-uniform coverage when appropriate.

SV detection rates were compared among a set of over 40 samples known to harbor pathogenic deletions and duplications using augmented exome and standard WES approaches. Additionally, the accuracy of our approach relative to standard WES is estimated by comparing SVs detected in a reference sample to a “gold set” of SVs developed in-house. This “gold-set” is derived from a deeply sequenced 16-member pedigree where SVs in the reference sample are vetted by breakpoint, inheritance state, and variant type consistency among related pedigree members. Using both the known samples and the “gold-set”, we demonstrate increased accuracy in detecting SVs with the augmented exome over a range of SV sizes (1KB- >1MB).

## 69

**Mutations in TENM4, a regulator of axon guidance and central myelination, cause essential tremor.** *H. Hor[1,2,3,4], L. Francescatto[5], L. Bartesaghi[6], S. Ortega-Cubero[7], M. Kousi[8], O. Lorenzo- Betancor[7], F.J. Jiménez-Jiménez[8], A. Gironell[9,10], J. Clarimón[10,11], O. Drechsel[1,2], J.A.G. Agúndez[12], D. Kenzelmann Broz[13], R. Chiquet-Ehrismann[13], A. Lleó[10], F. Coria[14], E. García-Martin[15], H. Alonso-Navarro[8], M.J. Martí[16], J. Kulisevsky[9,11], C.N. Hor[1,2,3,4], S. Ossowski[1,2], R. Chrast[6], N. Katsanis[5], P. Pastor[7], X. Estivill[1,2,3,4,17].* 1) Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain; 2) Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain; 3) Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain; 4) CIBER de Epidemiología y Salud Pública (CIBERESP), Barcelona, Catalonia, Spain; 5) Center for Human Disease Modeling, Duke University, Duke University Medical Center, Durham, USA; 6) Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland; 7) Neurogenetics Laboratory, Division of Neurosciences, Center for Applied Medical Research (CIMA), and Department of Neurology, Clínica Universidad de Navarra, University of Navarra School of Medicine and Centro de Investigación Biomédica; 8) Section of Neurology, Hospital Universitario del Sureste, Arganda del Rey, Madrid, Spain; 9) Movement Disorders Unit, Neurology Department, Hospital de Sant Pau, Barcelona, Catalonia, Spain; 10) Sant Pau Biomedical Research Institute, Barcelona, Catalonia, Spain; 11) Universitat Autònoma de Barcelona and CIBERNED, Barcelona, Catalonia, Spain; 12) Department of Pharmacology, University of Extremadura, Cáceres, Spain; 13Friedrich Miescher Institute of Biomedical Research, Novartis Research Foundation and University of Basel, Faculty of Sciences and Department of Biomedicine, Basel, Switzerland; 13) Friedrich Miescher Institute of Biomedical Research, Novartis Research Foundation and University of Basel, Faculty of Sciences and Department of Biomedicine, Basel, Switzerland; 14) Clinic for Nervous Disorders, Service of Neurology, Son Espases University Hospital, Palma de Mallorca, Spain; 15) Department of Biochemistry and Molecular Biology, University of Extremadura, Cáceres, Spain; 16) Movement Disorders Unit, Neurology Service, Hospital Clinic, CIBERNED and Institut d'Investigacions Biomédiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain; 17) Dexeus Women’s Health, University Hospital Quiron-Dexeus, Barcelona, Catalonia, Spain.

Introduction: Essential tremor (ET) is the most common movement disorder with an estimated prevalence of 5% in individuals older than 65 years. To date, genetic studies have failed to identify a common genetic cause for this condition. Methods: We performed exome sequencing in an ET family and identified a candidate causal genetic variant in *TENM4*, which segregated in an autosomal dominant fashion. We then used targeted resequencing in 299 unrelated familial ET cases to identify additional families with segregating mutations in *TENM4*. We performed in vitro assays in HEK cells and oligodendrocyte precursor cells to elucidate the properties of TENM4, and used a zebrafish model to interrogate the effect of the identified *TENM4* mutations. Results: We identified a missense mutation (C4100A) in *TENM4* that segregated with ET in a nuclear family. Subsequent targeted resequencing of *TENM4* led to the discovery of two novel missense mutations, each of which segregated with ET in two additional families. Consistent with a dominant mode of inheritance, in vitro analysis in either HEK cells or oligodendrocyte precursor cells showed that mutant proteins are both unstable and mislocalize. Finally, expression of human mRNA harboring any of three patient mutations in zebrafish embryos induced defects in axon guidance, confirming a dominant-negative mode of action for these mutations. Conclusion: Our genetic and functional data, which is corroborated by the existence of a *Tenm4* knock-out mouse displaying an essential tremor phenotype, implicates TENM4 in ET as a major contributor to the etiopathology of the disorder. Together with previous studies of TENM4 in model organisms, our studies intimate that processes regulating myelination in the central nervous system and axon guidance might be significant contributors to the genetic burden of this disorder.

## 70

**Mitochondrial serine protease HTRA2 p.G399S in a 6-generation kindred with Essential Tremor and Parkinson's Disease.** *H. Unal Gulsuner[1], S. Gulsuner[2], N. Durmaz Mercan[3], O.E. Onat[4], T. Walsh[2], H. Shahin[5], O. Dogu[6], T. Kansu[7], H. Topaloglu[8], B. Elibol[7], C. Akbostanci[3], M.-C. King[2], T. Ozcelik[4], A.B. Tekinay[1].* 1) Institute of Materials Science and Nanotechnology, National Nanotechnology Research Center (UNAM), Bilkent University, Ankara, Turkey; 2) Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA; 3) Department of Neurology, Faculty of Medicine, Ankara University, Ankara, Turkey; 4) Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey; 5) Department of Life Sciences, Bethlehem University, Bethlehem, Palestinian Territory; 6) Department of Neurology, Faculty of Medicine, Mersin University, Mersin, Turkey; 7) Department of Neurology, Faculty of Medicine, Hacettepe University, Ankara, Turkey; 8) Department of Pediatrics, Neurology Unit, Faculty of Medicine, Hacettepe University, Ankara, Turkey.

Essential tremor (ET) is one of the most common movement disorders in humans. It is characterized primarily by postural and kinetic tremor of the arms and hands, but the head, legs, voice, and other regions of the body may also be affected. Genes responsible for most of ET are not yet known, with mutation of FUS identified as responsible for the phenotype in one family. Our study focused on a 6-generation family of Turkish ancestry including 16 relatives with varying degrees of tremor. Three of these relatives developed ET as children and signs of Parkinson's disease (PD) in middle age. We carried out whole exome sequencing using DNA from three severely affected relatives, and then genotyped all shared putatively damaging variants in 24 informative family members. Both dominant and recessive modes of inheritance were considered. Exactly one potentially damaging variant, HTRA2 p.G399S, co-segregated with ET in the family. Of 16 affected individuals, 5 were homozygous and 11 were heterozygous for the variant allele 399Ser. The 5 homozygous individuals included the 3 relatives with both ET and PD and 2 relatives younger than age 30 who developed ET at ages 10 and 12 years. HTRA2 genotype was significantly associated with age at onset of tremor (mean onset for homozygotes age 19 and for heterozygotes age 40, P = 0.010), severity of postural tremor (P = 0.0002) and severity of kinetic tremor (P = 0.011). No mutations in the complete HTRA2 sequence were detected in probands of 25 other Turkish families with ET. Among Turkish controls (ages 20-30), 2 of 364 were heterozygous for HTRA2 p.G399S (allele frequency 0.0027). HtrA2 deficiency is responsible for motor neuron degeneration in the mouse model Mnd2. Heterozygosity for HTRA2 p.G399S was associated with PD in some, but not all, studies. Our results suggest that HTRA2 p.G399S is responsible for hereditary ET in this kindred and that homozygosity for this variant leads to development of parkinsonian features. Although ET and PD have been known as distinct entities, our results show the co-existence of both disorders. These results might reveal shared etiologic factors underlying ET and PD phenotypes.

## 71

**De novo mutations in SIK1 dysregulate HDAC5-MEF2C activity and cause Ohtahara syndrome and infantile spasms.** *J.N. Hansen[1], C. Snow[2], E. Tuttle[1], D. Ghoneim[1], C. Smyser[3], C.A. Gurnett[3], M. Shinawi[4], W.B. Dobyns[5], J. Wheless[6], M.W. Halterman[1,7], L.A. Jansen[8], B.M. Paschal[2,9], A.R. Paciorkowski[1,7,10].* 1) Center for Neural Development and Disease, University of Rochester, Rochester, NY; 2) Center for Cell Signaling, University of Virginia, Charlottesville, VA; 3) Department of Neurology, Washington University, St. Louis, MO; 4) Department of Pediatrics, Division of Genomic Medicine, Washington University, St. Louis, MO; 5) Department of Neurology and Division of Genetic Medicine, Department of Pediatrics, University of Washington and Center for Integrative Brain Research, Seattle Research Institute, Seattle, WA; 6) LeBonheur Children's Hospital and the University of Tennessee, Memphis, TN; 7) Department of Neurology, University of Rochester Medical Center, Rochester, NY; 8) Department of Neurology, University of Virginia, Charlottesville, VA; 9) Departments of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA; 10) Departments of Pediatrics and Biomedical Genetics, University of Rochester Medical Center, Rochester, NY.

Ohtahara syndrome and infantile spasms are severe childhood epilepsies for which genetic causes have been increasingly identified, including mutations in the forebrain transcription factors *ARX*, *FOXG1* and *MEF2C*. Here we report 3 individuals with Ohtahara syndrome and 3 individuals with infantile spasms with de novo mutations in the salt-inducible-kinase *SIK1* that dysregulate downstream HDAC5 and MEF2C activity. A subset of SIK1 mutants are expressed at higher levels than wild-type SIK1 and are also more resistant to proteasomal degradation. Downstream signaling of SIK1 is also disrupted by these mutations, with SIK1 mutants exhibiting reduced MEF2C transcriptional activity and/or increased HDAC5 phosphorylation in vitro. These observations suggest that *SIK1* mutations causing Ohtahara syndrome and infantile spasms are likely gain-of-function mutations. This report is the first example of a human disease associated with *SIK1* mutations, and a novel cause of developmental epilepsy associated with dysregulation of the forebrain transcription factor *MEF2C*.

## 72

**Haploinsufficiency of Pumilio1 leads to SCA1-like neurodegeneration by increasing wild-type Ataxin1 levels in a miRNA-independent manner.** *V.A. Gennarino[1,2], R. Singh[3], J.J. White[2,3,4], K. Han[1,2,5], A. De Maio[2,6], P. Jafar-Nejad[1,2], A. di Ronza[1,2], H. Kang[1,2,7], H.T. Orr[8], R.V. Sillitoe[2,3,4,6], H.Y. Zoghbi[1,2,3,5,6,9].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, 77030, USA; 2) Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, Texas, 77030, USA; 3) Department of Pathology and Immunology, Baylor College of Medicine, Houston, Texas, 77030, USA; 4) Department of Neuroscience, Baylor College of Medicine, Houston, Texas, 77030, USA; 5) Howard Hughes Medical Institute, Baylor College of Medicine, Houston, Texas, 77030, USA; 6) Program in Developmental Biology, Baylor College of Medicine, Houston, Texas, 77030, USA; 7) National Institute of Supercomputing and Networking, Korea Institute of Science and Technology Information, Daejeon, South Korea; 8) Institute for Translational Neuroscience, Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, Minnesota; 9) Department of Pediatrics, Baylor College of Medicine, Houston, Texas, 77030, USA.

Accumulation of mutant disease-driving proteins in specific brain regions promotes different neurodegenerative disorders including Alzheimer, Parkinson, and spinocerebellar ataxias. Spinocerebellar ataxia type 1 (SCA1) is a fatal dominant neurodegenerative disease characterized by progressive loss of motor abilities primarily due to degeneration of Purkinje neurons. SCA1 is caused by expansion of an unstable CAG repeat in *Ataxin1* (*ATXN1*). The resulting protein harbors an expanded polyQ tract rendering the protein toxic through a gain-of-function mechanism. A striking feature of *ATXN1* is the extraordinarily long 3'UTR (~7kb) that we hypothesized must harbor key post-transcriptional regulatory elements. We rationalized that deciphering the post-transcriptional mechanisms regulating ATXN1 levels *in vivo* will provide a better insight into factors that might contribute to SCA1 pathogenesis. We identified a conserved binding motif for the RNA-binding protein Pumilio1 (PUM1) in the 3'UTR of *ATXN1* mRNA. Using RNA-Clip we found that Pum1 physically interacts with the conserved binding site of the *Atxn1*-3'UTR in mouse brain. It is known that PUM1 can regulate microRNA-dependent gene silencing by induction of a conformational switch in the 3'UTR of its targets. However, by mutating miRNA targets sites in *ATXN1* as well as knocking down *AGO2*, we found that PUM1 modulates ATXN1 independent of miRNA interaction. Importantly, *Pum1* heterozygous (*Pum1*$^{\pm}$) mice show an increase of both Atxn1 protein and mRNA levels by 30 to 50% in the brain and a wide range of neurological defects, such as impaired motor coordination and Purkinje cells degeneration, similar to those in SCA1 mouse model. Moreover, removing one allele of *Pum1* greatly enhanced the disease progression in SCA1 mice. Interestingly, breeding *Pum1*$^{\pm}$ mice to mice lacking one wild-type allele of *Atxn1* (*Atxn1*$^{\pm}$) normalized Atxn1 levels and rescued most of the behavioral and pathological phenotypes observed in *Pum1*$^{\pm}$ mice. These data demonstrate that precise levels of Pum1 are critical for neuronal maintenance, that *Atxn1* is a key target of Pum1 and mediates crucial neuropathological features of its haploinsufficiency, and that a mild increase in Atxn1 levels is detrimental to neurons. Lastly, these data underscore the potential for *PUM1* and *ATXN1* as candidate neurodegenerative disease genes either through haploinsufficiency or duplication, respectively, or through mutations in their regulatory regions.

## 73

**Exome sequencing unveils novel disease-causing variation in Charcot-Marie-Tooth disease and suggests genetic burden contributes to phenotypic variability and complex neuropathy.** *C. Gonzaga-Jauregui[1,2], T. Harel[1], T. Gambin[1], M. Kousi[2], L.B. Griffin[3,4], M.N. Bainbridge[5], K.S. Lawson[6], D. Pehlivan[1], Y. Okamoto[1], M. Withers[1], P. Mancias[7], A. Slavotinek[8], P.J. Reitnauer[9], M. Shy[10], T.O. Crawford[11], M. Koenig[12,13], M.T. Goksungur[14], S. Jhangiani[5], J. Willer[2], B.N. Flores[3], W. Wiszniewski[1], A. Antonellis[3,15,16], N. Katsanis[2], D.M. Muzny[5], E. Boerwinkle[5,6], R.A. Gibbs[1,5], J.R. Lupski[1,17,18], Baylor-Hopkins Center for Mendelian Genomics.* 1) Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA; 2) Center for Human Disease Modeling, Duke University, Durham, NC, 27710, USA; 3) Cellular and Molecular Biology Program, University of Michigan Medical School, Ann Arbor, MI, 48109, USA; 4) Medical Scientist Training Program, University of Michigan Medical School, Ann Arbor, MI, 48109, USA; 5) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, 77030, USA; 6) Human Genetics Center and Institute of Molecular Medicine, University of Texas-Houston Health Science Center, Houston, TX, 77030, USA; 7) Department of Neurology and Pediatrics, Division of Child & Adolescent Neurology, University of Texas Medical School at Houston, Houston, TX, 77030, USA; 8) Division of Genetics, Department of Pediatrics, University of California, San Francisco, CA, 94158, USA; 9) Pediatric Teaching Program, Cone Health System and UNC-Chapel Hill, Greensboro NC 27401, USA; 10) Department of Neurology, Carver College of Medicine, University of Iowa, Iowa City, IA, 52242, USA; 11) Departments of Neurology and Pediatrics, Johns Hopkins University, Baltimore, Maryland, USA; 12) Institut de Genetique et de Biologie Moleculaire et Cellulaire (IGBMC), CNRS-INSERM-Universite de Strasbourg, Illkirch, 67404, France; 13) INSERM UMR_S 827, Institut Universitaire de Recherche Clinique, and Laboratoire de Génétique Moléculaire, Centre Hospitalier Universitaire de Montpellier, Montpellier, France; 14) Department of Neurology, Istanbul University, Istanbul Medical Faculty, Istanbul, Turkey; 15) Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, 48109, USA; 16) Department of Neurology, University of Michigan Medical School, Ann Arbor, MI, 48109, USA; 17) Department of Pediatrics, Baylor College of Medicine, Houston, TX, 77030, USA; 18) Texas Children's Hospital, Houston, TX, 77030, USA.

Charcot-Marie-Tooth (CMT) disease is the most common hereditary neuropathy affecting approximately 1/2500 individuals. It is a clinically heterogeneous distal symmetric polyneuropathy (DSP) with two major groups distinguished electrophysiologically and pathologically into demyelinating CMT1 and axonal CMT2. CMT shows extensive underlying genetic heterogeneity, with ~50 loci identified or linked to date to different subtypes of the disease. Exome sequencing (ES) allows assessment of most of the coding variation in the human diploid genome, but the interpretation can be complicated by the presence of extensive allelic and genetic heterogeneity and oligogenic inheritance with rare variants in more than one causative gene. We performed ES at high coverage in a cohort of 40 patients from 37 unrelated families with CMT-like peripheral neuropathy, in whom the genetic cause had not been previously identified using a multitude of molecular genetic analyses. We identified the apparent disease-causative variants in 47.5% of patients, accounting for 18 of 37 of the families; and potentially disease causing variants in novel genes in three additional families. We also show that affected individuals can have an enrichment of rare variants in multiple CMT genes as compared to unaffected control individuals, and we propose that this mutation burden likely contributes to phenotypic variability of the disease in the population. These findings are consistent with the proposed Clan Genomics hypothesis which posits that new mutations in patients or those that arose in recent ancestors, and novel combinations from the proband's parents, rather than common/ancient alleles in populations, can have a more pronounced effect on the proband's phenotypic presentation and account for medically actionable variants. We performed functional studies in yeast and zebrafish to elucidate the functional impact of the novel identified variants and the genetic interactions and modifier effects of different genes in the neuropathy disease network. We observed that a fraction of non-neuropathy affected control individuals can carry an excess of rare "carrier" variants in recessive neuropathy associated genes which perhaps in the presence of further genetic or environmental factors can contribute to the presentation of complex neuropathy disease. Our findings suggest that rare variants can contribute susceptibility to DSP and other common neuropathies beyond the known Mendelian forms.

## 74

**Genome-wide association study identifies common variants associated with general and MMR vaccine-related febrile seizures.** *B. Feenstra[1], B. Pasternak[1], F. Geller[1], L. Carstensen[1], T. Wang[2], F. Huang[2], J.L. Eitson[3], M.V. Hollegaard[4], H. Svanström[1], M. Vestergaard[5], D.M. Hougaard[4], J.W. Schoggins[3], L.Y. Jan[2], M. Melbye[1], A. Hviid[1,6].* 1) Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark; 2) Departments of Physiology, Biochemistry and Biophysics, University of California, San Francisco, Howard Hughes Medical Institute, San Fransisco, CA; 3) Department of Microbiology, University of Texas Southwestern Medical School, Dallas, TX; 4) Danish Centre for Neonatal Screening, Department of Clinical Biochemistry, Immunology and Genetics, Statens Serum Institut, Copenhagen, Denmark; 5) Research Unit and Section for General Practice, Department of Public Health, Aarhus University, Aarhus, Denmark; 6) Department of Medicine, Stanford University School of Medicine, Stanford, CA.

Fever is a common reaction to immunization, and febrile seizures (FS) occasionally occur after vaccination, especially with live-virus vaccines such as the measles, mumps, and rubella (MMR) vaccine. Overall, FS occur in 2-5% of children before 5 years of age, and it has been estimated that 3 to 16 FS cases per 10,000 children can be attributed to MMR vaccination. To identify genetic risk factors associated with MMR-related FS and with FS in general, we conducted a series of genome-wide association scans based on samples in the Danish National Biobank comparing 929 children with MMR-related FS, 1,070 children with FS unrelated to vaccination, and 4,118 controls with no history of FS. After replication genotyping in an additional set of 408 children with MMR-related FS, 1,035 children with MMR-unrelated FS, and 1,647 controls, six independent genetic loci were associated with $P < 5\times10^{-8}$ in one or more of the analyses. Two loci were distinctly associated with MMR-related FS, harbouring the interferon-stimulated gene *IFI44L* (odds ratio (OR) = 1.41, $P = 5.9\times10^{-12}$ vs. controls; OR = 1.42, $P = 1.2\times10^{-9}$ vs. MMR-unrelated FS) and the measles virus receptor *CD46* (OR = 1.43, $P = 9.6\times10^{-11}$ vs. controls; OR = 1.48, $P = 1.6\times10^{-9}$ vs. MMR-unrelated FS). As compared with controls, four loci were associated with FS in general implicating two sodium channel genes (*SCN1A*, OR = 1.34, $P = 2.2\times10^{-16}$ and *SCN2A*, OR = 1.22, $P = 3.1\times10^{-10}$), a TMEM16 family gene (*TMEM16C*, OR = 2.09, $P = 3.7\times10^{-20}$), and a region associated with serum magnesium levels (12q21.33, OR = 1.25, $P = 3.4\times10^{-11}$). To further investigate the functional relevance of *TMEM16*, we conducted a follow-up study in an animal knockout model. Electrophysiological recordings in brain slices from wild-type and knockout rats revealed that in the absence of *Tmem16C*, hypothalamic neurons were less responsive to heat, which could lead to impaired homeostatic control when body temperature rises, and hippocampal neurons became hyperexcitable, which could possibly contribute to FS genesis. In conclusion, these findings provide new insights into genetic mechanisms and biological pathways associated with FS occurring as an adverse event following MMR vaccination and in general.

## 75

**A genome-wide meta-analysis of migraine in more than 59,000 cases and 313,000 controls reveals 29 new loci, increasing the total number of risk loci to 42.** *P. Gormley[1,2], V. Anttila[2,3], M. Muona[4], A. Palotie[1,2,4] on behalf of the International Headache Genetics Consortium (IHGC).* 1) Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA; 2) Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 3) Analytical and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA; 4) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.

Migraine is a complex disorder of neurogenic and/or neurovascular origin affecting around 14% of the population worldwide. It is characterized by chronic, recurrent headaches, often with related symptoms including nausea, photophobia, and phonophobia. One third of cases experience a sensory aura just prior to the headache, usually a visual phenomenon known as scintillating scotoma but other forms also exist. Family and twin studies estimate high heritability for migraine, pointing to a strong genetic component of the disease. Despite this, relatively little is known about the molecular mechanisms that contribute to migraine pathophysiology. Understanding has been limited because, to date, only 13 genome-wide significant risk loci have been identified for common migraine (Nature Genetics, 2013, 45:912). In this study we aimed to increase our understanding by carrying out a large-scale meta-analysis of migraine, consisting of 59,043 cases and 313,781 controls collected from 6 tertiary headache clinics and 26 population-based cohorts from multiple locations throughout the world. We performed a fixed-effects meta-analysis of genome-wide association data in 32 cohorts from the International Headache Genetics Consortium (IHGC) after imputing missing genotypes in each dataset to a common 1000 Genomes reference panel (March 2012 release, Phase I, v3). Our primary analysis investigated all general forms of common migraine, including self-reported cases and those diagnosed according to International Headache Society (IHS) diagnostic criteria. For the common migraine phenotypes we now implicate 42 risk loci in total, replicating 10 of the 13 previously reported associations and detect 29 new loci at levels of genome-wide significance. In follow-up analyses we investigated migraine sub-types and found 7 out of the 42 loci could be associated specifically to migraine without aura. No loci were robustly associated to migraine with aura, reinforcing the suggestion that rare variation plays more of a role in this form of the disease. In summary, we report 29 new risk loci and increase the total number of genome-wide significant loci to 42 for the common forms of migraine. Preliminary results point to a role for many molecular processes, including cell-signaling and ion-channel dysfunction, but also implicate several genes with previous vascular-disease associations, thus supporting a role for both neuropathic and neurovascular mechanisms in migraine pathophysiology.

## 76

**Brain somatic mutations cause focal cortical dysplasia typeII in human and mouse.** *J.S. Lim[1], W.I. Kim[1], H.C. Kang[2], S.H. Kim[3], A.H. Park[4], S. Kim[5], D. Kim[4], D.S. Kim[6], J.H. Lee[1].* 1) Graduate School of Medical Science and Engineering, KAIST, Daejeon 305-701, Korea; 2) Department of Pediatrics, Pediatric Epilepsy Clinics, Severance Children's Hospital, Epilepsy Research Center, Yonsei University College of Medicine, Seoul, Korea; 3) Department of Pathology, Brain Korea 21 project for medical science, Yonsei University College of Medicine, Seoul, Korea; 4) Department of Biological Sciences, KAIST, Daejeon 305-701, Korea; 5) Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul, Korea; 6) Pediatric Neurosurgery, Severance Children's Hospital, Department of Neurosurgery, Yonsei University College of Medicine, Seoul, Korea.

Focal cortical dysplasia type II (FCDII) is a developmental malformation of cerebral cortex and an important cause of medically refractory epilepsy. FCDII is characterized by dysmorphic neurons interspersed with normal cells and disrupted cortical lamination in affected regions. In addition, FCD sporadically occur and show funnel-shape appearance on neuroimaging, implying that dispersed abnormal neurons are derived from progenitors at the ventricular zone. These findings suggested that FCD is caused by somatic mosaic mutations only in affected brain regions. However, no such mutations have been identified. Here, we report de novo somatic mutations of MTOR in the affected brains of FCDII patients. Deep whole exome sequencing (the median read depth of 492×) of paired brain-blood DNA from 4 FCDII patients revealed brain somatic mutations in 3 patients including MTOR c.4448G>A (p.Cys1483Tyr), MTOR c.7255G>A (p.Glu2419Lys) and c.7280T>C (p.Leu2427Pro). We also performed deep targeted sequencing (the median read depth of 135,424×) of the codons encoding mTOR p.Cys1483, pGlu2419, and p.Leu2427 residues in brain tissues obtained from an additional 76 FCDII patients. In total, We identified 13 FCDII patients carrying somatic missense mutations in MTOR including mTOR p.Cys1483Tyr or Arg, pGlu2419Lys or Gly , and p.Leu2427Pro or Gln, accounting for 16.3% of all FCDII participants (13 of 80). The prevalence of the mutant allele in affected brain tissues ranged from 1.0% to 12.6%. In immunoblotting and kinase assay, the identified mutations induced the constitutive activation of mTOR kinase. Next, to determine whether the FCDII patients had the mTOR hyperactivation, we performed immunostaining in brain tissues sections obtained from FCDII patients carrying identified mutations. The results showed a marked increase in the number and size of neuronal cells positive for phosphorylated S6 in FCDII patients carrying MTOR mutations, not in non-FCD patient. Furthermore, the focal cortical expression of MTOR mutants in in utero electroporated mice was sufficient to interfere with proper neuronal migration and cause spontaneous seizures and cytomegalic neurons. Therefore, this study provides the first evidence that brain somatic mutations in MTOR cause focal cortical dysplasia.

## 77

**Parallel Studies in Humans and Dogs Implicate *ADAMTS20* in Cleft Lip and Palate Formation.** *Z. Wolf[1], B. Arzi[2], E. Leslie[3], J. Shaffer[4], H. Brand[3,4], C. Willet[5], N. Karmi[1], T. McHenry[3], E. Feingold[4], X. Wang[3], J. Murray[6], M. Marazita[3,7], C. Wade[5], D. Bannasch[1].* 1) Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, Davis, CA; 2) Department of Surgical and Radiological Sciences, School of Veterinary Medicine, University of California, Davis, Davis, CA; 3) Center for Craniofacial and Dental Genetics, Department of Oral Biology, University of Pittsburgh School of Dental Medicine, Pittsburgh, PA; 4) Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA; 5) Faculty of Veterinary Science, University of Sydney, Sydney, New South Wales, Australia; 6) Department of Pediatrics, University of Iowa, Iowa City, IA; 7) Clinical and Translational Science and Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA.

Cleft lip with or without cleft palate (CL/P) and cleft palate (CP) are commonly occurring birth defects in people. In order to better elucidate the molecular mechanisms responsible for formation of these birth defects, we utilized the domestic dog. Much like human orofacial clefts, CL/P and CP in dogs are naturally occurring, genetically heterogeneous, and follow the highly conserved steps of palatal development. Notably, the population structure of purebred dogs provides a powerful resource to study a complex trait on a simple genetic background. Previous work in dogs investigating *DLX5* and *DLX6* as candidate genes identified a LINE-1 insertion within *DLX6* in 12 CP cases from the Nova Scotia Duck Tolling Retriever (NSDTR) breed. This prompted the sequencing of the same candidate genes in a cohort of 197 humans with CL/P or CP, where a missense mutation in the highly conserved functional domain of *DLX5* was identified in one of 30 patients with Pierre Robin Sequence (Wolf, Leslie et al. 2014). Since *DLX5/6* only explain one human and 12 of 22 NSDTRs with orofacial clefts, we performed a genome-wide association study (GWAS) in 7 CL/P cases and 112 controls within NSDTRs to identify additional genes involved in orofacial cleft formation. This GWAS identified an associated region on canine chromosome 27 (9.29 - 10.73 Mb) and whole-genome sequencing of 3 cases and 4 controls identified a frameshift mutation within *ADAMTS20* (c.1360_1361delAA (p.Lys453Ilefs*3)) that is predicted to truncate 75% of the functional protein. A parallel study in a human cohort of 937 Guatemalans (545 from CL/P case families and 392 controls) also identified allelic (p =2.67 x 10-6) and gene-level (p=5.3 x 10-5) associations within *ADAMTS20*, further implicating *ADAMTS20* in CL/P formation. *DLX5/6* and *ADAMTS20* only explain 20 of the 22 NSDTR cases, leaving additional loci to be identified. Overall, these results demonstrate the power of the canine animal model as a genetically tractable approach to understanding naturally occurring and heterogeneous birth defects in humans.

## 78

**Identification of a novel susceptibility locus for nonsyndromic cleft lip and palate at chromosome 15q13.** *K.U. Ludwig[1,2], A.C. Boehmer[1,2], H. Peters[3], D. Graf[4], P. Gültepe[1,2], P.A. Mossey[5], R.P. Steegers-Theunissen[6,7], M. Rubini[8], M.M. Nöthen[1,2], M. Knapp[9], E. Mangold[1].* 1) Institute of Human Genetics, University of Bonn, Bonn, Germany; 2) Department of Genomics, University of Bonn, Bonn, Germany; 3) Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK; 4) School of Dentistry, University of Alberta, Edmonton, Canada; 5) Orthodontic Unit, Dental Hospital & School, University of Dundee, Dundee, UK; 6) Erasmus Medical Center, University Medical Center, Rotterdam, The Netherlands; 7) Radboud University Medical Center, Nijmegen, The Netherlands; 8) Department of Biomedical and Specialty Surgical Sciences, Medical Genetics Unit, University of Ferrara, Ferrara, Italy; 9) Institute of Medical Biometry Informatics and Epidemiology, University of Bonn, Bonn, Germany.

Nonsyndromic cleft lip with or without cleft palate (nsCL/P) is one of the most common congenital malformations worldwide. The etiology of nsCL/P is multifactorial. There is a considerable phenotypic variability associated with the trait, the main subgroups being nonsyndromic cleft lip only (nsCLO) and nonsyndromic cleft lip with cleft palate (nsCLP). Data from both epidemiology and embryology suggest that genetic factors might contribute to this variability. In the last years, genome-wide studies have identified 15 nsCL/P susceptibility loci explaining about 20% of nsCL/P heritability. In our recent meta-analysis (Ludwig et al. 2012, Nature Genetics) we observed a number of variants in the range of $10^{-05}>P>5\times10^{-08}$, and some of these might be true genetic risk loci. As lack of power could explain why true risk loci fail to reach genome-wide significance, we investigated some of the variants in an independent European trio cohort (n=793). One variant, rs1258763 on chr. 15q13 ($P=1.81\times10^{-06}$), was replicated in that sample ($P=0.038$), resulting in $P=4.83\times10^{-07}$ in the combined analysis. Integration of subgroup-information on nsCLO or nsCLP further decreased that *P*-value, now reaching genome-wide significance in the nsCLP group ($P=1.04\times10^{-08}$, odds ratio = 1.38 per allele). The top variant rs1258763 has previously been shown to be associated with an increased nose width (Boehringer et al. 2011, EJHG, Liu et al. 2012, PloS Genetics), suggesting that it also contributes to human facial variation. The associated region maps intergenically, between the Gremlin-1 (*GREM1*) and Formin-1 (*FMN1*) genes. GREM1 is a known antagonist of the BMP4 pathway which is relevant to craniofacial genesis. Sequencing the *GREM1*coding region plus UTR in about 400 individuals revealed a significant overrepresentation of rare variants within patients ($P=0.02$). Based on analyses of the murine *Grem1* expression pattern during embryonic craniofacial development, we further subdivided our nsCLP patient cohort and identified a strong increase in the odds ratio, which reached 3.76 per allele in a group of patients with a specific, and clinically relevant subtype of nsCLP. Notably, in four of six multiplex-families with rare mutations in *GREM1*, the rare allele also cosegregated with this particular clinical subtype. Our results demonstrate that the combination of increasing sample sizes and analysing sub-phenotypes might help to identify further risk loci for genetically complex traits.

## 79

**Variants in developmental genes confer risk of hypospadias.** *F. Geller[1], B. Feenstra[1], L. Carstensen[1], T.H. Pers[2,3,4], I.A.L.M. van Rooij[5], I.B. Körberg[6], S. Choudhry[7], J. Karjalainen[8], T.H. Schnack[1], M.V. Hollegaard[9], W.F.J. Feitz[10], N. Roeleveld[5,11], D.M. Hougaard[9], J.N. Hirschhorn[2,3,12], L.S. Baskin[7], A. Nordenskjöld[6], L.F.M. van der Zanden[5], M. Melbye[1,13].* 1) Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark; 2) Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital,Boston, USA; 3) Medical and Population Genetics Program, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, USA; 4) Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark; 5) Department for Health Evidence, Radboud university medical center, Nijmegen, The Netherlands; 6) Department of Women's and Children's Health and Center of Molecular Medicine, Karolinska Institutet, Stockholm, Sweden; 7) Department of Urology, University of California, San Francisco, USA; 8) Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands; 9) Danish Centre for Neonatal Screening, Department of Clinical Biochemistry, Immunology and Genetics, Statens Serum Institut, Copenhagen, Denmark; 10) Department of Urology, Pediatric Urology, Amalia Children's Hospital, Radboud university medical center, Nijmegen, The Netherlands; 11) Department of Pediatrics, Amalia Children's Hospital, Radboud university medical center, Nijmegen, The Netherlands; 12) Department of Genetics, Harvard Medical School, Boston, USA; 13) Department of Medicine, Stanford School of Medicine, Stanford, USA.

Hypospadias is a common congenital condition where the urethra opens on the underside of the penis. Family studies have indicated a strong genetic component in the etiology of hypospadias. We performed a genome-wide association study on 1,006 surgery-confirmed cases and 5,486 controls from Denmark. After replication genotyping of additional 1,972 cases and 1,812 controls from Denmark, the Netherlands and Sweden, 18 genomic regions showed independent association with $P < 5\times10^{-8}$; four more regions were associated with $P < 1\times10^{-6}$. Together these loci explain 9.4% of the liability to this malformation. Investigating the potential of the 548,642 genotyped SNPs resulted in an overall estimate of 56.9% of the variance explained. It was striking that several of the identified loci point to genes with roles in embryonic development, including four loci close to different members of the homeobox gene family (*HOXA* cluster, *IRX5*, *IRX6*, *ZFHX3*). Another strong candidate is *EYA1*, because it is known that deletion of *Eya1* in mice is associated with multiple genitourinary tract defects including severe hypospadias. The associated loci near *ADK* and *EEFSEC* are connected to GWAS findings for tooth development and menarche, respectively, suggesting that these genes remain important after embryogenesis. Given these connections, we decided to perform comprehensive pathway analyses with GRAIL and DEPICT. The GRAIL analysis confirmed that multiple genes in different associated regions are functionally connected. We performed DEPICT analyses for all autosomal loci with $P < 1\times10^{-5}$ in the GWA scan. In a tissue cell type enrichment analysis, the three categories with the lowest *P* were particularly relevant for hypospadias: "mesenchymal stem cells" develop into "stromal cells" and "fibroblasts" are among the most common stromal cells, playing a key role in closing the urethral groove. Analyzing physiological systems gave significant results for the urogenital and the musculoskeletal system, warranting further study of genes with potential roles in both skeletal and urogenital development in embryos. Finally, a gene set analysis resulted in a large number of sets associated with development, morphology and abnormal growth showing significant enrichment. Overall, our study provides valuable insight into the genetic architecture of hypospadias by identifying many new risk loci and connecting nearby genes in developmental pathways that could also be important for other conditions.

## 80

**Increased frequency of de novo predicted deleterious variants in non-isolated congenital diaphragmatic hernia.** *L. Yu[1], A. Sawle[2], J. Wynn[1], G. Aspelund[3], C. Stolar[4], M. Arkovitz[5], D. Potoka[6], K. Azarow[7], G. Mychaliska[8], Y. Shen[2], W. Chung[1].* 1) Division of Molecular Genetics, Department of Pediatrics, Columbia University Medical Center, New York, NY 10032, USA; 2) Departments of Systems Biology and Biomedical Informatics, Columbia University Medical Center, New York, NY 10032, USA; 3) Department of Surgery, Columbia University Medical Center, New York, NY 10032, USA; 4) California Pediatric Surgery Group, Santa Barbara, California 93105 USA; 5) Division of Pediatric Surgery, Tel Hashomer Medical Center, Tel Hashomer, Israel; 6) Department of Pediatric Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA; 7) Pediatric Surgery Division, Department of Surgery, Oregon Health Science University, Portland, OR 97239, USA; 8) Section of Pediatric Surgery, Department of Surgery, University of Michigan Health System, Ann Arbor, MI 48109, USA.

Congenital diaphragmatic hernia (CDH) is a serious birth defect that accounts for 8% of all major birth anomalies. Approximately 40% of CDH occurs in association with other anomalies (non-isolated CDH). We hypothesized that de novo variants would account for a significant fraction of sporadic non-isolated CDH since this likely has a significant effect on reproductive fitness. We performed exome sequencing in 36 non-isolated CDH trios to detect rare and de novo variants. We compared the frequency of de novo variants to 340 unaffected controls from the Simons Simplex Collection, and found that CDH patients were more likely to carry deleterious de novo variants ($p = 0.007$, OR = 1.7, 95% CI: 1.14-2.39). 20/36 of our CDH patients carry de novo deleterious variants. After accounting for the frequency of de novo variants in the control population, we estimate that 23% of sporadic non-isolated CDH patients carry CDH associated de novo variants. By investigating the protein-protein interaction network using genes with de novo predicted pathogenic variants and genes that are known to cause abnormal diaphragmatic phenotypes in mice, we found that SIN3A and MYBBP1A form an interacting network with other known CDH genes. We have identified several genes with deleterious de novo variants that fall into common categories of transcription factors and cell migration which have been implicated in the pathogenesis of CDH. These data provide evidence for novel genes in the pathogenesis of CDH associated with other anomalies and suggest that de novo variants are more common compared to the general population.

## 81

**A mutation in transferrin receptor 1 that disrupts iron internalization causes a novel immunodeficiency.** *S.E. Boyden[1,2], H.H. Jabara[3,4], J. Chou[3,4], N. Ramesh[3,4], M.J. Massaad[3,4], L. Notarangelo[3,4], M.D. Fleming[5], W. Al-Herz[6], L.M. Kunkel[2,4], R.S. Geha[3,4].* 1) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 2) Division of Genetics and Genomics, and the Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, and Department of Genetics, Harvard Medical School, Boston, MA; 3) Division of Immunology, Boston Children's Hospital, Boston, MA; 4) Department of Pediatrics, Harvard Medical School, Boston, MA; 5) Department of Pathology, Boston Children's Hospital and Harvard Medical School, Boston, MA; 6) Department of Pediatrics, Faculty of Medicine, Kuwait University, Kuwait.

Fourteen patients in a consanguineous Kuwaiti pedigree suffered from a previously undescribed syndrome of combined immunodeficiency, intermittent neutropenia, thrombocytopenia, and mild anemia (CINTA). Patients had normal numbers of T and B cells but impaired T and B cell proliferation and immunoglobulin production, resulting in agammaglobulinemia and recurrent sinopulmonary infections. Several patients died prior to the recognition of a genetic immunodeficiency in the family; subsequently, patients were successfully treated with hematopoietic stem cell transplantation. Linkage analysis and whole genome sequencing revealed patients were homozygous for a c.58T>C missense mutation in TFRC, which encodes transferrin receptor 1 (TFR1), resulting in a p.Y20H substitution. The mutation co-segregated with the CINTA phenotype, altered a perfectly conserved residue, and is absent from variant databases and 396 ancestry-matched control subjects from Kuwait. In one obligate carrier, the mutation was the only heterozygous variant within an extended homozygous background, suggesting a recent de novo origin for the mutation and segregation within the family of both mutant and non-mutant versions of otherwise identical haplotypes. The p.Y20H mutation disrupted the internalization motif critical for TFR1 endocytosis. Correspondingly, TFR1 surface expression was markedly increased and TFR1 internalization was decreased in patient lymphocytes. Transduction of wild-type but not mutant TFRC restored transferrin uptake by patient fibroblasts, and supersaturation of transferrin by iron citrate in vitro, allowing transferrin-independent iron uptake, fully rescued patient T and B cell defects. In contrast to lymphocytes, patient erythroblasts had only a modest increase in surface TFR1 expression. The metalloreductase STEAP3, which possesses an internalization sequence similar to that of TFR1, is selectively expressed in erythroblasts and interacted with TFR1. Overexpression of murine Steap3, but not of an internalization-defective Steap3 mutant, partially rescued transferrin uptake in patient fibroblasts, suggesting that STEAP3 could provide an accessory TFR1 endocytosis signal in erythroblasts that spares patients from severe anemia. Disruption of cellular iron transport represents a novel pathogenic mechanism for primary immunodeficiencies.

## 82

***TRNT1* missense mutations define an autoinflammatory disease characterized by recurrent fever, severe anemia, and B-cell immunodeficiency.** *M. Stoffels[1], Q. Zhou[1], A. Giannelou[2], D. Stone[1], A. Sediva[3], S. Rosenzweig[4], J. Edwan[2], M. Pelletier[2], K. Bishop[5], B. Carrington[5], R. Sood[5], E.F. Remmers[1], K. Barron[6], I. Aksentijevich[1], D.L. Kastner[1].* 1) Inflammatory Disease Section, NHGRI, Bethesda, MD., USA; 2) National Institute for Arthritis and Musculoskeletal and Skin Diseases, Bethesda, MD, USA; 3) University Hospital Motol, Prague, Czech Republic; 4) National Institutes of Health, Bethesda, MD, USA; 5) Zebrafish Core, NHGRI, Bethesda, MD, USA; 6) National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA.

We observed a syndrome characterized by recurrent fever, severe anemia, gastrointestinal symptoms, and a spectrum of immunologic and neurologic symptoms in five children from four unrelated families. Neurologic manifestations ranged from mild developmental delay to nystagmus, spasticity, optic nerve atrophy, and sensorineural hearing loss. Sideroblastic anemia was identified by bone marrow biopsies in two of the children. We performed whole-exome sequencing in three unrelated families. After filtering for novel and rare variants (allele frequency <1:1000) and homozygous recessive inheritance in the first two families, we observed that the patients carried missense mutations in *TRNT1*, encoding tRNA nucleotidyl transferase, CCA-adding, 1. By additional exome and Sanger sequencing we found two other patients with mutations in *TRNT1*. The TRNT1 enzyme catalyzes the addition of the CCA terminus to the 3-prime end of tRNA precursors. All disease associated mutations affect highly conserved amino acid residues and are predicted to be damaging to the protein function. The first consanguineous family from Saudi Arabia had two affected daughters, both homozygous for the p.H215R mutation; the second family of mixed Czech and British background had one affected son, carrying a compound heterozygous p.I223T/p.D163V mutation; two families of mixed European ancestry from the US each had one affected daughter, both compound heterozygous for a p.R99W/p.D163V mutation. Two out of five patients died. The p.H215R mutation was not found in any public database or in 1061 Arabian control DNA samples. The three Caucasian mutations are either novel, or found at a very low allele frequency, consistent with recessive inheritance. Cytokine profiling revealed increased IL-6 serum levels in two out of three tested patients. We observed impaired maturation of $CD10^{-}CD20^{+}$ B cells in bone marrow aspirates. Knockdown of the zebrafish *TRNT1* homologue caused hydrocephaly, defects in tail development, anemia, and a reduction in the number of hair cells present in the lateral line, which subserves functions of the inner ear in zebrafish. In conclusion, our data demonstrate that missense mutations in *TRNT1* are associated with an autoinflammatory disease manifesting with fevers, transfusion dependent anemia, gastrointestinal symptoms, immunologic, and neurologic features.

## 83

**COPA mutations disrupt intracellular transport and cause a novel autoimmune syndrome characterized by chronic pulmonary disease with pulmonary hemorrhages.** *W. Wiszniewski[1], L.B. Watkin[2], B. Jessen[3], T. Vece[2], L. Forbes[2], C. Gonzaga-Jauregui[1], S.N. Jhangiani[4], D.M. Muzny[4], E. Boerwinkle[5], R.A. Gibbs[4], A. Shum[3], J. Orange[2], J.R. Lupski[1].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Texas Children's Hospital, Houston, TX; 3) UCSF School of Medicine, San Francisco, CA; 4) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 5) Human Genetics Center and Institute of Molecular Medicine, University of Texas-Houston Health Science Center, Houston, TX,.

Autoimmune disorders are a group of clinically heterogeneous diseases that result from dysregulation of immunologic mechanisms. The genetic contribution to autoimmune disease ranges from simple Mendelian inheritance of causative alleles to the complex interactions of multiple weak loci influencing risk. Despite the rapid advances in genome-wide genetic analysis, substantial components of the heritable risk remain unexplained. Here, we report a novel hereditary autoimmune syndrome where affected individuals develop chronic pulmonary disease with pulmonary hemorrhages and other autoinflamatory manifestations including arthritis and nephropathy. We performed whole-exome sequencing in five affected subjects from three large pedigrees. We identified missense mutations in COPA (OMIM# 601924), a novel gene encoding a cargo protein involved in intracellular protein transport in the affected members of all three families. COPA is a component of the COPI coatomer complex that is important for retrograde Golgi to ER transport. All mutations were localized to a functionally important WD40 domain, predicted to be pathogenic by available bioinformatics tools and shown to segregate with the disease phenotype in multi generation families. These predictions were further supported by results of functional studies that demonstrated evidence of i) increase in ER stress, ii) abnormal size of endolysosomes, and iii) impaired autophagy. These latter cellular abnormalities have been described in other autoimmune diseases. In conclusion we demonstrated that abnormal intracellular transport caused by specific COPA mutations may trigger increased ER stress and impaired autophagy that leads to dysregulation of immunologic mechanisms and secondary autoimmune disorder affecting the lung and joints.

## 84

**Mendelian genetic studies of immune disorders to identify novel targets for therapeutic intervention.** *J. McElwee[1], X. Chen[1], J. Lyons[2], G. Sun[2], X. Yu[2], J. Milner[2], Y. Liuv[3], Z. Deng[3], A. Almeida de Jesus[3], R. Goldbach-Mansky[3], Y. Zhang[4], H. Matthews[4], H. Su[4], M. Lenardo[4].* 1) Merck Research Laboratories, Department of Genetics & Pharmacogenomics (GpGx), Boston, MA; 2) Laboratory of Allergic Diseases, NIAID, NIH, Bethesda, MD; 3) Translational Autoinflammatory Disease Section, NIAMS, NIH, Bethesda, MD; 4) Laboratory of Immunology, NIAID, NIH, Bethesda, MD.

Mendelian disorders that selectively affect the immune system provide a means to identify and understand new genes and pathways involved in human immunology which could represent new routes for therapeutic intervention for human disease. As part of a collaborative effort to identify novel therapeutic opportunities in immunobiology, Merck and the NIH Clinical Center have undertaken a focused study of Mendelian forms of primary immunodeficiencies, syndromic atopic disease, and severe auto-inflammatory conditions using next generation sequencing approaches. We have sequenced 711 exomes representing over 200 independent families exhibiting a range of suspected Mendelian immune diseases, including autoimmune lymphoproliferative syndrome, chronic active EBV infection, atopic disorders (such as familial tryptasemia, atopic dermatitis, or severe asthma), and diverse forms of primary immunodeficiencies or severe auto-inflammatory diseases. Analysis of these data using Mendelian filtering techniques, further supported by extensive molecular and cellular validation studies, has led to the successful identification of causal mutations in known and novel immune genes for over 25% of kindreds that have been analyzed to date. We will present highlights from the analysis of this unique cohort, including several newly-identified disease-causing genes. First we will describe a new form of immunodeficiency and atopic disease caused by loss-of-function mutations in the phosphoglucomutase-3 gene (*PGM3*). Affected individuals from 2 families exhibit severe atopy with marked serum IgE level increases, recurrent bacterial and viral infections, and neurocognitive impairment. The causal mutations in *PGM3* affect an enzyme crucial in the generation of UDP-GlcNAc, pointing to an underappreciated role for glycosylation in the regulation of immune cell function. Next, we will describe a new auto-inflammatory disorder identified in 6 unrelated patients exhibiting severe skin vasculitis, chronic interstitial lung disease and systemic inflammation. Analysis of these families has identified *de novo* gain-of-function mutations in the ds-DNA sensor STING (encoded by *TMEM173*), a key adapter molecule that links innate nucleic acid sensing to IFN pathway activation. Finally, we will present an overview of our experience and success rate for Mendelian genetic studies within this cohort, and future plans for analysis and expansion of this collaborative discovery effort.

## 85

**The Human Knockout Project: systematic discovery of loss-of-function variants in humans.** *K.J. Karczewski[1,2], V. Narasimhan[3], M. Lek[1,2], M. Rivas[4], S. Balasubramanian[5], M. Gerstein[5], B. Keating[6], T. Lappalainen[7], A. Palotie[1,2], M. Daly[1,2], D. van Heel[8], R. Trembath[8], R. Durbin[3], D.G. MacArthur[1,2].* 1) Massachusetts General Hospital, Boston, MA; 2) Broad Institute, Cambridge, MA; 3) Wellcome Trust Sanger Institute, Hinxton, UK; 4) Wellcome Trust Centre for Human Genetics Research, University of Oxford, Oxford, UK; 5) Yale University, New Haven, CT; 6) University of Pennsylvania, Philadelphia, PA; 7) New York Genome Center, New York, NY; 8) Queen Mary University of London, London, UK.

Every human carries at least a hundred loss-of-function (LoF) variants predicted to severely disrupt the function of protein-coding genes, including many in the homozygous state. These variants represent "experiments of nature" that can cast light on the function of currently uncharacterized human genes: indeed, much novel biology has already been learned from the involvement of rare LoF variants in severe Mendelian disease. Additionally, these variants have also proved valuable in identifying potential therapeutic targets: LoF variants in PCSK9 have been causally linked to low LDL cholesterol levels, leading to the development of PCSK9 as a therapeutic target for cardiovascular disease. However, discovering LoFs in the human population remains a significant challenge, as these variants are enriched for sequencing and annotation errors, and typically have very low frequency, confounding their discovery and interpretation. For this reason, large sample sizes are required to discover LoFs in every possible gene. Alternatively, two distinct strategies, the use of bottlenecked populations and populations with a high rate of consanguineous mating, are established to significantly enrich for discovery of homozygous rare LoF variants, effectively identifying "knockout" humans. To this end, we have developed an open-source tool, LOFTEE (Loss Of Function Transcript Effect Estimator), to annotate loss-of-function variants. In order to characterize the landscape of homozygous LoF variants (knockouts) across humans, we have applied LOFTEE to a number of large datasets from collaborative efforts, including over 91,000 exomes aggregated from a variety of rare and complex disease consortia, deeply phenotyped samples from Finnish national biobanks, and over 1,000 parentally-related individuals from the UK. We validate these methods using databases of known disease variants, and investigate the role of LoF variants on splicing and gene expression by intersecting exomes with matched RNA-Seq data from over 500 individuals from the GTEx consortium. Finally, we describe the aggregation of these variants into a database of LoF variants, dbLoF, providing a resource for pharmaceutical development, transplant biology, and understanding of rare Mendelian diseases.

## 86

**Using compressed data structures to capture variation in thousands of human genomes.** *S.A. McCarthy[1], Z. Lui[1], J.T. Simpson[2], Z. Iqbal[3], T.M. Keane[1], R. Durbin[1].* 1) Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom; 2) Ontario Institute for Cancer Research, Toronto, Ontario, Canada; 3) Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom.

Currently the most widely used approach to catalogue variation amongst a set of samples is to align the sequencing reads to a single linear reference genome. This principle has been at the core of the 1000 Genomes data processing pipeline since the pilot phase of the project. However, there is now an increased awareness of the limitations of this approach, such as alignment artefacts, reference bias and unobserved variation on non-reference haplotypes. The Burrows-Wheeler transform and FM-index are compact data structures that have been successfully used in sequence alignment and assembly. One of the key features of these structures is that they are a searchable and reference-free representation of the raw sequencing reads. Our project aims to build a web server based on BWT data structures containing all the reads from many thousands of samples so as to efficiently retrieve matching reads and information about samples and populations. Enticingly, it is expected that data storage for this system would plateau as we collect more data since most new sequencing reads will have already been observed. We expect this to enable powerful new ways to query variation data from thousands of individuals. For the first phase of this project, we include all 87 Tbp of the low-coverage and exome data from the 2,535 samples in 1000 Genomes Phase 3. We envisage this would provide a means for researchers to easily check the prevalence of any human sequence in a control set of thousands of putatively healthy samples. We present our approaches and initial benchmarks on variant sensitivity and specificity against truth datasets and explore several applications for these structures such as validation of short insertion/deletion and structural variant calls, and rapid searching for traces of viral DNA.

## 87

**Exploring Genetic Variation and Genotypes Among Millions of Genomes.** *R.M. Layer, A.R. Quinlann.* University of Virginia, Charlottesville, VA.

Rare, and thus largely unknown, variants are a major reason that, typically, less than 10% of the heritability of complex diseases currently can be explained by known genetic variation. While increasing the number of sequenced genomes may improve our ability to reveal this "hidden heritability," the scale of the resulting dataset poses substantial storage and computational demands. Current efforts to sequence 100,000 genomes, and combined efforts that are likely to surpass 1 million genomes will identify hundreds of millions to billions of polymorphic loci. The minimum storage requirement for directly representing the variability found by these projects (1 bit per individual per variant, ignoring the necessary metadata) will range from terabytes to petabytes. Like most big-data problems, a balance must be found between optimizing storage and computational efficiency. For example, while compression can minimize storage by reducing file size, it can also cause inefficient computation since data must be decompressed before it can be analyzed. Conversely, highly structured data can reduce analysis times but typically require extra metadata that increase file size. Current variation storage schemes were not designed to quickly analyze massive datasets and fail to balance these competing goals. We present GENOTQ, an open source API and toolkit that reduces file size and data access time through use of a succinct data structure, a class of data structures that compress data such that operations can be performed without requiring the full decompression. Word aligned hybrid (WAH) bitmap compression is one such data structure that was developed to improve query times for relational databases. Binary values are encoded such that logical operations (AND, OR, NOT) can be performed on the compressed data. This encoding results in file sizes that are 20X smaller than uncompressed versions, and only 50% larger than the compressed version. Queries, such as finding shared variants among a subpopulation, are also 21X faster. Furthermore, representing the genotypes in this manner makes our method well suited to both distributed architectures like BigQuery and parallel processors like GPUs. We stress that this method is only part of a larger solution that would incorporate genomic annotations, medical histories, and pedigrees. Incorporating fast genotype queries with this web of metadata will provide a rich information source to both clinicians and researchers.

## 88

**Databases, genome repositories, and clinical applications to interpret personal genome for precision and preventative therapies.** *R. Chen.* Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY.

With the advance of next generation sequencing technologies, we can sequence a personal genome in a few days under one thousand dollars. Hundreds of thousands of individuals have been sequenced and released into public repositories, with millions of genomes coming in the next few years. There is an urgent need to build automated systems to interpret personal genome for clinical diagnosis, precision medicine, and preventative therapies. We have built an automated system APOLLO to process next generation sequencing data and interpret personal genome. First, we curated and analyzed hundreds of thousands of human genomes, exomes, and genotyping data and built a central variant store, which contains over 110 million distinct genetic variants with unique IDs across studies. Second, we used Hadoop and MapReduce to calculate the frequencies of these variants across hundreds of disease states and ethnic populations. Third, we built hundreds of annotation databases using text mining, manual curation, and public repositories. Fourth, we used these primary databases to annotate 110 million variants and built a secondary annotation database called ActiVar. Last, we developed a tool called VARA to integrate variant calls from multiple sequencing platforms and variant calling algorithms to identify reliable and actionable variants. We further developed a series of clinical applications to interpret personal genome and exome for precision and preventative therapies. For each cancer patient with solid tumor, we sequenced the DNA and RNA from the tumor and blood, built a decision tree for each FDA approved drug by integrating variants, fusions, CNV, and RNA, and created a clinical report to recommend personalized precision medicine, clinical trials, and immunotherapy vaccines. To search for preventative therapies, we launched the resilience project to search for "Unexpected Heroes": healthy individuals with resilience to deleterious mutations commonly leading to severe childhood diseases. We curated 674 founder or recurrent disease causing variants with extremely high penetrance from 162 genes for 125 distinct Mendelian disorders, screened 596,610 personal genome, exome, and genotyping data, filtered with Sanger confirmation and clinical review, and identified 9 final unexpected heroes. In summary, the explosion of big data has enabled clinical applications to interpret personal genome for clinical diagnosis, precision medicine, and preventative therapies.

## 89

**DbGaP Genotype Fingerprint Collection.** *Y. Jin, S. Stefanov, S. Dracheva, Z. Wang, N. Sharopova, A. Sturcke, S. Sherry, M. Feolo.* National Center for Biotechnology Information , National Library of Medicine, National Institutes of Health , Bethesda, MD., USA.

The database of Genotypes and Phenotypes (dbGaP) has accessioned more than one million samples from over 750,000 human individuals. At this scale, it is not uncommon that multiple, independent samples were collected from the same individual (or subject) for different research purposes and submitted to dbGaP under different studies. The dbGaP has established a genotype fingerprint collection to detect these cryptic duplicates in the database. Theoretically, a few dozen independent and informative SNPs are enough to uniquely determine an individual. However, since genotypes submitted to dbGaP are obtained using different methods and cover differing genomic regions, many more SNPs are needed to ensure a sufficient number of informative genotyped SNPs overlap between any two samples. We have selected 11,000 SNPs for fingerprinting using the following requirements: 1) the SNP is covered by at least 80% of the genotyping methods used by dbGaP studies; 2) the SNP is biallelic with a minimum minor allele frequency 0.17 as reported by the 1000 Genome Project; 3) the SNP is well separated from its nearest neighbor in the set with physical distance of at least 50,000 bps; 4) the SNP is not palindromic; 5) the SNP is autosomal. Non-palindromic SNPs were selected to avoid the DNA strand orientation problem. For example, if one genotype chip determines that the two alleles for a certain SNP are A/G, and another chip reports T/C for the same SNP, then we know genotype AA in the first chip is the same as TT in the second one. We have created computer programs to read genotypes from different formats, including PLINK *ped* and *bed* files, transposed datasets, and other matrix formats. To minimize the footprint of the collection, we use four binary numbers to represent the three genotypes and one missing state and store genotypes from four SNPs in one byte. We have loaded genotypes of about 600,000 samples into the fingerprint collection. We have also developed algorithms to identify duplicates quickly. Using these algorithms we have found about 70,000 pairs of cryptic duplicate samples that were collected either from the same subjects or identical twins. This presentation will introduce dbGaP genotype fingerprint collection and describe how we use it to discover sample/subject overlaps between dbGaP studies, find inconsistencies across the submitted subject-sample mapping files, pedigree files, genotype datasets, as well as estimate per study genotyping error rates.

## 90

**Integrated analysis of microRNA expression, UTR binding sites, and human variation in ocular tissues.** *T. Gaasterland[1,2], A.N. Dubinsky[3], L.E. Edsall[2,4], T.S. Mondala[5], P. Ordoukhanian[5], S.R. Head[5].* 1) Institute for Genomic Medicine, Univ California San Diego, La Jolla, CA; 2) Scripps Institution of Oceanography, Univ California San Diego, La Jolla, CA; 3) Pediatrics Department, School of Medicine, Univ California San Diego, La Jolla, CA; 4) University Program in Genetics and Genomics, Duke University, Durham, NC; 5) Next Generation Sequencing Core, The Scripps Research Institute, La Jolla, CA.

We present a platform to evaluate and rank genome variants in untranslated regions (UTR) of mRNA transcripts and their potential relevance to eye diseases through disruption of microRNA binding. We evaluate microRNA and mRNA expression in human eye tissues to identify UTR variants with the potential to disrupt microRNA regulatory activity and contribute to risk of, or progression in, glaucomatous optic nerve degeneration, a blinding eye disease affecting over 60 million people worldwide. We overlaid eye tissue gene expression patterns for microRNA and mRNA, microRNA binding sites in UTRs, and UTR variants detected through exome sequencing. We sequenced ~500 human exomes with capture probes that include UTR targets (Nimblegen SeqCap EZ Exome). 90,358 variant sites located in UTR regions were observed in at least one exome. Total RNA was extracted from optic nerve, optic disc with lamina cribrosa, retina, trabecular meshwork, ciliary body, and choroid from human donor eyes and subjected to high-throughput sequencing (Illumina, Hiseq 2000; Nugen) to measure microRNA and mRNA expression levels. MicroRNA reads were matched against the 2,555 unique, known human microRNAs (mirbase.org, v20), counted, and normalized. mRNA reads were mapped to the reference human genome, counted per gene, and normalized (genome.ucsc.edu, hg19; bowtie; cuffdiff). MicroRNAs and mRNAs were stratified by high, medium, and low expression and ranked within tissue by expression level. The first 8-9 bases of a microRNA can guide its binding to an mRNA. Generally, these seed sites start at base 2 of the microRNA and match their mRNA targets in reverse-complement. With our tool, ZoomMiR, we tabulated and scored all seed sites for the known human microRNAs in 124,315 5-prime UTR and 194,503 3-prime UTR sequences from all mRNA transcript isoforms annotated in hg19. Exome variants, microRNA and mRNA eye-tissue expression levels, and microRNA seed sites were merged based on genomic chromosome positions to identify all UTR variants within or near microRNA seed sites and detected through exome sequencing. Known variants received dbSNP identifiers and population frequencies from the 1000 Genomes and Exome Sequencing Project sites (1000genomes.org; evs.gs.washington.edu). Together, these data provide a tool to evaluate the impact of UTR variants with alternate alleles over-represented in patients with disease compared to general or control populations.

## 91

**Second-generation PLINK: rising to the challenge of larger and richer datasets.** *C.C. Chang[1,2], C.C. Chow[3], L.C.A.M. Tellier[2,4], S. Vattikuti[3], S.M. Purcell[5,6,7,8], J.J. Lee[3,9].* 1) BGI Hong Kong, 16 Dai Fu Street, Tai Po Industrial Estate, Tai Po, N.T., Hong Kong; 2) BGI Cognitive Genomics Lab, Building No. 11, Bei Shan Industrial Zone, Yantian District, Shenzhen, China 518083; 3) Mathematical Biology Section, NIDDK/LBM, National Institutes of Health, Bethesda, MD 20892; 4) Bioinformatics Centre, University of Copenhagen, 2200 Copenhagen, Denmark; 5) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142; 6) Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029; 7) Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029; 8) Analytic and Translational Genetics Unit, Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA 02114; 9) Department of Psychology, University of Minnesota Twin Cities, Minneapolis, MN 55455.

PLINK 1 is a widely used open-source C/C++ toolset for genome-wide association studies (GWAS) and research in population genetics. However, the steady accumulation of data from imputation and whole-genome sequencing studies has exposed a strong need for even faster and more scalable implementations of key functions. In addition, GWAS and population-genetic data now frequently contain probabilistic calls, phase information, and/or multiallelic variants, none of which can be represented by PLINK 1's primary data format.

To address these issues, we are developing a second-generation codebase for PLINK. The first major release from this codebase, PLINK 1.9, introduces extensive use of bit-level parallelism, O(sqrt(n))-time/constant-space Hardy-Weinberg equilibrium and Fisher's exact tests, and many other algorithmic improvements. In combination, these changes accelerate most operations by 1-4 orders of magnitude, and allow the program to handle datasets too large to fit in RAM. This will be followed by PLINK 2.0, which will introduce (a) a new data format capable of efficiently representing probabilities, phase, and multiallelic variants, and (b) extensions of many functions to account for the new types of information.

The second-generation versions of PLINK will offer dramatic improvements in performance and compatibility. For the first time, users without access to high-end computing resources can perform several essential analyses of the feature-rich and very large genetic datasets coming into use.

## 92

**Microtask crowdsourcing for annotating diseases in PubMed abstracts.** *A.I. Su, B.M. Good, M. Nanis.* Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA.

The scientific literature is massive, but only a tiny percentage of that biomedical knowledge is structured in a way that is amenable to computational analysis. Comprehensively annotating the literature in the form of concepts and relationships between concepts would result in a powerful knowledgebase for biomedical research. Although many biological natural language processing (BioNLP) projects attempt to address this challenge, the state of the art in BioNLP still leaves much room for improvement. Expert curators are vital to the process of knowledge extraction but are always in short supply.

Recent studies have shown that workers on general-purpose microtasking platforms such as Amazon's Mechanical Turk (AMT) can, in aggregate, generate high-quality annotations of biomedical text. Here, we investigated the use of the AMT in capturing disease mentions in PubMed abstracts, comparing to the NCBI Disease corpus as a gold standard. After merging the responses from 5 AMT workers per abstract with a simple voting scheme, we were able to achieve a maximum F measure of 0.815 (precision 0.823, recall 0.807) as compared to expert annotations on the same abstracts. These results can also be tuned to optimize for precision (up to 0.98) or recall (up to 0.89) by adjusting the voting procedure among AMT workers. We also found that providing AMT workers continuous feedback on performance led to improved performance over time. Finally, using AMT for biocuration had clear benefits in terms of both time and cost, requiring just 7 days and less than $200 to complete all 593 abstracts in the test corpus (at $.06/abstract).

This experiment demonstrated that microtask-based crowdsourcing can be successfully used to identify disease mentions in biomedical text. Although there is room for improvement in the crowdsourcing protocol, overall AMT workers are clearly capable of performing this annotation task. Our experience using AMT motivates two orthogonal lines of ongoing research. First, we are investigating the use of AMT to perform other biocuration tasks, including relationship extraction. Second, these results strongly suggest that Citizen Science workers have both the skill and motivation to help structure biomedical knowledge.

## 93

**Automating literature reviews: Predicting variant pathogenicity using the bibliomic index.** *C.A. Cassa[1], D.M. Jordan[2], S.R. Sunyaev[1].* 1) Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, BOSTON, MA; 2) Graduate Program in Biophysics, Harvard University, Cambridge, MA.

Clinical geneticists and researchers rely on the medical and scientific literature to interpret potential disease variants. While the GWAS community has developed stringent assessment standards to avoid false positive associations, Mendelian disease variants are typically assessed using inconsistent platforms and validation standards. Many variants are identified in small, symptomatic populations, so their effect size may be incorrect, or they may be erroneously associated with disease due to limited validation or unmatched control populations. The result is an admixture of trusted associations with unverified, or even incorrect variants.

The consequence is that clinical interpretation of these variants often requires manual review to ascertain effect size and clinical significance. This approach will not scale with the exponential growth of clinical sequencing programs, as we observe previously reported disease variants at substantial rates in sequenced individuals, many of which require manual review.

Based on the idea that there are valuable published disease associations, but that it is difficult to distinguish the importance of any specific citation, we attempt to use the literature - in aggregate, by gene - to predict the pathogenicity of individual variants. Using a large set of publications that describe disease associations (HGMD), we develop a novel statistic called the bibliomic index, which uses publication impact and citation frequency (Thompson Reuters) to predict variant pathogenicity. Using an independent dataset that is restricted to known disease genes, variants that have higher bibliomic index scores are more likely to be rare, pathogenic variants. The three features in our bibliomic index have very strong predictive value, achieving an AUC of 0.8584. This information compliments existing computational methods, which rely on structural and evolutionary factors; when combined with PolyPhen-2, we achieve an AUC of 0.9408.

This demonstrates that aggregate bibliomic data can substantially improve the current arsenal of in silico predictors, mitigating the challenges traditionally associated with the accession and parsing of manuscripts. These features may be used to prioritize and contextualize candidate disease variants in known disease genes.

## 94

**Integrated analysis of protein-coding variation in over 90,000 individuals from exome sequencing data.** *D.G. MacArthur[1,2], M. Lek[1,2,3,4], E. Banks[2], R. Poplin[2], T. Fennell[2], K. Samocha[1,2], B. Thomas[1,2], K. Karczewski[1,2], S. Purcell[1,2,5], P. Sullivan[6], S. Kathiresan[1,2], M.I. McCarthy[7], M. Boehnke[8], S. Gabriel[2], D.M. Altshuler[1,2], G. Getz[1,2], M.J. Daly[1,2], Exome Aggregation Consortium.* 1) Massachusetts General Hospital, Boston, MA., USA; 2) Broad Institute of Harvard and MIT, Cambridge, MA, USA; 3) University of Sydney, Sydney, NSW, Australia; 4) Institute for Neuroscience and Muscle Research, Sydney, NSW, Australia; 5) Mt Sinai School of Medicine, New York, NY, USA; 6) University of North Carolina, Chapel Hill, NC, USA; 7) University of Oxford, Oxford, UK; 8) University of Michigan, Ann Arbor, MI, USA.

The discovery of genetic variation has been empowered by the growing availability of DNA sequencing data from large studies of common and rare diseases, but these data are typically inconsistently processed and largely inaccessible to most genetics researchers. We have developed an efficient and scalable pipeline for the joint analysis of exome sequencing data from tens of thousands of samples and have applied it to a collection of over 90,000 individuals sequenced in diverse population genetic and disease studies. Using extensive independent validation data we demonstrate that our joint variant calling approach improves accuracy, sensitivity and consistency of rare variant detection. Our results provide an unprecedented view of the spectrum of human functional genetic variation extending down to extremely low population frequencies. We observe >8 million single nucleotide polymorphisms (SNPs), including over 3.5 million rare (<1%) missense variants and >15,000 previously reported severe disease-causing mutations. We show that the frequency spectrum of rare variants can be used to assess the accuracy of functional annotation approaches, and to identify likely misannotated disease mutations. We describe the distribution of >150,000 predicted loss-of-function variants across human genes and the functional assessment of over 1,000 of these with independent RNA sequencing data. We also demonstrate the benefits of large joint-called reference panels for identifying gene regions subject to strong functional constraint and for the discovery of rare causal variants in both complex and Mendelian diseases. Finally, we announce the public release of observed variants, population frequencies and gene-level summary statistics for a subset of over 55,000 reference exomes. These summary results are publically available via an intuitive browser.

## 95

**Identification of a large set of rare complete human knockouts.** *P. Sulem[1], H. Helgason[1,2], A. Oddson[1], H. Stefansson[1], SA. Gudjonsson[1], F. Zink[1], E. Hjartasson[1], G. Sigurdsson[1], A. Jonasdottir[1], A. Sigurdsson[1], O. Magnusson[1], A. Kong[1,2], A. Helgason[1,3], U. Thorsteinsdottir[1,4], G. Masson[1], D. Gudbjartsson[1,2], K. Stefansson[1,4].* 1) Dept Statistics, DeCode Genetics/Amgen, Reykjavik, Iceland; 2) School of engineering and natural sciences,University of Iceland.Reykjavik, Iceland; 3) Department of anthropology, University of Iceland, 101 Reykjavik, Iceland; 4) Faculty of medicine,University of Iceland,101 Reykjavik, Iceland.

Mutations that cause a loss of function (LoF) of a gene are the causes of many Mendelian diseases. We sequenced the whole genomes of 2,636 Icelanders and imputed the sequence variants identified in this set into 101,584 additional chip typed and phased Icelanders. A total of 6,795 autosomal LoF variants, 3,979 SNPs and 2,816 indels, were found in 4,924 genes. Most LoF mutations are rare, with 85% having minor allele frequency (MAF) below 0.5%. Only 41% of the LoF variants detected in Iceland are present in the Exome Sequencing Project or dbSNP. Recessive diseases are often caused by losing the function of both copies of a gene - the gene being completely knocked out. Here we focus on variants with MAF below 2%, which is the MAF of mutations causing cystic fibrosis, the most common recessive disease in Europeans. It is usually not necessary to sequence the individual affected by a recessive disease in order to observe the causative mutation since it will be present in the heterozygous state in unaffected relatives. Of the 104,220 genotyped individuals, 8,041 (7.7%) have at least one gene completely knocked out by LoF variants with MAF under 2%. These individuals are homozygotes or compound heterozygotes for LoF mutations in 1,171 genes (6.1%) and only 9.9% are children of parents who are second cousins or closer. We assessed whether a gene being completely knocked out depends on its tissue of expression and found that genes that are highly expressed in the brain are less often completely knocked out. We found that LoF mutations are less frequently observed in the offspring of heterozygous parents than Mendelian inheritance would predict. In particular, homozygous LoF offspring of two heterozygous parents are substantially fewer than expected. Detailed phenotyping of the individuals with the extreme genotype of having a gene completely knocked out offers a way of understanding the function of the completely knocked out genes.

## 96

**Making Sense Of Nonsense : consequence of premature Stop mutations.** *S. Balasubramanian[1], Y. Fu[1], M. Pawashe[1], M. Jin[1], J. Liu[1], D. MacArthur[2], M. Gerstein[1].* 1) MB & B, Yale University, New Haven, CT; 2) ATGU, Massachussets General Hospital, Boston, MA.

Loss-of-function variants (LOF) attract great clinical interest, as it is believed that most of them are potentially pathogenic. However, some LOF variants are also known to be beneficial. For example, LOF variants in PCSK9 lead to low LDL levels. Therefore, several pharmaceutical companies are actively pursuing the inhibition of PCSK9. Recent sequencing efforts demonstrate the presence of null variants in seemingly healthy people. Consequently, there is great interest in understanding putative LOF variants. While there are several methods for inferring the effect of nonsynonymous variants, there is no predictor for nonsense variants. Moreover, most methods predict pathogenicity of a variant without taking into consideration its genotype. Here, we present a method to infer the effect of LOF variants, primarily those due to premature Stop codons. We have developed ALOFT (Annotation of Loss-of-Function Transcripts), a pipeline to annotate putative LOF variants with a variety of functional and evolutionary features. Using these features, we developed a predictive model to classify nonsense variants into those that are benign, lead to recessive disease and those that lead to dominant disease. We applied this method to infer the effect of nonsense mutations in studies from the Center For Mendelian Genomics and correctly predict the mode of inheritance and pathogenicity of all of the truncating variants. We also validate our method by applying our classifier to four different autism studies. De-novo LOF SNPs have been implicated in autism. Our method shows that de-novo LOF events are significantly higher in autism cases versus controls. To our knowledge, this is the first method that predicts the impact of nonsense SNPs in the context of a diploid model, i.e. whether nonsense SNP will lead to recessive or dominant disease.

## 97

**Analysis of stop-gain and frameshift variants in human innate immunity genes.** *A. Rausell[1,2,3,4], P. Mohammadi[1,5], PJ. McLaren[1,2,6], I. Bartha[1,2,6], I. Xenarios[1,4,7], J. Fellay[1,6], A. Telenti[2,3].* 1) Swiss Institute of Bioinformatics (SIB) and University Hospital of Lausanne, Lausanne, Vaud, Switzerland; 2) Department of Laboratories, University Hospital of Lausanne, Switzerland; 3) University of Lausanne, Lausanne, Switzerland; 4) Vital-IT group, SIB Swiss Institute of Bioinformatics Lausanne, Switzerland; 5) Computational Biology Group, ETH Zurich, Switzerland; 6) School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; 7) Swiss-Prot group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland.

Loss-of-function variants in innate immunity genes are associated with Mendelian disorders in the form of primary immunodeficiencies. Recent resequencing projects report that stop-gains and frameshifts are collectively prevalent in humans and could be responsible for some of the inter-individual variability in innate immune response. Current computational approaches evaluating loss-of-function in genes carrying these variants rely on gene-level characteristics such as evolutionary conservation and functional redundancy across the genome. However, innate immunity genes represent a particular case because they are more likely to be under positive selection and duplicated. To create a ranking of severity that would be applicable to innate immunity genes we evaluated 17764 stop-gain and 13915 frameshift variants from the NHLBI Exome Sequencing Project and 1000 Genomes Project. Sequence-based features such as loss of functional domains, isoform-specific truncation and nonsense-mediated decay were found to correlate with variant allele frequency and validated with gene expression data. We integrated these features in a Bayesian classification scheme and benchmarked its use in predicting pathogenic variants against Online Mendelian Inheritance in Man (OMIM) disease stop-gains and frameshifts. The classification scheme was applied in the assessment of 335 stop-gains and 236 frameshifts affecting 227 interferon-stimulated genes. The sequence-based score ranks variants in innate immunity genes according to their potential to cause disease, and complements existing gene-based pathogenicity scores. Specifically, the sequence-based score improves measurement of functional gene impairment, discriminates across different variants in a given gene and appears particularly useful for analysis of less conserved genes.

## 98

**Insights into protein truncating variation from high-quality indel calling in 1000 UK population exomes - implications for disease gene discovery and clinical utility.** *E. Ruark[1], A. Renwick[1], E. Ramsay[1], S. Seal[1], K. Snape[1], S. Hanks[1], A. Rimmer[2], M. Munz[2], A. Elliott[1], G. Lunter[2], N. Rahman[1].* 1) Institute of Cancer Research, Sutton, United Kingdom; 2) Wellcome Trust Centre for Human Genomics, Oxford, UK.

The advent of next-generation sequencing (NGS) has provided the potential to comprehensively capture coding variation in large numbers of individuals fast and affordably. This has furthered both research to discover gene mutations that predispose to disease and opportunities to expand clinical gene testing. However, in both contexts, the spectrum of coding variation in the general population is a necessary consideration when evaluating the impact on association with disease. Many disease predisposition genes are characterized by multiple different mutations that result in premature protein truncation, termed protein truncating variants (PTV) or loss-of-function variants (LoF). The predominant mechanism for generating PTVs is base insertions or deletions (indels). Unfortunately, accurate detection of indels in short-read NGS data has proved challenging, with sub-optimal sensitivity and specificity and low concordance between callers. We have developed and validated a pipeline for exome analysis (base substitutions and indels) with 95% sensitivity and 94% specificity for indels. Application of the pipeline to 1000 UK population controls unselected with respect to disease reveals multiple insights into PTV architecture. PTVs were identified in one third of genes (5627/17588); however, for the majority (51%), one PTV in one individual was identified. Only 139 genes (0.8%) had multiple (5 or more), different PTVs in the 1000 individuals. Furthermore, these data have allowed us to describe the genetic variation of the average UK individual, who carries 22,000 coding variants of which 160 are rare (≤ 0.1% of the population). On average each person in UK has 211 PTVs of which 6 are rare (≤ 0.1%) and 91 are homozygous. The data also provide a framework for disease gene identification and clinical characterisation studies. Most importantly, we show that rare protein truncating variants are an expected part of the normal spectrum of an individual's genomic variation. As such, more caution than typically applied is appropriate in the evaluation of the likely causal link between a rare PTV and a phenotype in a given individual. Conversely, multiple different PTVs within the same gene is an unusual pattern in the general population, but a common pattern in people with genetic diseases, and may serve as a useful mutational signature in disease gene discovery studies.

## 99

**Exome sequencing of fit adults with high parental relatedness identifies over 600 rare human gene knockouts.** *V. Narasimhan[1,6], KJ. Karczewski[4,5], KA. Hunt[2], Y. Xue[1], P. Danecek[1], S. Mccarthy[1], C. Tyler-Smith[1], C. Griffiths[2], J. Wright[3], ER. Maher[6], DG. Macarthur[4,5], RC. Trembath[2], DA. van Heel[2], RM. Durbin[1].* 1) Wellcome Trust Sanger Institute, Hinxton, UK; 2) Queen Mary University of London, London, UK; 3) Bradford Institute for Health Research, Bradford, UK; 4) Massachusetts General Hospital, Boston, USA; 5) Broad Institute of MIT and Harvard, Cambridge, USA; 6) University of Cambridge, Cambridge, UK.

The majority of human genes have poorly characterized function. Naturally occurring loss of function variants (LoFs) provide insight into the phenotypic impact of gene inactivation. Homozygous LoFs are responsible for many rare diseases, but are also present in healthy adult individuals. The examination of such variants in health and disease can provide insight into human gene function. To explore the impact of homozygous LoFs on human phenotypes we have exome sequenced 2,625 individuals, ascertained as fit adults, with a range of parental relatedness. We developed an algorithm to detect autozygous stretches and observed 165 (6%) samples with >12.5% autozygosity (expected of double first cousin relatedness) and an additional 689 (26%) individuals with autozygosity around 6.25% (first cousins). The remainder had a range of more distant relatedness around 3-4%. From estimates across all individuals we find that essentially every site in the coding region of the genome can be effectively homozygozed in healthy individuals. We observed 657 rare (<1%) LoF homozygotes in 639 genes, 96.4% within long autozygous sections. Population analysis revealed a total of 17,520 LoFs, on average 148 per individual, including 40 homozygous; 0.5 of them rare. Rare knockout genes include single gene drug targets in current preclinical to phase II development and 43 genes where knockout of the mouse homolog is lethal. Additionally, we observe knockouts in genes present in high penetrance Mendelian diseases having no apparent effect in these fit adult samples, thus providing potential information about their penetrance. By measuring the reduction in rare LoF mutation density in autozygous stretches compared to that in heterozygous regions, we have been able to estimate the fraction of genes for which loss of function is incompatible with healthy life. We describe large-scale validation experiments for all observed knockouts at the DNA level, and planned analyses at the RNA, protein and functional levels. We are now expanding beyond this pilot to the sequencing of 15,000 individuals with self-stated parental relatedness with consent to recontact for deeper phenotyping, allowing us to investigate in vivo the phenotypic consequences of knockouts in a large number of genes. Based on our pilot data we expect to find knockouts in a third of human genes, including multiple knockouts in many cases, providing phenotypic annotation for many currently uncharacterized human genes.

## 100

**Analysis of Loss-of-Function Variants in 8,612 Deeply-Phenotyped Individuals Identifies Novel Loci for Common Chronic Disease.** *A.H. Li[1], A.C. Morrison[1], G. Metcalf[2], L.A. Cupples[3,4], J.A. Brody[5], L.M. Polfus[1], B. Yu[1], N. Veeraraghavan[2], X. Liu[1], T. Lumley[5,6], D. Muzny[2], T.H. Mosley[7], R.A. Gibbs[2], E. Boerwinkle[1,2].* 1) Human Genetics Center, University of Texas Health Science Center, Houston, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3) National Heart, Lung, and Blood Institute (NHLBI) Framingham Heart Study, Framingham, MA; 4) Department of Biostatistics, Boston University School of Public Health, Boston, MA; 5) Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA; 6) Department of Statistics, University of Auckland, Auckland, New Zealand; 7) The Memory Impairment and Neurodegenerative Dementia Research Center, University of Mississippi, Jackson, MS.

A typical human exome analysis reveals more than 100 loss-of-function (LOF) variants, approximately 20 of which are homozygous and predicted to abolish gene function. The effects of these variants have been explored in clinical samples ascertained for rare phenotypes but have not been examined in population-based samples measured for a broad spectrum of common risk factor phenotypes and clinical outcomes. We sequenced the exomes of 8,612 ethnically-diverse individuals (2,849 African American, 5,763 European American) and identified 50,259 predicted LOF variants (splice, stopgain, frameshift indel). Gene-based burden analyses were performed on over 40 chronic disease risk factor phenotypes including serum electrolytes, liver enzymes, serum lipids, diabetes biomarker, markers of lung function and anthropomorphic measurements. Our analysis framework replicated known phenotypic associations of two well-studied genes (PCSK9, APOC3) and identified more than 11 novel associations. For example, individuals with LOF variants in SCNN1D, which encodes the delta subunit of the non-voltage-gated sodium channel 1 protein, presented elevated creatinine across multiple clinical visits (T5 burden, 2.5 x 10-8). We also demonstrate evidence for recessive effects, as demonstrated by the presentation of abnormal fasting glucose in samples with homozygous LOF genotypes in TYW1B, a gene within the Williams-Beuren syndrome deletion (MIM: 194050; 7q11.23), while heterozygotes appear normal (Wilcoxon, 2.6×10-4). These data demonstrate the utility of applying detailed functional annotation of whole exome sequence to a large sample of deeply-phenotyped individuals for novel gene discovery.

## 101

**Loss-of-Function Variants Influence the Human Metabolome.** *B. Yu[1], AH. Li[1], G. Metcalf[2], DM. Muzny[2], AC. Morrison[1], TH. Mosley[3], RA. Gibbs[2], E. Boerwinkle[1,2].* 1) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3) The Memory Impairment and Neurodegenerative Dementia Research Center, University of Mississippi, Jackson, MS.

Loss-of-function (LoF) variants are more frequent in individuals of African descent and are powerful instruments for identifying disease susceptibility genes. The metabolome is a collection of small molecules resulting from a multitude of biologic processes and can act as biomarkers of disease. We sequenced and annotated the exomes and measured 308 serum metabolites in a sample of 1,361 African-Americans (AAs) from the Atherosclerosis Risk in Communities (ARIC) Study. On average, each individual had 112 LoF variants and were homozygous for 7. Single SNP tests (for MAF > 5%) and gene burden tests (for cMAF ≤ 5%) were performed for each metabolite. We identified 12 novel genes ($p < 5.0E-7$) harboring more than 50 LoF variants affecting the metabolome. Depending on the metabolite, these loci were associated with 19-51% of the difference in metabolite levels, with an average effect of 38%. For example, six LoF mutations in SLCO1B1 were consistently associated with high levels of hexadecanedioate (cMAF = 2.7%, p = 9.3E-10), a C16 dicarboxylic acid. SLCO1B1 is an organic ion transporter and follow-up studies showed pleiotropic effects on other dicarboxylic acids (e.g. tetradecanedioate, p = 9.0E-5). These associations were replicated in an independent sample of 616 ARIC AAs ($p = 4.6E-5$ and $p = 1.6E-5$, respectively). Reflecting an important role of fatty acid oxidation in myocardial energy metabolism, these two metabolites were significant predictors of incident heart failure beyond the traditional risk factors, whereby higher levels of hexadecanedioate and tetradecanedioate were associated with increased risk ($p = 3.0E-7$ and $p = 4.3E-3$, respectively). In a second example, a LoF mutation in CD36 (Y325X, MAF = 8.4%), which encodes a membrane-bound fatty acid transporter, was associated with reduced levels of octanoylcarnitine and decanoylcarnitine ($p = 3.9E-8$ and $p = 2.7E-7$, respectively), and the associations were replicated in an independent sample of ARIC AAs ($p = 0.02$ and $p = 0.01$, respectively). These two fatty acylcarnitines compounds are biomarkers for medium-chain acyl-CoA dehydrogenase (MCAD) deficiency (MIM: 201450), and CD36 is an important component of platelet and monocyte biology. Our findings suggest a role of CD36 in regulating acylcarnitine levels. Taken together, these results provide new insights into gene function and the understanding of disease etiology by integrating -omic technologies in a deeply phenotyped population.

## 102

**Capture of 390,000 SNPs in dozens of ancient central Europeans reveals a population turnover in Europe thousands of years after the advent of farming.** *I. Lazaridis[1,2,*], W. Haak[3,*], N. Patterson[2], N. Rohland[1,2], S. Mallick[1,2], B. Llamas[3], S. Nordenfelt[1,2], E. Harney[1,2,4], A. Cooper[3], K.W. Alt[5,6,7], D. Reich[1,2,4].* 1) Department of Genetics, Harvard Medical School, Boston, MA, USA; 2) Broad Institute of Harvard and MIT, Cambridge, MA, USA; 3) Australian Centre for Ancient DNA, University of Adelaide, Australia; 4) Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA; 5) State Office for Heritage Management and Archaeology Saxony-Anhalt and Heritage Museum, Halle, Germany; 6) Center of Natural and Cultural History of Teeth, Danube Private University, Krems-Stein, Austria; 7) Hightech Research Center, University of Basel and Integrative Prehistory and Archaeological Science, Basel University, Switzerland.

To understand the population transformations that took place in Europe since the early Neolithic, we used a DNA capture technique to obtain reads covering ~390 thousand single nucleotide polymorphisms (SNPs) from a number of different archaeological cultures of central Europe (Germany and Hungary). The samples spanned the time period from 7,500 BP to 3,500 BP (Early Neolithic to Early Bronze Age periods) and most of them were previously studied using mtDNA (Brandt, Haak et al., Science, 2013). The captured SNPs include about 360,000 SNPs from the Affymetrix Human Origins Array that were discovered in African individuals, as well as about 30,000 SNPs chosen for other reasons (that are thought to have been affected by natural selection, or to have phenotypic effects, or are useful in determining Y-chromosome haplogroups). By analyzing this data together with a dataset of 2,345 present-day humans and other published ancient genomes, we show that late Neolithic inhabitants of central Europe belonging to the Corded Ware culture were not a continuation of the earlier occupants of the region. Our results highlight the importance of migration and major population turnover in Europe long after the arrival of farming. * Contributed equally to this work.

## 103

**Insights into British and European population history from ancient DNA sequencing of Iron Age and Anglo-Saxon samples from Hinxton, England.** *S. Schiffels[1], W. Haak[2], B. Llamas[2], E. Popescu[3], L. Loe[4], R. Clarke[3], A. Lyons[3], P. Paajanen[1], D. Sayer[5], R. Mortimer[3], C. Tyler-Smith[1], A. Cooper[2], R. Durbin[1].* 1) Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) Australian Centre for Ancient DNA, University of Adelaide, Australia; 3) Oxford Archaeology East, Cambridge, United Kingdom; 4) Oxford Archaeology South, Oxford, United Kingdom; 5) University of Central Lancashire, Preston, United Kingdom.

British population history is shaped by a complex series of repeated immigration periods and associated changes in population structure. It is an open question however, to what extent each of these changes is reflected in the genetic ancestry of the current British population. Here we use ancient DNA sequencing to help address that question. We present whole genome sequences generated from five individuals that were found in archaeological excavations at the Wellcome Trust Genome Campus near Cambridge (UK), two of which are dated to around 2,000 years before present (Iron Age), and three to around 1,300 years before present (Anglo-Saxon period). Good preservation status allowed us to generate one high coverage sequence (12x) from an Iron Age individual, and four low coverage sequences (1x-4x) from the other samples.

By providing the first ancient whole genome sequences from Britain, we get a unique picture of the ancestral populations in Britain before and after the Anglo-Saxon immigrations. We use modern genetic reference panels such as the 1000 Genomes Project to examine the relationship of these ancient samples with present day population genetic data. Results from principal component analysis suggest that all samples fall consistently within the broader Northern European context, which is also consistent with mtDNA haplogroups. In addition, we obtain a finer structural genetic classification from rare genetic variants and haplotype based methods such as FineStructure. Reflecting more recent genetic ancestry, results from these methods suggest significant differences between the Iron Age and the Anglo-Saxon period samples when compared to other European samples. We find in particular that while the Anglo-Saxon samples resemble more closely the modern British population than the earlier samples, the Iron Age samples share more low frequency variation than the later ones with present day samples from southern Europe, in particular Spain (1000GP IBS). In addition the Anglo-Saxon period samples appear to share a stronger older component with Finnish (1000GP FIN) individuals. Our findings help characterize the ancestral European populations involved in major European migration movements into Britain in the last 2,000 years and thus provide more insights into the genetic history of people in northern Europe.

## 104

**Fine-Scale Population Structure in Europe.** *S. Leslie[1], G. Hellenthal[2], S. Myers[3], P. Donnelly[3, 4], International Multiple Sclerosis Genetics Consortium.* 1) Statistical Genetics, Murdoch Childrens Research Institute, Parkville, Victoria, Australia; 2) University College London Genetics Institute, Darwin Building, Gower Street, London, WC1E 6BT, UK; 3) University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, UK; 4) The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK.

There is considerable interest in detecting and interpreting fine-scale population structure in Europe: as a signature of major events in the history of the populations of Europe, and because of the effect undetected population structure may have on disease association studies. Population structure appears to have been a minor concern for most of the recent generation of genome-wide association studies, but is likely to be important for the next generation of studies seeking associations to rare variants. Thus far, genetic studies across Europe have been limited to a small number of markers, or to methods that do not specifically account for the correlation structure in the genome due to linkage disequilibrium. Consequently, these studies were unable to group samples into clusters of similar ancestry on a fine (within country) scale with any confidence. We describe an analysis of fine-scale population structure using genome-wide SNP data on 6,209 individuals, sampled mostly from Western Europe. Using a recently published clustering algorithm (fineSTRUCTURE), adapted for specific aspects of our analysis, the samples were clustered purely as a function of genetic similarity, without reference to their known sampling locations. When plotted on a map of Europe one observes a striking association between the inferred clusters and geography. Interestingly, for the most part modern country boundaries are significant i.e. we see clear evidence of clusters that exclusively contain samples from a single country. At a high level we see: the Finns are the most differentiated from the rest of Europe (as might be expected); a clear divide between Sweden/Norway and the rest of Europe (including Denmark); and an obvious distinction between southern and northern Europe. We also observe considerable structure within countries on a hitherto unseen fine-scale - for example genetically distinct groups are detected along the coast of Norway. Using novel techniques we perform further analyses to examine the genetic relationships between the inferred clusters. We interpret our results with respect to geographic and linguistic divisions, as well as the historical and archaeological record. We believe this is the largest detailed analysis of very fine-scale human genetic structure and its origin within Europe. Crucial to these findings has been an approach to analysis that accounts for linkage disequilibrium.

## 105

**The population structure and demographic history of Sardinia in relationship to neighboring populations.** *J. Novembre[1], C. Chiang[2], J. Marcus[1], C. Sidore[3, 4, 5], M. Zoledziewska[3], M. Steri[3], H. Al-asadi[1], G. Abecasis[4], D. Schlessinger[6], F. Cucca[3, 5].* 1) Dept of Human Genetics, University of Chicago, Chicago, IL; 2) Deptartment of Ecology and Evolution, University of California - Los Angeles, Los Angeles, CA; 3) Instituto di Ricerca Genetica e Biomedica, CNR, Monserrato, Cagliari, Italy; 4) Center for Statistical Genetics University of Michigan Ann Arbor, MI; 5) Universita degli Studi di Sassari Sassari, Italy; 6) Laboratory of Genetics National Institute on Aging National Institutes of Health Baltimore, MD.

Numerous studies have made clear that Sardinian populations are relatively isolated genetically from other populations of the Mediterranean, and more recently, intriguing connections between Sardinian ancestry and early Neolithic ancient DNA samples have been made. In this study, we analyze a whole-genome low-coverage sequencing dataset from 2120 Sardinians to more fully characterize patterns of genetic diversity in Sardinia. The study contains one subsample that contains individuals from across Sardinia and a second subsample that samples 4 villages from the more isolated Ogliastra region. We also merge the data with published reference data from Europe and North Africa. Overall Fst values of Sardinia to other European populations are low (<0.015); however using a novel method for visualizing genetic differentiation on a geographic map, we formally show how Sardinia is more differentiated than would be expected given its geographic distance from the mainland, consistent with periods of isolation. Applications of the software Admixture show how Sardinia populations differ in the levels of recent admixture with mainland European populations and that there are only minor contributions from North African populations to Sardinian ancestry. Notably the Sardinians from Ogliastra contain a distinct genetic cluster with minimal evidence of recent admixture with mainland Europe. We found frequency-based f3 tests and the tree-based algorithm Treemix both also show minimal evidence of recent admixture. Given the relative isolation, one might expect to see a unique demographic history from neighboring populations. Using coalescent-based approaches, we find Sardinian populations have had more constant effective sizes over the past several thousand years than mainland European populations, which typically show evidence for rapid growth trajectories in the recent past. This unique demographic history has consequences for the abundance of putatively damaging and deleterious variants, and we use our data to address the prediction that the genetic architecture of disease traits is expected to involve fewer loci with a greater proportion of variants at common frequencies in Sardinia.

## 106

**Population structure in African-Americans.** *S. Gravel[1], M. Barakatt[1], B. Maples[2], M. Aldrich[4], E.E. Kenny[3], C.D. Bustamante[2], S. Baharian[1].* 1) Human Genetics, McGill University, Montreal, QC, Canada; 2) Genetics, Stanford University, Stanford, CA; 3) Department of Genetics and Genomic Sciences, The Charles Bronfman Institute for Personalized Medicine, New York, NY; 4) Department of Thoracic Surgery, Vanderbilt University, Nashville, TN.

We present a detailed population genetic study of 4 African-American cohorts comprising over 6000 genotyped individuals across US urban and rural communities: two nation-wide longitudinal cohorts, one biobank cohort, and the 1000 genomes ASW cohort. Ancestry analysis reveals a uniform breakdown of continental ancestry proportions across regions and urban/rural status, with 79% African, 19% European, and 1.5% Native American/Asian ancestries, with substantial between-individual variation. The Native Ancestry proportion is higher than previous estimates and is maintained after self-identified hispanics and individuals with substantial inferred Spanish ancestry are removed. This strongly supports direct admixture between Native Americans and African Americans on US territory, and linkage patterns suggest contact early after African-American arrival to the Americas. Local ancestry patterns and variation in ancestry proportions across individuals are broadly consistent with a single African-American population model with early Native American admixture and ongoing European gene flow in the South. The size and broad geographic sampling of our cohorts enables detailed analysis the geographic and cultural determinants of finer-scale population structure. Recent Identity-by-descent analysis reveals fine-scale structure consistent with the routes used during slavery and in the great African-American migrations of the twentieth century: east-to-west migrations in the south, and distinct south-to-north migrations into New England and the Midwest. These migrations follow transit routes available at the time, and are in stark contrast with European-American relatedness patterns.

## 107

**Genetic Testing of 400,000 Individuals Reveals the Geography of Ancestry in the United States.** *Y. Wang, J.M. Granka, J.K. Byrnes, M.J. Barber, K. Noto, R.E. Curtis, N.M. Natalie, C.A. Ball, K.G. Chahine.* Ancestry.com DNA LLC 153 Townsend Street, Ste. 800 San Francisco, CA 94107.

The population of the United States is formed by the interplay of immigration, migration and admixture. Recent research (R. Sebro et al., ASHG 2013) has shed light on the U.S. demography by studying the self-reported ethnicity from the 2010 U.S. Census. However, self-reported ethnicity may not accurately represent true genetic ancestry and may therefore introduce unknown biases. Since launching its DNA service in May 2012, AncestryDNA has genotyped over 400, 000 individuals from the United States. Leveraging this huge volume of DNA data, we conducted a large-scale survey of the ancestry of the United States. We predicted genetic ethnicity for each individual, relying on a rigorously curated reference panel of 3,000 single-origin individuals. Combining that with birth locations, we explored how various ethnicities are distributed across the United States Our results reveal a distinct spatial distribution for each ethnicity. For example, we found that individuals from Massachusetts have the highest proportion of Irish genetic ancestry and individuals from New York have the highest proportion of Southern European genetic ancestry, indicating their unique immigration and migration histories. We also performed pairwise IBD analysis on the entire sample set and identified over 300 million shared genomic segments among all 400,000 individuals. From this data, we calculated the average amount of sharing for pairs of individuals born within the same state or from two different states. In general, we found the genetic sharing decreases as the geographic distance between two states increases. However, the pattern also varies substantially among the 50 states. In summary, our analysis has provided significant insight on the biogeographic patterns of the ancestry in the United States.

## 108

**Statistical inference of archaic introgression and natural selection in Central African Pygmies.** *P. Hsieh[1], J.D. Wall[5], J. Lachance[4], S.A. Tishkoff[4], R.N. Gutenkunst[3], M.F. Hammer[2].* 1) Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ; 2) Arizona Research Laboratories Division of Biotechnology, University of Arizona, Tucson, AZ; 3) Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ; 4) Department of Biology and Genetics, University of Pennsylvania, Philadelphia, PA; 5) Institute for Human Genetics, University of California, San Francisco, CA.

Recent evidence from ancient DNA studies suggests that genetic material introgressed from archaic forms of Homo, such as Neanderthals and Denisovans, into the ancestors of contemporary non-African populations. These findings also imply that hybridization may have given rise to some of adaptive novelties in anatomically modern humans (AMH) as they expanded from Africa into various ecological niches in Eurasia. Within Africa, fossil evidence suggests that AMH and a variety of archaic forms coexisted for much of the last 200,000 years. Here we present preliminary results leveraging high quality whole-genome data (>60X coverage) for three contemporary sub-Saharan African populations (Biaka, Baka, and Yoruba) from Central and West Africa to test for archaic admixture. With the current lack of African ancient DNA, especially in Central Africa due to its rainforest environment, our statistical inference approach provides an alternative means to understand the complex evolutionary dynamics among groups of the genus Homo. To identify candidate introgressive loci, we scan the genomes of 16 individuals and calculate S*, a summary statistic that was specifically designed by one of us (JDW) to detect archaic admixture. The significance of each candidate is assessed through extensive whole-genome level simulations using demographic parameters estimated by ∂a∂i to obtain a parametric distribution of S* values under the null hypothesis of no archaic introgression. As a complementary approach, top candidates are also examined by an approximate-likelihood computation method. The admixture time for each individual introgressive variant is inferred by estimating the decay of the genetic length of the diverged haplotype as a function of its underlying recombination rate. A neutrality test that controls for demography is performed for each candidate to test the hypothesis that introgressive variants rose to high frequency due to positive directional selection. Several genomic regions were identified by both selection and introgression scans, and we will discuss the possible genetic and functional properties of these "double-hits". The present study represents one of the most comprehensive genomic surveys to date for evidence of archaic introgression to anatomically modern humans in Africa.

## 109

**Inferences about human history and natural selection from 280 complete genome sequences from 135 diverse populations.** *S. MALLICK[1,2,3], D. REICH[1,2,3], Simons Genome Diversity Project Consortium.* 1) Harvard Medical School, Boston, MA, USA; 2) Broad Institute of Harvard and MIT, Boston, MA, USA; 3) Howard Hughes Medical Institute, Chevy Chase, MD, USA.

The most powerful way to study population history and natural selection is to analyze whole genome sequences, which contain all the variation that exists in each individual. To date, genome-wide studies of history and selection have primarily analyzed data from single nucleotide polymorphism (SNP) arrays which are biased by the choice of which SNPs to include. Alternatively they have analyzed sequence data that have been generated as part of medical genetic studies from populations with large census sizes, and thus do not capture the full scope of human genetic variation. Here we report high quality genome sequences (~40x average) from 280 individuals from 135 worldwide populations, including 45 Africans, 26 Native Americans, 27 Central Asians or Siberians, 46 East Asians, 25 Oceanians, 46 South Asians, and 71 West Eurasians. All samples were sequenced using an identical protocol at the same facility (Illumina Ltd.). We modified standard pipelines to eliminate biases that might confound population genetic studies. We report novel inferences, as well as a high resolution map that shows where archaic ancestry (Neanderthal and Denisovan) is distributed throughout the world. We compare and contrast the genomic landscape of the Denisovan introgression into mainland Eurasians to that in island Southeast Asians. We are making this dataset fully available on Amazon Web Services as a resource to the community, coincident with the American Society of Human Genetics meeting.

## 110

**Galanin mutations in temporal lobe epilepsy.** *M. Guipponi[1], A. Chentouf[2], K.E.B. Webling[3], K. Freimann[3], A. Crespel[4], C. Nobile[5], T. Dorn[6], J. Hansen[6], J. Lemke[7,8], G. Lesca[9,10,11], F. Becker[12], U. Stephani[13], H. Muhle[13], I. Helbig[13], P. Ryvlin[14], E. Hirsch[15], G. Rudolf[15], C. Gehrig[1], F. Santoni[1], M. Pizzato[16], U. Langel[3], S.E. Antonarakis[1,17].* 1) Genetic Medecine, University of Geneva Medical School, Geneva, Switzerland; 2) Service of Neurology, CHU Oran, Oran, Algeria; 3) Department of Neurochemistry, Arrhenius Laboratories for Natural Science, Stockholm University, Stockholm, Sweden; 4) Departement of Neurology, Montpellier University Hospital, France; 5) Institute of Neurosciences, Section of Padua, Padova, Italy; 6) Swiss epilepsy center, Zürich, Switzerland; 7) Division of Human Genetics, University Children's Hospital Inselspital, Bern, Switzerland; 8) Institute of Human Genetics, University Hospital Leipzig, Germany; 9) Department of Medical Genetics, Hospices Civils de Lyon, France; 10) Claude Bernard Lyon I University, Lyon, France; CRNL, CNRS UMR 5292; 11) CRNL, CNRS UMR 5292; INSERM U1028, Lyon, France; 12) Department of Neurology and Epileptology, University of Tübingen, 72076 Tübingen, Germany; 13) Department of Neuropediatrics, University Medical Center Schleswig-Holstein (UKSH), Kiel, Germany; 14) Service of Neurology and Epileptology, CHU lyon-GH Est, Lyon, France; 15) Department of Neurology, INSERM UMR 7191, CHU Strasbourg-Hôpital Civil, Strasbourg, France; 16) Centre for integrative Biology (CIBIO), University of Trenton, Trenton, Italy; 17) Institute of Genetics and Genomics in Geneva (iGE3), Geneva, Switzerland.

Temporal lobe epilepsy (TLE) is a common and heterogeneous epilepsy syndrome with a complex etiology. Despite strong evidence for the participation of genetic factors, the genetic basis of TLE remains largely unknown. Here, we studied a family with a pair of monozygotic twins affected by temporal lobe epilepsy and two unaffected siblings born to healthy parents. Exome sequencing revealed that the twins carried a de novo missense mutation (A39E) in the galanin/GMAP prepropeptide (GAL) gene. In an independent cohort of 591 unrelated TLE patients, we observed 2 cases carrying heterozygous missense mutations in the GAL gene. One patient had the same mutation (A39E) as found in the monozygotic twins and the other carried an E65K substitution. Both mutations were predicted as damaging and were absent from dbSNPv138, 1000Genomes and EVS databases. Galanin is a 30-amino acid neuropeptide produced from the cleavage of the 123-amino acid preprogalanin protein, which acts as a potent anticonvulsivant and regulates epileptic seizures in animal models. The A39E mutation is located within the 5'-half of galanin which is highly conserved and critical for binding to receptors (GalR) and biological activity. Consistently, competitive binding assays showed that the A39E mutant had a significantly reduced binding affinity for GalR2 and GalR3 but not for GalR1 when compared to that of wild type galanin. The E65K mutant is located just downstream of the 3' dibasic tryptic cleavage site of the galanin neuropeptide. Using western blot analysis, we obtained preliminary data suggesting that the E65K mutation impaired galanin cleavage and would lead to the generation of a 3'uncleaved peptide with potentially altered binding and/or signaling activities. Over the last decade, galanin has emerged as a potent inhibitor of epileptic activity in animal models. Here, we report the identification of mutations interfering with galanin activity in individuals with temporal lobe epilepsy. Given the availability of galanin agonists, our findings could potentially have direct implication for the treatment of individuals with TLE.

## 111

**Homozygous mutations in *SLC6A17* are causative for autosomal recessive intellectual disability.** *H. van Bokhoven[1,2], Z. Iqbal[1], M. H. Willemsen[1], MA. Papon[3], H. Venselaar[4], W. M. Wissink-Lindhout[1], M. Benevento[1,2], A. T. Vulto-van Silfhout[1], L. E. L. M. Vissers[1], A. P.M de Brouwer[1], N. Nadif Kasri[1,2], T. F. Wienker[5], H. Hilger Ropers[5], L. Musante[5], K. Kahrizi[6], H. Najmabadi[6], F. Laumonnier[3], T. Kleefstra[1].* 1) Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; 2) Department of Cognitive Neurosciences, Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, Nijmegen, The Netherlands; 3) Institut National de la Santé et de la RechercheMédicale, Inserm U930, Tours, France; 4) Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, The Netherlands; 5) Max Planck Institute for Molecular Genetics, Berlin, Germany; 6) Social Welfare and Rehabilitation University, Tehran, Iran.

The combination of homozygosity mapping and next generation sequencing (NGS) has proven to be a powerful technique to identify autosomal recessive genetic defects. Here, we studied a Dutch family with three affected adult females presenting moderate to severe intellectual disability (ID) and neurologic-tremor. A combination of NGS and homozygosity mapping revealed a single homozygous mutation, c.484G>A, p.(Gly162Arg) in the gene *SLC6A17* (NM_001010898). Simultaneously, by following the similar approach, we identified another homozygous mutation c.1898C>G, p.(Pro633Arg) in an Iranian consanguineous family presenting with comparable phenotypic features including severe ID. SLC6A17 protein is exclusively expressed in the brain, and is a synaptic vesicular transporter of neutral amino acids. It plays an important role in the regulation of monoaminergic as well as glutamatergic synapses. The mutations are located on the 3rd and 12th transmembrane domains of the protein, respectively. Most of the prediction programs classified the identified mutations to be pathogenic. 3D modeling predicted that introduction of the Arginine at both locations will disrupt the conformation of the protein. To directly test the functional consequences, we investigated the neuronal subcellular localization of the wildtype and mutant proteins in mouse primary hippocampal neuronal cells. Our data revealed that the wildtype protein is present in soma, axons, dendrites and dendritic spines. The $Slc6a17^{Gly162Arg}$ mutant protein overexpression was associated with an abnormal neuronal morphology mainly characterized by the loss of dendritic spines, whereas, $Slc6a17^{Pro633Arg}$ mutant protein was found in soma and in proximal dendrites but did not reach spines. Because of these dramatic cellular phenotypes, it was not possible to extend the experiments to record electrophysiological measurements. In summary, our genetic findings are further strengthened with in-silico and in-vitro functional analyses, leading to assign a novel pathogenic role to *SLC6A17* implicated in autosomal recessive intellectual disability.

## 112

**De Novo KCNB1 Mutations in Epileptic Encephalopathy.** *A. Torkamani[1], K. Bersell[4], B.S. Jorge[4], R.L. Bjork[2], J.R. Friedman[3], C.S. Bloss[1], S.E. Topol[1], G. Zhang[1], J. Lee[1], J. Cohen[5], S. Gupta[6], S. Naidu[6], C.G. Vanoye[7], A.L. George[7], J.A. Kearney[8].* 1) The Scripps Translational Science Institute, San Diego, CA; 2) Pediatrics, Scripps Health, San Diego, CA; 3) Departments of Neurosciences and Pediatrics, University of California, San Diego, San Diego, CA; 4) Departments of Medicine and Pharmacology, Vanderbilt Brain Institute, Vanderbilt University, Nashville, TN; 5) Kennedy Krieger Institute, Baltimore, MD; 6) Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD; 7) Department of Pharmacology, Northwestern University Feinberg School of Medicine, Chicago, IL; 8) Division of Genetic MedicineVanderbilt University, Nashville, TN.

**Purpose**: An 8-year-old female presented with a sporadic severe partial seizure disorder with an unusual pattern including intermittent lapses into stupor, tantrums, and oppositional behavior followed by cataplexy with no memory of the events. Her overall condition includes a complex neurological history of global delay, hypotonia, epileptic encephalopathy, vision impairment, poor modulation of motor movement, blood pressure and pulse lability, long QT, and potential cerebral folate deficiency. The condition did not appear to fit any diagnostic category, was deteriorating and demonstrated breakthroughs to most seizure medications. Thus, a family-based genome sequencing study was pursued in order to identify the cause of her condition. **Methods**: Combined whole genome sequencing (WGS) and whole exome sequencing (WES) was performed on the affected 8-year-old female and her unaffected parents and sibling in order to identify the genetic cause of her complex neurological condition. A combination of inheritance-based, population-based, functional-impact-based and variant annotation-based filters were applied to small variants and copy number variants identified in the family in order to isolate the potential molecular cause of the proband's disorder. **Results**: We identified a de novo missense mutation in KCNB1 that encodes the KV2.1 voltage-gated potassium channel. Functional studies demonstrated a deleterious effect of the mutation on KV2.1 function leading to a loss of ion selectivity and gain of a depolarizing inward cation conductance. Subsequently, we identified two additional patients with epileptic encephalopathy and de novo KCNB1 missense mutations that result in a similar pattern of KV2.1 dysfunction. Our genetic and functional evidence demonstrate that KCNB1 mutation is a novel genetic cause of early onset epileptic encephalopathy.

## 113

**A *Drosophila* genetic resource facilitates the identification of variants in *ANKLE2* in a unique family with severe microcephaly.** *W.-L. Charng[1,2], M. Jaiswal[2,3], N. Link[2], S. Yamamoto[1,2,4], T. Gambin[2,5], E. Karaca[2], G. Mirzaa[6,7], W. Wiszniewski[2,8], B. Xiong[1], V. Bayat[1], T. Harel[2,8], D. Pehlivan[2], S. Penney[2,8], L.E. Vissers[9], J. de Ligt[9], S. Jhangiani[10], D. Muzny[2,10], R.D. Clark[11], C.J. Curry[12], E. Boerwinkle[10,13], W.B. Dobyns[6,7,14], R.A. Gibbs[2,10], R. Chen[1,2,10], M.F. Wangler[2,8], H.J. Bellen[1,2,3,4, 9,15], J.R. Lupski[2,8,10,16].* 1) Program in Developmental Biology, Baylor College of Medicine (BCM), Houston, TX, 77030; 2) Department of Molecular and Human Genetics, BCM, Houston, TX, 77030; 3) Howard Hughes Medical Institute, Houston, TX, 77030; 4) Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital (TCH), Houston, TX, 77030; 5) Institute of Computer Science, Warsaw University of Technology, 00-661 Warsaw, Poland; 6) Department of Pediatrics, University of Washington, Seattle, WA, 98195; 7) Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, 98101; 8) Texas Children's Hospital, Houston, TX, 77030; 9) 9Department of Human Genetics, Radboudumc, PO Box 9101, 6500 HB, Nijmegen, The Netherlands; 10) Human Genome Sequencing Center, BCM, Houston, TX, 77030; 11) Division of Medical Genetics, Department of Pediatrics, Loma Linda University Medical Center, Loma Linda, CA, 92354; 12) Department of Pediatrics, University of California San Francisco, San Francisco, CA, 94143, and Genetic Medicine Central California, Fresno, CA, 93701; 13) Human Genetics Center, University of Texas, Health Science Center at Houston, Houston, TX, 77030; 14) Department of Neurology, University of Washington, 98195; 15) Department of Neuroscience, BCM, Houston, TX, 77030; 16) Department of Pediatrics, Baylor College of Medicine, Houston, TX, 77030.

To create a resource for the study of human disease genes, we conducted a large scale forward genetic screen on fly *X*-chromosome and isolated lethal mutations in 165 genes involved in neuronal development, function, or maintenance. We then explored the corresponding 250 human homologs/orthologs in 1,929 human exomes from the Baylor Hopkins Center for Mendelian Genomics (see Wangler *et al.*). Many study subjects have neurological disease phenotypes. We identified disease associated mutations in *ANKLE2*, which is the only homolog of fly *dAnkle2*, and has not been associated with human disorders. In a single family, two children exhibit severe microcephaly, and hyper and hypo-pigmented macules of the skin. The head size of the proband was over 9 standard deviations below the mean and MRI revealed a polymicrogyria-like cortical malformation. *ANKLE2* is one of the 10 genes that segregate in a pattern consistent with Mendelian recessive expectations and one of four genes with high expression in the central nervous system (CNS). In our screen, *dAnkle2* mutant flies exhibit bristles loss in the peripheral nervous system due to an underdevelopment of sensory organs. In addition, the brain of *dAnkle2* mutant larva is much smaller than controls, indicating defective CNS development. Indeed, the number of dividing *dAnkle2* mutant neuroblasts in the developing brain is largely reduced and these neuroblasts undergo apoptosis. All the fly phenotypes can be rescued by human *ANKLE2*, indicating a functional conservation across species. Interestingly, *ANKLE2* is an interactor of *VRK1*, a gene implicated in pontocerebellar hypoplasia and reported by us in association with microcephaly and sensorimotor axonal distal symmetric polyneuropathy. We are currently culturing fibroblasts of proband and studies are underway to determine the defects observed in fly. Overall, our approach represents a unique combination of fly forward genetic screens and subsequent disease gene identification in human genomic data sets with additional functional studies in the fly and human cell lines. Recently, a mutation in *CLP1* has been reported by us to cause apoptosis of neuronal progenitors and lead to microcephaly. In summary, the finding of neuroblast susceptibility to apoptosis in *dAnkle2* mutant flies supports an emerging theme in neurodevelopmental disorders associated with both *VRK1* and *CLP1* variant alleles.

## 114

**Additive toxicity of SOX10 mutation underlies a complex neurological phenotype of PCWH.** *K. Inoue[1], Y. Ito[1,2], N. Inoue[1], Y.U. Inoue[3], S. Nakamura[1], Y. Matsuda[4], M. Inagaki[4], T. Ohkubo[1], J. Asami[3], Y.W. Terakawa[3], S. Kohsaka[5], Y. Goto[1], C. Akazawa[5,6], T. Inoue[3].* 1) Dept MR & BD Res, Natl Inst Neurosci, NCNP, Kodaira, Tokyo, Japan; 2) Dept Molecular Neuroscience, Med Res Inst, Tokyo Med & Dent Univ., Tokyo, Japan; 3) Dept. Biochemistry & Cellular Biology, Natl Inst Neurosci, NCNP Tokyo, Japan; 4) Dept. Developmental Disorders, Natl Inst Neurosci, NCNP Tokyo, Japan; 5) Dept. Neurochemistry, Natl Inst Neurosci, NCNP Tokyo, Japan; 6) Dept Biochemistry & Biophysics, Grad School of Health Care Sciences, Tokyo Med & Dent Univ., Tokyo, Japan.

Distinct classes of *SOX10* mutations result in peripheral demyelinating neuropathy, central dysmyelinating leukodystrophy, Waardenburg syndrome, and Hirschsprung disease, collectively known as PCWH. Meanwhile, *SOX10* haploinsufficiency caused by allelic loss-of-function mutations leads to a milder non-neurological disorder, Waardenburg-Hirschsprung disease. The cellular pathogenesis of more complex PCWH phenotypes in vivo has not been elucidated until now. Here we determined the pathogenesis of PCWH by constructing a transgenic mouse model. A PCWH-causing *SOX10* mutation, c.1400del12, was introduced into mouse *Sox10*-expressing cells by means of bacterial artificial chromosome (BAC) transgenesis. By crossing the multiple transgenic lines, we examined the effects produced by various copy numbers of the mutant *SOX10* transgene. In the nervous systems, transgenic mice revealed delay in integration of Schwann cells in the sciatic nerve and terminal differentiation of oligodendrocytes in the spinal cord. Transgenic mice also showed defects of melanocytes presenting as neurosensory deafness and abnormal skin pigmentation, and a loss of the enteric nervous system. Phenotypes in each lineage were more severe in mice carrying higher copy numbers, suggesting a gene dosage effect of toxic mutant *SOX10*. By uncoupling the effects of gain-on-function and haploinsufficiency in vivo, we have demonstrated that the effect of a PCWH-causing *SOX10* mutation is solely toxic in all *SOX10*-expressing cells in dosage-dependent manner. In both peripheral and central nervous systems, primary consequence of *SOX10* mutations is hypomyelination. The complex neurological phenotypes in PCWH patients likely result from a combination of haploinsufficiency and additive dominant toxicity.

## 115

**Paving the road to elaborate the genetics of intellectual disabilities.** *H. Najmabadi[1], H. Hu[2], Z. Fattahi[1], S. Abedini[1], M. Hosseini[1], F. Larti[1], R. Jazayeri[1], M. Oladnabi[1], M. Mohseni[1], T. Wienker[2], L. Musante[2], K. Kahrizi[1], H.H. Ropers[2].* 1) Genetics Research Center, University of Social Welfare & Rehabilitation Sciences, Tehran, Iran; 2) Max Planck Institute for Molecular Genetics, Berlin, Germany.

The complex human brain has been derived from very simple system and through evolution has adapted a role to an extremely powerful and capable part of our body to make us what we are. It is believed that more than fifty percent of our genes needed for its function. Failure of most of these genes could create catastrophic effects for the individual resulting in intellectual disability (ID). Many of these changes could be inherited while some could be de nova. Over eleven years ago we decided to structure a system to identify these genetic causes; "A comprehensive approach". Many families from different ethnicities (Persian, Turk, Arab, Kurdish, and...) with complete clinical profile having two and more affected individuals were recruited. Karyotying, FMR1 testing, and families with ID and microcephaly were also checked for the mutation by linkage analysis and conventional sequencing. In the first few years of the study we indentified number of novel genesand later on in October 2011 we reported 50 novel ID genes using next generation sequencing. Since then additional 240 families with two and more affected have been investigated in our group. This work has been continuing last couple of years and here we are reporting additional 52 novel genes either causing syndromic or non-syndromic ID and many previously reported ID genes. In over 50% of thesefamilies the pathologicalchanges have been identified. In 19 families we could not conclude a single candidate gene because they had at least two candidate genes.Functional analyses have been conducted in many of these genes, and theresults showthe connection between these genes and the involved pathways in the human brain, as well as the role of these genes in brain size and its functions.These findings have contributed to improve the diagnosis of ID and understanding the human brain function.

## 116

**KIRREL3, associated with intellectual disability and autism, functions as a presynaptic organizer and interacts with proteins with roles in neurodevelopment.** *A.K. Srivastava[1], Y.F. Liu[1], Y. Luo[1,3], A. Chaubey[1], H-G. Kim[2], S.M. Sowell[1].* 1) J.C. Self Research Institute of Human Genetics, Greenwood Genetic Center, Greenwood, SC; 2) Department of OB/GYN, Institute of Molecular Medicine and Genetics, Georgia Regents University, GA; 3) Present address: Department of Pediatrics, Emory University School of Medicine, Atlanta, GA.

A large body of evidence indicates dysfunction of the synapse (synapse formation and plasticity) as a major contributing factor in autism spectrum disorder (ASD) and intellectual disability (ID). Defects of the *KIRREL3* gene, located at 11q24.2, which encodes a synaptic cell-adhesion molecule of the immunoglobulin (Ig) superfamily, have recently been identified in ID, ASD and in the neurocognitive delay associated with Jacobsen syndrome. However, the molecular mechanisms of its physiological actions remain largely unknown. We have functionally characterized KIRREL3 and determined that in neuronal cells, KIRREL3 localizes on the cell membrane, in selected areas in the cytoplasm as well as in punctate structures along the entire length of neurite processes. We further confirmed its co-localization with markers for the Golgi apparatus and secretory vesicles. Using deletion constructs, we determined that the fourth Ig-like domain in the KIRREL3 extracellular domain (ECD) might be crucial for the secretory vesicle sublocalization of KIRREL3. Importantly, we found that the ECD of KIRREL3 can be cleaved after transient transfection in neuronal cells. Using a mixed culture system, we observed that both the full-length KIRREL3 and cleaved ECD can promote clustering of synaptic vesicles indicating their potential roles as presynaptic organizers. To further gain an understanding of the physiological role of KIRREL3 in neurodevelopment, we have shown that its intracellular domain interacts with the X-linked ID-associated synaptic scaffolding protein CASK and KIRREL3-ECD interacts with brain expressed proteins MAP1B and MYO16. In addition, we identified a genomic deletion encompassing MAP1B in one patient with ID, microcephaly and seizures. We also identified deletions encompassing MYO16 in two unrelated patients with ID, autism and microcephaly. MAP1B is involved in the development of the actin-based membrane skeleton. MYO16 has been shown to indirectly affect actin cytoskeleton through its interaction with WAVE1 complex. Together these findings suggest a potential contribution of KIRREL3 in the local assembly of the F-actin cytoskeleton at presynaptic sites that potentially initiates synapse formation. Our findings provide further insight into understanding the molecular mechanisms underlying the physiological action of KIRREL3 and its role in neurodevelopment.

## 117

**The importance of neurosteroid hormones in the pathogenesis of Protocadherin 19 female limited epilepsy and intellectual disability (PCDH19-FE).** *J. Gecz[1,2,3,4], C. Tan[1], E. Ranieri[2], D. Pham[1,3], C. Shard[4], K. Hynes[1], E. Douglas[2], L.S. Nguyen[1], M. Corbett[1], D. Leach[5], G. Buchanan[5], E. Haan[6], L.G. Sadleir[7], C. Depienne[8], R.S. Moller[9], R. Guerrini[10], C. Marini[10], S.F. Berkovic[11], I.E. Scheffer[11,12].* 1) Paediatrics, The University of Adelaide at Women's & Children's Hosp, Adelaide, South Australia, Australia; 2) SA Pathology, Adelaide Australia; 3) Robinson Institute, The University of Adelaide, Adelaide, SA, Australia; 4) School Molecular and Biomedical Sciences, The University of Adelaide, Adelaide, Australia; 5) Basil Hetzel Institute for Translational Health Research, The Queen Elizabeth Hospital, Adelaide, Australia; 6) South Australian Clinical Genetics Service, SA Pathology (at Women's and Children's Hospital), North Adelaide, Australia; 7) Department of Paediatrics and Child Health, School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand; 8) INSERM UMR 975, Hôpital Pitié-Salpêtrière, Paris, France; 9) Danish Epilepsy Centre, Dianalund and Institute for Regional Health Services, University of Southern Denmark, Odense, Denmark; 10) Pediatric Neurology Uni, Children's Hospital, Florence, Italy; 11) Epilepsy Research Centre, The University of Melbourne, Heidelberg, Australia; 12) Florey Institute of Neuroscience and Mental Health, Melbourne, Australia.

PCDH19-Female-Epilepsy (PCDH19-FE) is an unusual X-linked disorder that primarily affects females. PCDH19-FE encompasses a broad clinical spectrum from infantile epileptic encephalopathy resembling Dravet syndrome to epilepsy with or without intellectual disability and autism spectrum disorders. PCDH19-FE is highly, but not fully penetrant. We have used transcriptome profiling of primary skin fibroblasts of PCDH19-FE females (n=12 and n=3 age and passage matched normal controls) and unaffected transmitting males (n=3 and n=3 age and passage matched control males) to study PCDH19-FE pathology. We found 94 significantly de-regulated genes between PCDH19-FE and control females (One-way ANOVA, $p < 0.05$, fold change $> \pm 2$), of which 73 were annotated. Among these 94 genes there were 43 genes, which showed sex-biased expression in our control male versus control female comparison (223 genome-wide sex-biased genes). The enrichment of sexually biased genes among our significantly de-regulated genes was highly significant, $p = 2.51 \times 10\text{-}47$, Two-tail Fisher's exact test. Followup studies using additional patient skin fibroblast cell lines (including the cell line from one affected somatic mosaic male) validated the majority of deregulated genes, among them the aldo-keto reductase family 1, members C1-3 (AKR1C1-3) genes. The AKR1C genes play a crucial role in neurosteroid hormone metabolism. Human skin is endowed to metabolise neurosteroids. Of relevance to this is, that germline mutations of AKR1C genes cause disorders of sexual development (ie. sex reversal). We subsequently showed that AKR1C protein levels are affected and as a result of this the PCDH19-FE girls are allopregnanolone deficient (based on peripheral blood allopregnanolone tests). Additional support for the role of steroid hormones in the pathology of PCDH19-FE came from the age of onset (mean ~10 months) and offset (mean ~12.5 years) of epilepsy (n=140 patients), both of which coincide with dramatically varying sex hormone levels (onset - after 'minipuberty' at ~8-9 months and offset - with the advent of puberty at ~12 years). Our data together with the broadly discussed role of steroids in epilepsy led us to postulate that steroid hormones and specifically neurosteroids like allopregnanolone are involved in the pathology of PCDH19-FE. These findings open realistic opportunities for targeted therapeutic interventions.

## 118

**Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut rhythm.** *N. Gosset[1], P. Chetaille[2], J-M. Côté[2], C. Houde[2], C. Preuss[1], S. Burkharda[3], J. Castilloux[2], J. Piché[1], S. Leclerc[1], F. Wünnemann[1], M. Thilbault[1], C. Gagnon[1], A. Galli[4], E. Tuck[4], G-R.X. Hickson[5], N. El Amine[5], F. LeDeist[5], E. Lemyre[5], P. De Santa Barbara[6], S. Faure[6], A. Jonzon[7], M. Cameron[1], H. Dietz[8], E. Gallo-McFarlane[8], W. Benson[9], Y. Shen[10], M. Jomphe[11], S-J.M. Jones[10], J. Bakkers[3], G. Andelfinger[1].* 1) Cardiovascular Genetics, CHU Sainte Justine Research Center, Montreal, Quebec, Canada; 2) Centre Mère Enfants Soleil, CHU de Québec, Québec, QC, Canada; 3) Hubrecht Institute, Utrecht, The Netherlands; 4) The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; 5) Department of Pediatrics, University of Montréal, Montréal, Québec, Canada; 6) INSERM U1046, Montpellier Cedex, France; 7) Astrid Lindgren's Children's Hospital, Uppsala University, Uppsala, Sweden; 8) Johns Hopkins University School of Medicine, Howard Hughes Medical Institute, Baltimore, MD, USA; 9) Children's Hospital of Wisconsin, Milwaukee, WI, USA; 10) Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada; 11) Projet BALSAC, Université du Québec à Chicoutimi, QC, Canada.

Disturbances of pacemaker activity in the heart and the gut can have varied clinical manifestations. In the heart, dysregulation of the sinus node results in sick sinus syndrome (SSS), the most common cause of pacemaker implantation. In the gut, pacemaking is mediated through the network of interstitial cells of Cajal and the autonomous enteric nervous system. Chronic intestinal pseudo-obstruction (CIPO) is a rare and severe disorder of gastrointestinal motility, in which intestinal obstruction occurs in the absence of a mechanical obstacle. Here, we describe a new syndrome characterized by Chronic atrial and intestinal dysrhythmia, termed CAID syndrome. We identified a cohort of 17 subjects in whom SSS and CIPO co-occurred during the first four decades of life including 16 French Canadian and one Swedish patient. We show that a single shared homozygous founder mutation in *SGOL1*, a component of the cohesin complex, causes CAID syndrome. Whole-exome sequencing and genetic fine mapping of the disease associated haplotype revealed a northern European origin for the rare haplotype on which the ultra-rare mutation (1/8598 in NHLBI exome dataset) recently arose. Genealogical analysis traced back the most common ancestors to a founder couple married in 1620 in France supporting the idea of a transatlantic founder effect. Cultured dermal fibroblasts from affected individuals showed accelerated cell cycling, increased senescence and enhanced activation of TGF-β signaling. Karyotypes show the typical railroad appearance of a centromeric cohesion defect. Patient tissues display pathological changes of both the enteric nervous system and smooth muscle. Morpholino-induced knockdown of *sgol1* in zebrafish recapitulated abnormalities seen in humans with CAID syndrome. Taken together, our findings identify CAID syndrome as a novel generalized dysrhythmia, suggesting a new role for *SGOL1* and the cohesin complex in mediating the integrity of human cardiac and gut rhythm.

## 119

**Functional Characterization of Long-QT Syndrome (LQT) and Sudden Infant Death (SIDS) Associated OLFML2B Mutations.** *T.A. Plötz[1], C.J. Gloeckner[1,2], A. Kiper[3], M. Vennemann[4], M. Kartmann[5], M. Schell[5], H. Prucha[6], C. Congiu[7], Z. Schäfer[1], S. Hauck[1], I. Sinicina[5], E. Kremmer[1], B.M. Beckmann[9], F. Domingues[7], T. Meitinger[1,6,10], A. Peters[1,5,10], M. Cohen[8], S. Kääb[9,10], J.J. Schott[11], E.R. Behr[12], T. Bajanowski[13], S. Just[14], H.W. Mewes[1,6], M. Ueffing[1,2], N. Decher[3], M. Näbauer[9], A. Pfeufer[1,6].* 1) Helmholtz Zentrum München, Neuherberg, Germany; 2) Universität Tübingen, Tübingen, Germany; 3) Universität Marburg, Marburg, Germany; 4) Universität Münster, Münster, Germany; 5) LMU München, München, Germany; 6) TU München, München, Germany; 7) EURAC, Bolzano, Italy; 8) University of Sheffield, Sheffield, UK; 9) Department of Medicine I, University Hospital Munich, Ludwig-Maximilians-University Munich, Munich, Germany; 10) Deutsches Zentrum für Herz-Kreislauf-Forschung e.V. (DZHK), partner site Munich Heart Alliance, Munich, Germany; 11) Universite de Nantes, Nantes, France; 12) St Georges University of London, London, UK; 13) Universität Essen, Essen, Germany; 14) Universität Ulm, Ulm, Germany.

We have mapped the strongest human QTL modifying cardiac repolarization (QT interval) to OLFML2B and NOS1AP in 1q23.3 by GWAS. OLFML2B encodes a secreted extracellular matrix (ECM) protein. By mutation screening we have identified overrepresentation of rare (MAF≤1%) nonsynonymous heterozygous mutations in 125 patients with long-QT Syndrome (LQT; OR=3.62 (1,46-8,93) p=2,9×10-3) and in 93 with sudden infant death syndrome (SIDS; OR=3,01 (1,05-8,65) p=3,2×10-2) but not in 94 adults with sudden cardiac death (SCD; OR=0,57 (0,07-4,41) p=5,9×10-1) compared to 702 population controls. Of 35 missense variants identified, we have selected 24 predicted to negatively affect protein structure (PolyPhen2, SIFT, Mutation taster) for functional analysis. Combining our 702 controls with in-silico data from 6503 WES sequenced individuals from the ESP, 10 variants of the 24 occurred among 7205 persons while the other variants were absent in the expanded control sample. All variants were equally expressed intracellularly. Their secretion into the extracellular space was impaired depending on the mutation ranging from mild reduction to non-secretion. Co-expression of wt and mut demonstrated dominant negative secretion impairment. Experiments were performed in triplicate at three temperatures and parametrized (LI-COR Image Studio) using wt-OLFML2B at 37°C as a reference. Protein secretion was temperature dependent (30°C>37°C>41°C; p<0,001). In addition secretion was significantly correlated with disease severity (wt>LQT>SIDS; p<0,05) and with allele frequency in the controls (p<0.01). Five out of the 24 variants were investigated by cellular electrophysiology in Xenopus oocytes. They showed significant reduction of the voltage gated KCNH2/Kv11.1 channel (IKr) but no other main cardiac ion channels. The degree of impairmant ranged from -10% to -50% and was also correlated with mutation secretion status. Taken together the functional proteomic investigation suggests a significant influence of OLFML2B and the ECM on myocardial repolarization. This assumption is supported by nonsecretion being associated with both disease severity and population allele frequency acting in an autosomal dominant manner. Our data support the hypothesis that rare nonsynonymous OLFML2B variants impair repolarization, most likely by failing to assume the correct topological position in the ECM, and confer genetic predisposition to long QT-Syndrome (LQT) and sudden infant death (SIDS).

## 120

***EIF2AK4* (*GCN2*) mutations cause pulmonary veno-occlusive disease, a severe form of pulmonary hypertension.** *F. SOUBRIER[1, 2, 3], M. EYRIES[1,2, 3], D. MONTANI[4,5, 6], B. GIRERD[4,5, 6], C. PERRET[1, 3], A. LEROY[2], C. LONJOU[7], N. CHELGHOUM[7], F. COULET[2, 3], D. BONNET[8,9], P. DORFMULLER[6,10], E. FADEL[6,11], O. SITBON[4, 5, 6], G. SIMONNEAU[4,5,6], D-A. TREGOUET[1,3], M. HUMBERT[4-6].* 1) UMR_S1166 UPMC and INSERM Paris France; 2) Genetics Department, Hôpital Pitié-Salpêtrière, APHP, Paris; 3) Institute for Cardiometabolism and nutrition (ICAN), Paris, France; 4) Université Pairs Sud, le Kremilin-Bicêtre, PAris France; 5) DHU Thorax innovation (TORINO), service de Pneumologie, Hôpital Bicêtre APHP, Le Kremlin-Bistre, France; 6) UMR_S 999 Labex LERMIT, Centre Chirurgical Marie-Lannelongue, Le Plessis-Robinson, France; 7) Post-Genomic Platform (P3S) UPMC; INSERM, Paris, France; 8) Pediatric Cardiology Dept, Hôpital Necker-Enfants malades, APHP, Paris, France; 9) UMR_S 765 INSERM; Université Paris-Descartes, Paris, France; 10) Dept of Pathology, Centre Chirurgical Marie-Lannelongue, Le Plessis-Robinson, France; 11) Thoracic Surgery Dept, Centre chirurgical Marie-Lannelongue, Le Plessis Robinson, France.

Pulmonary veno-occlusive disease (PVOD) is a rare and severe cause of pulmonary hypertension characterized histologically by widespread thickening and fibrous intimal proliferation of septal veins and preseptal venules. These lesions are frequently associated with pulmonary capillary dilatation and proliferation. PVOD presents either sporadically or as familial cases. In the French referral centre for severe pulmonary hypertension, we have identified 13 PVOD families: 5 with a confirmed diagnosis based on histological studies and 8 with a highly likely diagnosis, based on clinical, functional, and radiological criteria. All PVOD families were characterized by the presence of at least two affected siblings and unaffected parents, suggesting an autosomal recessive transmission. We used a whole-exome sequencing approach and detected recessive mutations (homozygous or compound heterozygous) in the EIF2AK4 (GCN2) gene that co-segregated with PVOD in all 5 families initially studied. We subsequently identified mutations in the 8 additional PVOD families. We also found bi-allelic EIF2AK4 mutations in 5 of 20 histologically confirmed sporadic PVOD cases. All identified mutations disrupted the function of the gene. In conclusion, we identified the first gene responsible for PVOD. Biallelic mutations in EIF2AK4 gene were found in 100% of familial cases and in 25% of sporadic cases of PVOD, making this new gene a major player linked to PVOD development. This discovery significantly contributes towards understanding the complex genetic architecture of pulmonary hypertension. Results obtained in the mouse model of EIF2AK4 inactivation will be presented.

## 121

**Delineation and Therapeutic Implications of a Modifier Locus of Aortic Aneurysm in Marfan Syndrome.** *A. Doyle[1,2,3], J. Doyle[1], K. Kent[1], L. Myers[1], N. Wilson[1], N. Huso[1], D. Bedja[4,5], M. Lindsay[6], J. Pardo-Habashi[1,7], B. Loeys[8], J. De Backer[9], A. De Paepe[9], H. Dietz[1,2,7].* 1) Institute of Genetic Medicine, Johns Hopkins Medical Institute, Baltimore, MD, USA; 2) Howard Hughes Medical Institute, Baltimore, MD, USA; 3) William Harvey Research Institute, Barts and The London School of Medicine, Queen Mary University of London, London, UK; 4) Department of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 5) Australian School of Advanced Medicine, Macquarie University, Sydney, Australia; 6) Massachusetts General Hospital Thoracic Aortic Center, Departments of Medicine and Pediatrics, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; 7) Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 8) Centre of Medical Genetics, University of Antwerp and Antwerp University Hospital, Antwerp, Belgium; 9) Centre of Medical Genetics, Ghent University Hospital, Ghent, Belgium.

Marfan syndrome (MFS) is a connective tissue disorder caused by mutations in the FBN1 gene. The major cause of mortality is aortic aneurysm, dissection and rupture. While the disorder shows high penetrance, there is also variable expression, which has been attributed to extreme allelic heterogeneity as well as variation dictated by the level of expression of the wild-type FBN1 allele. There is a growing body of evidence that promiscuous activation of TGFβ is responsible for multiple manifestations of the disease and that specific inhibition of the angiotensin II type 1 receptor (AT1R) and/or mitogen activated protein kinase (MAPK; JNK or ERK1/2) activation can ameliorate aortic aneurysm. We previously reported identification of a major protective modifier locus for MFS, encompassing a 5.5Mb linkage interval on chromosome 6 (LOD= 4.0), using 5 exceptional families with defined and typical FBN1 mutations showing discrete intrafamilial variation in the penetrance of vascular disease. While the protective haplotype varied between families, all patients with mild disease (20/20) shared a 3.9Mb familial haplotype that was only observed in 2/18 severely affected family members ($p<0.0001$). Of the 32 genes in the linkage interval, 2 emerged as strong candidates based on known function; MAS1 encoding the receptor for Ang1-7, a natural antagonist of AT1R signaling, and MAP3K4, a MAPK kinase kinase and effector of noncanonical TGFβ signaling. To date, direct sequencing of all exons and flanking intron boundaries has not identified causative variation. In the absence of additional families to narrow the linkage interval, we turned to functional analyses in a validated mouse model of MFS that shows fully penetrant postnatal aneurysm progression ($Fbn1^{C1039G/+}$). Targeted disruption of a single Map3k4 allele was sufficient to fully normalize the aortic root growth rate to wild-type levels in $Fbn1^{C1039G/+}$ mice ($p<0.01$). Similarly, systemic administration of Ang1-7 (the endogenous MAS1 receptor ligand) to $Fbn1^{C1039G/+}$ mice not only prevented abnormal aortic root and ascending aortic growth when compared to untreated mutant littermates ($p<0.001$ for both comparisons) but also induced a significant regression in the absolute aortic root size ($p<0.001$) over a 3-month time period. These results provide further evidence of a protective locus on the distal arm of chromosome 6 and indicate that both MAP3K4 and MAS1 represent novel therapeutic targets in patients with MFS.

## 122

**Mutations in FOXE3/Foxe3 Cause Familial Thoracic Aortic Aneurysms and Dissections.** *S.Q Kuang[1], O. Medina-Martinez2[2], D.C Guo[1], L. Gong[1], E.S. Regalado[1], C. Boileau[4], G. Jondeau[5], S.K. Prakash[1], A.M. Peters[1], H. Pannu[1], M.J. Bamshad[3], J. Shendure[3], D.A. Nickerson[3], C.L. Reynolds[6], M. Jamrich[2], D.M. Milewicz[1].* 1) Internal Medicine, The University of Texas Health Science Center at Houston, Houston, TX; 2) Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030; 3) 3Department of Genome Sciences, University of Washington, Seattle, Washington; 4) AP-HP, Hôpital Bichat, Centre National de Référence pour le syndrome de Marfan et apparentés, Paris, France; Université Paris 7, Paris, France; AP-HP, Hôpital Bichat, Laboratoire de Génétique moléculaire, Boulogne, France; INSERM, U1148, Paris, France; 5) AP-HP, Hôpital Bichat, Centre National de Référence pour le syndrome de Marfan et apparentés, Paris, France; Université Paris 7, Paris, France; AP-HP, Hôpital Bichat, Service de Cardiologie, Paris, France; INSERM, U1148, Paris, France; 6) Mouse Phenotyping Core, Baylor College of Medicine, Houston, Texas, USA.

Mutations in FOXE3 cause lens defects due to lack of proliferation and differentiation of epithelial cells. Exome sequencing of a large family with autosomal dominant thoracic aortic aneurysms and dissections (TAAD) identified a rare variant in FOXE3, c.457G>C (p. D153H) that altered a conserved amino acid, was predicted to disrupt protein function, and segregated with TAAD in the family with decreased penetrance in women. Sequencing of 354 unrelated probands with TAAD identified three additional FOXE3 variants, p.D156N, p.G137D, and p.R164S, which were novel and predicted to be damaging. These families did not have lens defects, and the mutations were in a different region of the DNA binding domain compared with mutations that cause lens defects. Morpholino (MO) knockdown of foxe3 in zebrafish disrupted aortic arch development in 70% of embryos. Co-injection of wild-type but not mutant foxe3 RNA resulted in partial rescue of the phenotype. Foxe3 is not expressed in adult mouse aortas but in situ hybridization detected Foxe3 expression in the mouse embryo pharyngeal arches from E9.5 to E10.5, suggesting that Foxe3 is involved in establishing neural crest-derived aortic smooth muscle cells (SMCs). Aortas from 4 week old Foxe3-/- mice had reduced SMC density and decreased differentiation of SMCs, but the mice did not form aneurysms. When exposed to increased pressures by constricting the transverse aorta (TAC), Foxe3-/- mice developed larger aneurysms compared with wildtype (WT) mice and aortic rupture occurred in 50% of Foxe3-/- mice. TUNEL staining of the ascending aorta showed more SMC apoptosis in the Foxe3-/- mice than WT mice ($p<0.05$). To drive cellular survival pathways, p53 activity was disrupted by administration of an inhibitor (pifithrin-$\alpha$) or crossing the Foxe3-/- mice into p53-/- mice, and both rescued aortic SMC apoptosis and aortic rupture in TAC-challenged Foxe3-/- mice. Aortic SMC density and differentiation were also rescued to WT levels in 4 week old Foxe3-/-p53-/- mice. These data indicate that loss of Foxe3 leads to decreased numbers and de-differentiated SMCs in the ascending aorta, and SMC apoptosis and aortic rupture with increased biomechanical stress in the adult ascending aorta. Interestingly, blocking p53 in the Foxe3-/- mice rescued these phenotypes by driving survival and differentiation of the SMCs.

## 123

**Titin truncations: dissection of genotype and cardiac phenotype.** *A. Roberts[1, 2, 3, 12], J. Ware[2, 3, 8, 9], D. Herman[8, 9], S. Schafer[4], J. Baksi[2, 3], R. Buchan[2, 3], R. Walsh[2, 3], S. John[2, 3], S. Wilkinson[2, 3], L. Felkin[2, 3], A. Bick[8, 9], F. Mazzarotto[2, 3], M. Radke[6], M. Gotthardt[6, 7], P. Barton[2, 3], N. Hubner[4, 5, 7], J. Seidman[8, 9], C. Seidman[8, 9, 10], S. Cook[2, 3, 11, 12].* 1) Clinical Sciences Centre, Imperial College London, UK; 2) NIHR Cardiovascular Biomedical Research Unit at Royal Brompton & Harefield NHS Foundation Trust and Imperial College London, UK; 3) National Heart and Lung Institute, Imperial College London, UK; 4) Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine, Berlin, Germany; 5) Charité-Universitätsmedizin, Berlin, Germany; 6) Neuromuscular and Cardiovascular Cell Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany; 7) DZHK (German Centre for Cardiovascular Research), partner site Berlin, Germany; 8) Department of Genetics, Harvard Medical School, Boston, USA; 9) Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; 10) Cardiovascular Division, Brigham and Women's Hospital and Howard Hughes Medical Institute, USA; 11) National Heart Centre Singapore, Singapore; 12) Duke-National University of Singapore, Singapore.

TTN truncating variants (TTNtv) cause severe dilated cardiomyopathy (DCM), but sometimes occur in healthy individuals, posing significant challenges for the interpretation of these variants in an era of accessible genome sequencing. The mechanism by which TTNtv impact clinical outcomes is poorly understood. Here, we integrated the power of quantitative cardiac MRI and capacity of next generation sequencing to assess the relationship between TTN genotype and cardiac phenotype. We sequenced TTN in 4,440 subjects including 308 healthy volunteers, 3,603 Framingham Heart Study (FHS) and Jackson Heart Study (JHS) participants, 374 prospective, unselected DCM cases and 155 end-stage retrospective DCM cases including 84 for whom left ventricular (LV) tissue was available for RNA and protein studies.

TTNtv were identified in 1.4% of controls (healthy volunteers, FHS and JHS participants), in 13% of unselected and 22% of end-stage DCM cases (OR 16.6, $P=4.8\times10^{-45}$, DCM vs controls). To improve TTN transcript annotations, we determined average cardiac TTN exon usage de novo from RNA-sequencing. TTNtv in DCM cases were enriched in highly utilised exons and isoforms ($P=2.5\times10^{-4}$) compared to controls. We estimate that TTNtv in highly utilised exons have >93% probability of pathogenicity (likelihood ratio 14) in DCM cases. TTNtv-positive DCM patients had more depressed LV ejection fraction (LVEF: P=0.02), thinner LV walls ($P<0.02$), and a higher incidence of sustained ventricular tachycardia (P=0.001). C-terminus TTNtv were associated with lower LVEF vs N-terminus ($\beta=-18\pm7\%$, p=0.006) and were more common in end-stage disease. TTNtv-positive FHS subjects had increased risk for DCM (RR=16, p=0.008). No change was detected in total TTN mRNA or protein levels in TTNtv-positive hearts.

Incorporation of variant position and exon-specific expression improves interpretation of TTNtv. Most individuals with TTNtv do not develop DCM, but TTNtv in highly utilised, particularly C-terminus exons commonly cause DCM, likely through dominant-negative mechanisms. In DCM patients, presence and position of TTNtv may aid prognostication and management.

## 124

***FLNC* is a novel gene for Dilated Cardiomyopathy in Two Families.** *R.L. Begay[1], A. Martin[2], S.L. Graw[1], D.B. Slavov[1], C.A. Tharp[1], M. Sweet[1], F. Brun[3], K.L. Jones[1], K. Gowan[1], D. Miani[3], G. Sinagra[3], L. Mestroni[1], D.M. Garrity[2], M.R.G. Taylor[1].* 1) Cardiovascular Institute and Adult Medical Genetics Program, University of Colorado Denver, Aurora, CO., United States; 2) Colorado State University, Ft. Collins, CO., United States; 3) University of Trieste Hospital, Trieste, and S. Maria della Misericordia Hospital, Udine, Italy.

Background - Dilated cardiomyopathy (DCM) is an important and frequently genetic form of heart failure. More than 30 DCM genes have been reported, the majority of which encode proteins involved in cytoskeletal and sarcomeric function. Currently, only 30-40% of cases can be attributed to a known DCM gene, motivating the ongoing search for novel disease genes. Methods and Results - We used whole exome sequencing (WES) in a multigenerational DCM family from Northern Italy in whom prior DCM genetic testing had been negative. Pathogenic variants were sought by a combination of bioinformatic filtering and cosegregation analysis of affected relatives. Thirteen gene candidates were identified including one novel variant in *FLNC* (filamin-C gene), previously linked to a skeletal muscle disease phenotype. WES in a second, smaller family from the same region of Northern Italy identified the identical *FLNC* variant present on the same haplotype. The variant is located in the 3' end of the gene and is predicted to disrupt splicing and produce haploinsufficiency for the FLNC protein. Our patients showed no evidence of skeletal myopathy, previously implicated in *FLNC* mutations. *In-situ* hybridization demonstrated cardiac filamin expression in zebrafish and morpholino knockdown of zebrafish *FLNC-b* produced a heart failure phenotype in zebrafish. Conclusion - Using WES, we have identified *FLNC* as a novel DCM gene leading to heart failure. The zebrafish *FLNC* model support a haploinsufficiency model leading to the DCM phenotype.

## 125

**Genome-wide association study on secundum atrial septal defects.** *L. Rodriguez-Murillo[1], M. Parfenov[5], I. Peter[2], W.K. Chung[4], L. Mitchell[3], A.J. Agopian[3], C. Seidman[5], J. Seidman[5], B.D. Gelb[1], Pediatric Cardiac Genomics Consortium.* 1) Mindich Child Health and Development Institute, Departments of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 3) Human Genetics Center, Division of Epidemiology, Human Genetics and Environmental Sciences, University of Texas School of Public Health, Houston, TX; 4) Departments of Pediatrics and Medicine, Columbia University, New York, NY; 5) Department of Genetics, Harvard Medical School, Boston, MA.

Secundum atrial septal defects (ASDs) are among the commonest forms of congenital heart disease (CHD), characterized by deficiencies in the septum primum that contributes to separation of the atria. To elucidate the contribution of common genetic variants to the etiology of ASD, the Pediatric Cardiac Genetics Consortium (PCGC) performed a genome-wide association study (GWAS) using a discovery cohort of 296 subjects with isolated ASD, genotyped with Illumina Omni2.5 microarrays. Our control cohort comprised two dbGaP datasets (916 individuals from a smoking cessation study genotyped with the Illumina Omni2.5 microarray and 2416 from the Framingham heart study genotyped with the Illumina Omni5 microarray). After quality control and principal component analysis (PCA) to correct for population stratification, 1,437,767 SNPs were tested for association on 165 ASD cases and 3330 controls (120 ASD cases showed mixed ethnicity after PCA, therefore were not included in the analysis). The strongest association in cases versus controls attained genome-wide significance using an additive model at SNP kgp10664470 ($P=3\times10^{-10}$; OR = 3.3). ~18 of our ASD cases harbor the minor allele for this SNP. SNPs in linkage disequilibrium with it also showed genome-wide significant P values ($P<1.4\times10^{-8}$), providing robustness to the signal. Subsequently, we replicated this finding by genotyping 144 additional PCGC ASD cases and testing for association against an independent dbGaP control dataset from an eye study (101 ASD cases and 1607 controls were included in the analysis after PCA correction for population stratification). All significant SNPs in the discovery dataset replicated our original findings with a consistent direction of association and a top combined P value of $7.8\times10^{-11}$. These SNPs are located within the *ROBO2* gene, which encodes an immunoglobulin superfamily protein that is evolutionarily conserved and has roles in cardiac cell polarity and morphogenesis in mice, zebrafish and Drosophila. This is a novel gene that had not been identified in previous GWAS performed either in Chinese population (Hu et al. 2013) or Caucasian population (Cordell et al. 2013). Confirmation in larger ASD cohorts and assessment of whether *ROBO2* genetic variation contributes to ASD is underway.

## 126

**Clinical comparison of Kabuki syndrome with *KMT2D* and *KDM6A* mutations.** *N. Miyake[1], E. Koshimizu[1], N. Matsumoto[1], N. Niikawa[2].* 1) Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan; 2) Research Institute of Personalized Health Sciences, Health Science University of Hokkaido, Hokkaido, Japan.

Kabuki syndrome (KS; MIM 147920) is a congenital anomaly syndrome characterized with the characteristic facial appearance including long palpebral fissures and ectropion of the lateral third of the lower eyelids, developmental delay, intellectual disability, prominent digit pads, skeletal anomalies and visceral abnormalities. Until now, KMT2D (previously known as MLL2) and KDM6A are known to cause this syndrome. In our cohort, 81 individuals clinically diagnosed as KS were incorporated. We screened them in two genes and identified a pathogenic mutation in KMT2D or KDM6A in 50 (61.7%) and five (6.2%) patients, respectively. To see the clinical difference among the mutation types, we compared 58 clinical features between two groups classified by these three conditions; (1) Mutation positive and negative groups, (2) KMT2D mutated and KDM6A mutated groups, (3) KMT2D truncating-type and non-truncating type mutation groups. Interestingly, mutation positive groups frequently showed cleft lip/palate. In addition, blue sclera, lower lip pits, lib abnormality, hip joint dislocation, kidney dysfunction, liver abnormality, spleen abnormality, premature thelarche were observed only in mutation positive group. When we compared the patients with KMT2D mutation and KDM6A mutation, high arched eyebrow, short fifth fingers and hypotonia in infancy were frequently observed in KMT2D mutated group. On the other hand, short stature was observed in all patients with KDM6A while about half of the patients with KMT2D mutation. As the KDM6A mutated male mice (Xutx-/Yyty+) were reported to show small body size (Shpargel et al., 2012), short stature could be one of the core feature of KMT2D mutated KS. When we compared the KMT2D truncating-type and non-truncating type mutation groups, prominent ears and hypotonia were statistically frequent in truncating-type group. However, the facial expression in truncating type seemed more typical than other non-truncating group. This might indicate that our current comparison is rather quantitative, not qualitative, so the degree of the each feature was not reflected.

## 127

**Mutations in *KMT2D*, *ZBTB24*, and *KMT2A* in patients with clinical diagnosis of Kabuki syndrome lead to shared epigenetic abnormalities of target genes.** *N. Sobreira[1], L. Zhang[1], C. Ongaco[2], J. Romm[2], M. Baker[1], K. Doheny[2], D. Bertola[3], K. Chong[3], A.B.A. Perez[4], M. Melaragno[4], V. Meloni[4], C. Ladd-Acosta[5], D. Valle[1], H.T. Bjornsson[1].* 1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 3) Unidade de Genética, Instituto da Criança, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil; 4) Genetics Division, Department of Morphology and Genetics, Universidade Federal de São Paulo, Brazil; 5) Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

Kabuki syndrome (KS) (MIM 147920) is a pleiotropic disorder characterized by intellectual disability, postnatal growth retardation, long palpebral fissures with eversion of the lateral third of the lower eyelids and persistence of the fetal fingerpads. About 60 to 74% of the individuals with KS are heterozygous for a LOF mutation in *KMT2D* (previously known as MLL2; MIM 602113) and about 6% have an X-linked dominant disorder caused by mutations *KDM6A* (MIM 300128). The genetic etiologies of the remaining cases are unknown. *KMT2D* is a histone methyltransferase that adds the open chromatin mark, H3K4me3 and *KDM6A* is a histone demethylase that removes the closed chromatin mark, H3K27me3. Both are components of the ASCOM complex involved in transcriptional coactivation of an undefined set of target genes. We used targeted next generation sequencing to sequence *KMT2D*, *KDM6A*, as well as two genes known to cause the ICF syndrome (*ZBTB24*, *DNMT3B*), a phenocopy of KS, plus 5 other candidate genes (*KDM6B*, *MEN1*, *KMT2A*, *KMT2B*, and *HCFC1*) selected on the basis of either interacting or having overlapping function with known KS genes on 29 individuals with a clinical diagnosis of KS. We identified 14 individuals with mutation in *KMT2D* (5/14 confirmed as de novo), 1 individual with a mutation in *ZBTB24* and 3 individuals with missense mutations in *KMT2A* (2 confirmed as de novo). Wiedemann-Steiner syndrome (WSS; MIM 60513) with hairy elbows, short stature, facial dysmorphism, and developmental delay is a rare autosomal dominant disorder caused by heterozygous LOF mutations in *KMT2A* (MIM 159555). Our data suggest that a subset of mutations in *KMT2A* is responsible for a Kabuki like phenotype distinct from classical WSS. We next postulated that the overlapping clinical phenotype might reflect shared epigenetic abnormalities of target genes. Therefore, we investigated genome-scale DNA methylation patterns in our cohort using Infinium 450K BeadChips. We found significant hypermethylation (FWER < 0.1; adjusted for blood cell composition) at 3 genomic regions, near *ZFP57*, *DEGS2* and *LCLAT1*, among our cohort compared to age and sex matched controls. Our results identify a new gene responsible for KS (*KMT2A*); suggest a mechanistic relationship between KS, ICF and WSS; and identify downstream methylation abnormalities in target genes that may be used as a marker of this clinical phenotype to guide the diagnostic process in this genetically heterogeneous group of patients.

## 128

**Noonan syndrome due to RIT1 mutations: further clinical and molecular delineation in 32 cases.** *A. Verloes[1,2], A. Caye[1], A. Dieux Coeslier[3], C. Baumann[1], C. Vincent-Delorme[4], P. Bouvagnet[5], A. David[6], D. Lacombe[7], P. Blanchet[8], B. Isidor[6], M. Rio[9], D. Héron[10], S. Sauvion[11], J.L. Alessandri[12], V. Drouin-Garraud[13], B. Doray[12], N. Pouvreau[13], A. Cavé[1,14].* 1) 1.Department of Genetics, APHP - Robert Debré University Hospital and Denis-Diderot-Robert DEBRE University Hopital and Paris VII University Medical School, Paris, France; 2) INSERM UMR 1141, Robert DEBRE Hospital, Paris, France; 3) Dept of Genetics, Jeanne de Flandre University Hospital, Lille; 4) Dept of Genetics, Regional Hospital, Arras; 5) Dept of Cardiology, University Hospital, Lyon; 6) Dept of Genetics, University Hospital, Nantes; 7) Dept of Genetics, University Hospital, Bordeaux; 8) Dept of Genetics, University Hospital, Montpellier; 9) Dept of Genetics, APHP - Necker-Enfants Malades Hospital, Paris; 10) Dept of Genetics, APHP - La Pitié-Salpétrière Hospital, Paris; 11) Dept of Pediatrics, Jean Verdier Hospital, Bondy; 12) Dept of Genetics, University hospital, Saint Denis de la Réunion; 13) Dept of genetics, University Hospital, Rouen; 14) INSERM U 1131, APHP - Saint Louis University Hospital, Paris.

Noonan syndrome is a heterogeneous dominant disorder, due to mutations in at least 8 different genes involved in the RAS/MAPK signaling pathway. Recently, *RIT1* was shown to be involved in the pathogenesis of some Noonan patients. We report a series of 32 patients from 22 pedigrees with mutations in *RIT1*. The patients show a typical Noonan Gestalt and facial phenotype. Among the 22 probands, 5 % showed postnatal growth retardation, 71 % had congenital heart defect, 37 % had hypertrophic cardiomyopathy, 52% had speech delay, 62 % have learning difficulties, but only 10% had intellectual disability. None of them has major skin anomalies. Compared to the canonic Noonan syndrome phenotype linked to *PTPN11* mutations, *RIT1* mutants appear to be less severely growth retarded and intellectually impaired. Incidence of cardiomyopathy was lower than previously observed. Based on our experience, we estimate that RIT1 may be the cause of 3 to 5 % of Noonan syndrome, and should be prioritized in patients with normal growth and cognitive development.

## 129

**Whole exome sequencing in 78 Noonan syndrome individuals identifies two new candidate genes.** *G.L. Yamamoto[1,2], R. Atique[2], M. Aguena[2], L. Testai[1], M. Buscarilli[1], A. Jorge[3], A.C. Pereira[4], A. Malaquias[3], C.A. Kim[1], M.R. Passos-Bueno[2], D.R. Bertola[1,2].* 1) Unidade de Genética do Instituto da Criança, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo/SP, Brazil; 2) Instituto de Biociências, Universidade de São Paulo, São Paulo/SP, Brazil; 3) Endocrinologia, LIM/25, Faculdade de Medicina da Universidade de São Paulo, São Paulo/SP, Brazil; 4) Instituto do Coração, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo/SP, Brazil.

Noonan syndrome is an autosomal dominant disorder characterized by facial dysmorphisms, short stature and cardiac abnormalities along with deregulation of the RAS/MAPK pathway. It exhibits great genetic heterogeneity, caused by heterozygous mutations in *PTPN11*, *KRAS*, *SOS1*, *RAF1*, *SHOC2*, *NRAS*, *CBL* and *RIT1*, observed in approximately 75% of the patients. We have studied 70 Noonan syndrome probands by WES that were tested negative for the most common mutations in *PTPN11*, *SOS1*, *RAF1*, *KRAS* and *SHOC2*, as well as nine relatives from five families. Segregation of the mutations in candidate genes with the phenotype was further investigated by Sanger sequencing in 10 more relatives. We have identified probable pathogenic mutations in two genes never before associated with Noonan syndrome in 6 probands. Mutations in gene A are segregating with the disease in two families and occurred de novo in one sporadic case. Mutations in gene B are segregating with the disease in one family, occurred de novo in one sporadic patient, and in another sporadic individual segregation status is under analysis. Gene A is not directly associated to the RAS pathway in the literature but it has been previously associated with a tumoral disease, while gene B is known to be directly associated to the RAS pathway involved with Noonan syndrome. We believe that these two new genes could be responsible for the disease in a small fraction of the molecularly undiagnosed individuals with Noonan syndrome and functional studies are underway to further confirm the pathogenicity of the mutations. FAPESP 2011/17299-3; CNPq.

## 130

**NSD1± DNA methylation (DNAm) signature: A novel functional diagnostic tool for Sotos syndrome.** *S. Choufani[1], C. Cytrynbaum[2], B.H.Y. Chung[3], A.L. Turinsky[4,5], D. Grafodatskaya[1], Y.A. Chen[1], H.M. Luk[6], I.F.M. Lo[6], S.T.S. Lam[6], D.J. Stavropoulos[7], B. Gibson[8], M. Reardon[9], M. Brudno[1,5,10], R. Mendoza-Londono[2], D. Chitayat[2], R. Weksberg[1,2].* 1) Program in Genetics and Genome Medicine, The Hospital for Sick Children, Toronto, On, Canada; 2) Div Clin & Metabolic Gen, The Hospital for Sick Children, Toronto, On, Canada; 3) Dept of Paediatrics & Adolescent Med, Li Ka Shing Faculty of Medicine, Hong Kong; 4) Molecular Structure & Function, The Hospital for Sick Children, Toronto, On, Canada; 5) Centre for Computational Medicine, The Hospital for Sick Children, Toronto, On, Canada; 6) Clinical Genetics Service, Department of Health, Hong Kong; 7) Paediatric Laboratory Medicine, Hosp Sick Children, Toronto, ON, Canada; 8) Dept. of Medical Genetics, UBC, Child and Family Research Institute, Vancouver, BC, Canada; 9) Our Lady's Hospital for Sick Children, Crumlin, Dublin 12, Ireland; 10) Department of Computer Science and Donnelly Centre, University of Toronto, Toronto, ON, Canada.

Sotos syndrome (SS) is characterized by somatic overgrowth and intellectual disability. Most SS cases have mutations in NSD1 (nuclear receptor-binding SET domain protein 1), a histone lysine methyltransferase. To determine if NSD1 mutations impact stable epigenetic marks such as DNAm, we compared DNAm in peripheral blood from SS cases with NSD1 mutations (NSD1±; n=20) to controls (n=53) using the Illumina Infinium450methylation BeadChip (450k array). Differential DNAm analysis using non-parametric statistics (with correction for multiple testing) identified a surprisingly high number of differentially methylated (DM) CG sites between SS and controls. The majority (99.3%) of these sites demonstrated loss of DNAm and were distributed across the genome. Using unsupervised hierarchical clustering of the significant DM sites, all SS cases with loss of function mutations in NSD1 clustered as a distinct group. The specificity of this signature was 100%; in comparison to DNAm profiles of 450k data from the GEO database for blood samples (n= 1200). The sensitivity of the NSD1± signature was tested in an independent replication cohort of 19 SS cases from Hong Kong with NSD1 loss of function mutations. The NSD1± signature demonstrated a sensitivity of 100%; highlighting its exceptional power in defining pathogenicity for mutations in NSD1. The classification of single nucleotide substitutions into benign or disease causing (damaging) variants represents an ongoing challenge in clinical diagnostics. To test the hypothesis that the NSD1± DNAm signature will inform the functional classification of NSD1 variants into benign or disease causing, we analyzed 16 cases with reported missense variants in NSD1. Using hierarchical clustering we classified these variants as pathogenic (n=9) or benign (n=7). Clinical re-assessment of 11/16 of these cases, for whom photos and adequate clinical information were available, was conducted by two experienced dysmorphologists (RW and DC) who were blinded to the methylation results. There was 100%; concordance between clinical impression and DNAm data. In comparison, 4/5 algorithms (SIFT, Polyphen-2, etc) were inconsistent in their prediction of pathogenicity. Our data suggest that the NSD1± DNAm signature reflects the functional effect of NSD1 variants on the methylome and can be used as a more robust predictor than existing algorithms for the functional classification of NSD1 variants in overgrowth disorders.

## 131

**A new neurodevelopmental-congenital heart disease syndrome caused by variants in a novel disease gene, TELO2.** *J. You[1,2], N. Sobreira[1], D. Gable[1,2], J. Jurgens[1,2], D. Valle[2,4,5], M. Armanios[2,6], J. Hoover Fong[3,5].* 1) Predoctoral Training Program in Human Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; 2) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; 3) McKusick-Nathans Institute of Genetic Medicine, Greenberg Center for Skeletal Dysplasias, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA; 4) Center for Inherited Disease Research, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21224, USA; 5) Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; 6) Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

TELO2, together with TTI1 and TTI2 forms an evolutionarily conserved complex, the TTT complex, which is thought to function as a chaperone. The TTT complex plays a critical role in maturing and stabilizing phosphoinositide-3-kinase-related protein kinases (PIKKs), which are important signal transducers, involved in genome stability, response to nutritional supply, nonsense mediated decay, gene expression and suppression of tumorigenesis. Langouët et al. (2013) have shown that a missense mutation in TTI2 causes a phenotype characterized by developmental delay and facial dysmorphisms. Here we report a family with three affected siblings (one set of dizygotic twins and a younger brother) with overlapping clinical features including global development delay, short stature, hearing loss, cardiac anomalies and photophobia with decreased visual acuity though structurally normal eyes. Using whole exome sequencing of the siblings and their parents, we identified compound heterozygous variants in TELO2 (p. C367F; p. D720V) segregating with the phenotype. No other candidate genes were identified for the autosomal recessive inheritance model, including homozygous and compound heterozygous. There was no sensitivity to irradiation or mitomycin C in fibroblasts of affected subjects, and the lymphocyte telomere length was normal. We next investigated the effect of these missense mutations on TELO2 protein stability. Immunoblot for TELO2 from the patients showed reduced levels at 30-40% of controls. TTI1 and TTI2 levels were also similarly reduced at 40-50%. This effect was seen in both fibroblasts and lymphoblastoid cells. Despite the reduction of the TTT proteins, the levels of PIKK proteins including ATM, ATR, DNA-PK, mTOR were unchanged. To our knowledge, this is the first report linking TELO2 mutations to a Mendelian disorder. Our findings support that TELO2 missense mutations result in loss of function and disturb TELO2 protein and TTT complex integrity. In light of the associated clinical phenotype, our findings indicate TELO2 functions are relevant to neurodevelopmental and cardiac disease genetics.

## 132

**A novel variant in tenascin-X may be associated with an Ehlers Danlos phenotype in patients with congenital adrenal hyperplasia.** *R. Morissette[1,3], W. Chen[2], Z. Xu[3], J. Dreiling[4], M. Quezado[4], N. McDonnell[3], D. Merke[1,5].* 1) Clinical Center, The National Institutes of Health, Bethesda, MD, USA; 2) PreventionGenetics, Marshfield, WI, USA; 3) Laboratory of Clinical Investigation, The National Institute on Aging, The National Institutes of Health, Baltimore, MD; 4) Laboratory of Pathology, The National Cancer Institute, The National Institutes of Health, Bethesda, MD, USA; 5) The Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA.

Background: Mutations in CYP21A2 result in congenital adrenal hyperplasia (CAH) due to 21-hydoxylase deficiency. Flanking CYP21A2 is TNXB, the gene encoding tenascin-X (TNX), an extracellular matrix glycoprotein that is highly expressed in connective tissue and regulates collagen fibrillogenesis and matrix maturation. TNX deficiency is a known cause of Ehlers Danlos syndrome (EDS), a connective tissue dysplasia. We previously reported that approximately 7% of CAH patients have haploinsufficiency for a contiguous deletion of CYP21A2 and TNXB resulting in CAH-X syndrome with characteristic clinical features of a connective tissue dysplasia, such as joint laxity, cardiac valvular abnormalities, and bifid uvula. In 10 of 330 patients, evidence of a CAH-X phenotype was found; however, a TNXB deletion was not identified. Therefore, we sought to determine the cause of the CAH-X phenotype in these CAH patients. Methods: Dermal fibroblasts and direct tissue from patients and controls were used for biochemical experiments. Sanger sequencing, western blotting, and immunohistochemistry are being used to investigate potential defects in TNX. Immunohistochemical analysis of elastin, fibrillin, and overall extracellular matrix organization in dermal tissue from patients and controls is ongoing. Results: We identified a potentially pathogenic variant c.12174C>G (p.Cys4058Trp) in TNXB in five CAH patients with an EDS phenotype. p.Cys4058Trp is in the fibrinogen, alpha/beta/gamma chain, and C-terminal globular domain of TNX. This variant currently is not found in the 1000 Genomes Project or dbSNP database. PolyPhen-2 software predicted that the variant is probably damaging. Software modeling showed that mutating this highly conserved cysteine disrupts a disulfide bond and likely results in protein misfolding. Western blot analysis in fibroblasts showed that this missense variant does not alter TNX expression compared to controls as expected. Immunohistochemical analysis is underway. Conclusion: A novel variant of a highly conserved cysteine in TNX is likely responsible for at least some of the connective tissue phenotypes found in patients with CAH. Patients with CAH due to 21-hydroxylase deficiency are at risk for also having a connective tissue dysplasia due to TNX deficiency. In addition to the known contiguous gene deletion, other types of TNXB genetic mutations may commonly occur in CAH patients.

## 133

**The impairment of MAGMAS function in human is responsible for a severe skeletal dysplasia.** *C. Mehawej[1,2], A. Delahodde[3], L. Legeai-Mallet[2], V. Delague[4,5], N. Kaci[2], J-P. Desvignes[4,5], Z. Kibar[6,7], J-M. Capo-Chichi[6,7], E. Chouery[1], A. Munnich[2], V. Cormier-Daire[2], A. Mégarbané[1].* 1) Unité de Génétique Médicale et Laboratoire International associé INSERM à l'Unité UMR_S 910, Faculté de Médecine, Université Saint-Joseph, Beirut, Lebanon; 2) Département de Génétique, Unité INSERM U781, Université Paris Descartes-Sorbonne Paris Cité, Fondation Imagine, Hôpital Necker Enfants Malades, Paris 75015, France; 3) University of Paris-Sud, CNRS, UMR 8621, Institute of Genetics and Microbiology, Orsay, 91405, France; 4) Inserm, UMR_S 910, 13385, Marseille, France; 5) Aix Marseille Université, GMGF, 13385, Marseille, France; 6) Center of Excellence in Neuroscience of Université de Montréal, Centre de Recherche du CHU Sainte-Justine, Montréal, Canada; 7) Department of Obstetrics and Gynecology, Université de Montréal, Montréal, Canada.

Impairment of the tightly regulated ossification process leads to a wide range of skeletal dysplasias (SD). Deciphering the molecular basis of a considerable number of SD contributes to the understanding of this complex process. Here, we report the identification of a homozygous mutation in the mitochondria-associated granulocyte macrophage colony stimulating factor-signaling gene *MAGMAS* (NM_016069: c.A226G; p.Asn76Asp) in a novel and severe spondylometaphyseal dysplasia, recently reported by Mégarbané et al. *MAGMAS*, also referred to as *PAM16* (presequence translocase-associated motor 16), is a mitochondria-associated protein involved in preprotein translocation into the matrix. We show that MAGMAS is specifically expressed in trabecular bone and cartilage at early developmental stages and that the mutation leads to an instability of the protein. We further demonstrate that the mutation described here confers to yeast strains a temperature-sensitive phenotype, impairs the import of mitochondrial matrix pre-proteins and induces cell death. Our finding of deleterious *MAGMAS* mutations in an early lethal skeletal dysplasia establishes for the first time a link between a mitochondrial protein and skeletal dysplasias and supports a key role for MAGMAS in the ossification process.

## 134

**Analysis of mutational landscape and genetic heterogeneity in liver cancer with whole genome sequencing.** *A. Fujimoto[1], M. Furuta[1], Y. Shiraishi[2], H.H. Nguyen[1], D. Shigemizu[1], K. Gotoh[3], Y. Kawakami[4], T. Nakamura[5], M. Ueno[6], S. Ariizumi[7], T. Shibata[8], H. Ojima[8], K. Shimada[8], S. Hayami[6], Y. Shigekawa[6], H. Aikata[4], K. Arihiro[4], H. Ohdan[4], S. Marubashi[3], T. Yamada[3], O. Ishikawa[3], M. Kubo[1], S. Hirano[5], M. Yamamoto[7], H. Yamaue[6], K. Chayama[1,4], S. Miyano[2], T. Tsunoda[1], H. Nakagawa[1].* 1) RIKEN Center for Integrative Medical Sciences; 2) Human Genome Center, The Institute of Medical Science, The University of Tokyo; 3) Osaka Medical Center for Cancer and Cardiovascular Diseases; 4) Hiroshima University School of Medicine; 5) Hokkaido University Graduate School of Medicine; 6) Wakayama Medical University; 7) Tokyo Women's Medical University; 8) National Cancer Center.

Primary liver cancers can be generally classified into hepatocellular carcinomas (HCCs) and liver cancer with biliary phenotype (LCB). To elucidate comprehensive genetic landscape of liver cancer, we performed whole genome sequence of sixty HCCs and fifteen LCBs, and identified point mutations, short indels, copy number alternation and rearrangements. While the genome-wide substitution pattern of LCBs that developed in livers affected by hepatitis overlapped with those of HCCs, the substitution patterns of LCBs without a hepatitis background diverged and were closer to liver fluke-related cholangiocarcinoma and pancreatic cancer, suggesting the influence of hepatitis and/or cellular origin on the substitution pattern. Whole genome sequencing and the subsequent validation study identified recurrent mutations in chromatin regulators, KRAS (specifically mutated in hepatitis-free LCBs), and other new pathways, in addition to known cancer-related genes such as TP53 and CTNNB1. Examination of intra-tumor heterogeneity by deep sequencing suggests that the distribution of clonal proportion reflect tumor type, and that mutations in various genes can contribute to tumor initiation in liver.

## 135

**Abundant somatic L1 retrotransposition occurs early during colorectal and pancreatic tumorigenesis.** *S. Solyom[1], A.D. Ewing[2], A. Gacita[1], L.D. Wood[3], F. Ma[1], A. Makohon-Moore[3], D. Xing[3], R. Hruban[3], C.A. Iacobuzio-Donahue[3], S.J. Meltzer[4], B. Vogelstein[5], K.W. Kinzler[5], H.H. Kazazian[1].* 1) Johns Hopkins University School of Medicine, Baltimore, MD; 2) Mater Research Institute, University of Queensland, Australia; 3) Department of Pathology, The Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins Medical Institutions, Baltimore, MD; 4) The Johns Hopkins Univ. School of Medicine & Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD; 5) The Ludwig Center and The Howard Hughes Medical Institute at Johns Hopkins Kimmel Cancer Center, Baltimore, MD.

The somatic mobilization of retroelements in the cancer genome has only recently been established as a new mutational phenomenon. In particular, Long INterspersed Element-1 (L1) retrotransposition has been observed in epithelial cancers. L1s are autonomous mobile elements that comprise 17% of the human genome and retrotranspose by a 'copy and paste' mechanism via an RNA intermediate. Here, we investigate the spatio-temporal map of these integration events in gastrointestinal tumors. We studied DNA from 4 colon cancer patients who had been previously diagnosed with colonic polyps, from 5 patients with colorectal dysplasia and cancer arising in inflammatory bowel disease, and from 7 patients with pancreatic carcinoma. Metastases were available in multiple instances. After dissection of abnormal from normal tissue, next generation L1-targeted resequencing (L1-seq) was carried out on DNA from these cases. After PCR-validation and Sanger sequencing of putative insertions, we found for the first time that pancreatic cancer is permissive for L1 mobilization and that certain pre-cancerous lesions are mutagenized by L1 insertions. We have so far validated 80 somatic insertions in these colon tumors, of which half occurred in adenomas and IBD dysplasias, while 24 insertions were validated in pancreatic cancers. However, by extrapolation, insertions in adenomas and matched colon cancers numbered in the hundreds. Surprisingly, multiple insertions in IBD dysplasias were also present in their paired carcinomas, and many insertions in primary colon and pancreatic cancers were also present in their paired metastases. In addition, among insertions tested in multiple sections of the same tumor, the majority were present in all sections of the tumors. Numerous genes were targeted by L1 insertions, including within exons. Together, these data suggest that: 1) insertions occur clonally; 2) somatic retrotransposition occurs early during the development of some gastrointestinal cancers; and 3) L1 insertions show potential as novel biomarkers of neoplastic disease progression. However, it is not yet known whether some insertions are cancer drivers, or to what extent retrotransposition contributes to genetic instability.

## 136

***Cis*-regulatory drivers in colorectal cancer.** *H. Ongen[1], C.L. Andersen[2], J.B. Bramsen[2], B. Oster[2], M.H. Rasmussen[2], P.G. Ferreira[1], J. Sandoval[3], E. Vidal[3], N. Whiffin[4], I. Tomlinson[5], R.S. Houlson[4], M. Esteller[3], T.F. Orntoft[2], E.T. Dermitzakis[1].* 1) Department of Genetics and Development, CMU, University of Geneva, Geneva., Switzerland; 2) Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark; 3) Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), 08908 Barcelona, Catalonia, Spain; 4) Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK; 5) Nuffield Department of Clinical Medicine and Oxford NIHR Comprehensive Biomedical Research Centre, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.

The *cis*-regulatory effects responsible for cancer development have not been as extensively studied as the perturbations of the protein coding genome in tumorigenesis. In order to better characterise colorectal cancer (CRC) development we are conducting an RNA-seq experiment of 300 matched tumour and adjacent normal colon mucosa samples from CRC patients of Danish origin, which are also germline genotyped. Preliminary analysis with 103 matched samples show that there are 1626 differentially expressed genes (FDR = 5%, fold change $\geq 2$) between normal colon and cancer. We identify multiple regions on nearly all chromosomes where the correlation of expression for proximal genes is significantly increased in the tumours when compared to normals. On average there are 688 significant allele-specific expression (ASE) signals (FDR = 1%) per sample. The proportion of sites that have an ASE effect is significantly more in tumours. By investigating ASE we show that the germline genotypes remain important determinants of allelic gene expression in tumours. Utilizing the changes in ASE in 103 matched pairs of samples we discover 71 genes with excess of somatic cis-regulatory effect in CRC, suggesting a cancer driver role. We correlated genotypes and gene expression to identify expression quantitative trait loci (eQTLs) in 103 normal and tumour tissues and find 1693 and 948 *cis*-eQTLs in normals and tumours, respectively. We estimate that 36% of the tumour eQTLs are exclusive to CRC and show that this specificity is partially driven by increased expression of specific transcription factors and changes in methylation patterns. We find tumour-specific eQTLs are more enriched for low CRC genome-wide association study (GWAS) p-values than shared eQTLs, which suggests some of the variants discovered in GWAS are *cis*-regulatory variants active specifically in the tumour. Importantly tumour specific eQTL genes also accumulate more somatic mutations when compared to the shared eQTL genes, raising the possibility that they constitute germline-derived cancer regulatory drivers. Collectively the integration of genome and the transcriptome reveals a substantial number of putative somatic and germline *cis*-regulatory drivers.

## 137

**Somatic mutations modulate ceRNA drivers of tumorigenesis.** *J. He[1,2,3], H.-S. Chiu[4], P. Sumazin[4], A. Califano[1,2,3].* 1) Department of Systems Biology, Columbia University, New York, NY; 2) Center for Computational Biology and Bioinformatics, Columbia University, New York, NY; 3) Department of Biomedical Informatics, Columbia University, New York, NY; 4) Texas Children's Cancer Center, Baylor College of Medicine, Houston, TX.

Pan-cancer studies have shown that competitive endogenous RNA (ceRNA) networks can cooperate with chromosome instability and abnormal DNA methylation in tumors to dysregulate tumor suppressors and oncogenes. However, ceRNA cooperative association with mutations in cancer has not been studied. Integrating data from TCGA and ENCODE, we show that the cooperation between ceRNA interactions and mutations of unknown function contribute to the dysregulation of cancer genes. We integrated ceRNA networks and mutations in an attempt to mechanistically recover missing genomic variability of cancer genes in TCGA breast cancer biopsies. Genes have missing genomic variability in a tumor dataset when their dysregulation cannot be explained through profiling of their DNA locus. Using a group lasso regression model we showed that ceRNA drivers cooperating with somatic mutations, CNV, and methylation, could account for a large fraction of the missing genomic variability of cancer genes in breast cancer. Moreover, using a greedy-forward optimization algorithm, we identified ceRNA driver mutations that could potentially drive tumorigenesis through the ceRNA mechanism. Furthermore, we showed that driver ceRNA mutations are enriched in known and predicted binding sites of transcription factors and microRNAs. In summary, our results suggest that somatic mutations, often of unknown function, cooperate with ceRNA regulators to alter the expression of cancer genes in breast cancer tumors.

## 138

**Divergence between high metastatic tumor burden and low circulating tumor DNA concentration in metastasized breast cancer.** *M. Heidary[1], M. Auer[1], P. Ulz[1], E. Heitzer[1], E. Petru[2], C. Gasch[3], S. Riethdorf[3], O. Mauermann[3], I. Lafer[1], G. Pristauz[2], S. Lax[4], K. Pantel[3], JB. Geigl[1], MR. Speicher[1].* 1) Institute of Human Genetics, Medical University of Graz, Harrachgasse 21/8, A-8010 Graz, Austria; 2) Department of Obstetrics and Gynecology, Medical University of Graz, Auenbruggerplatz 14, A-8036 Graz, Austria; 3) Institute of Tumor Biology, University Medical Center Hamburg Eppendorf, Martinistr. 52, D-20246 Hamburg, Germany; 4) Department of Pathology, General Hospital Graz West, Goestingerstrasse 22, A-8020 Graz, Austria.

Recently there has been considerable interest in circulating tumor DNA (ctDNA) as non-invasive biomarker in patients with cancer. Although recent studies have shown that ctDNA is an informative and highly sensitive biomarker for monitoring the tumor burden dynamics and treatment responses in metastatic cancer patients, current knowledge of the dynamic range of ctDNA in patients with metastatic breast cancer is limited. To address this issue, we studied the role of ctDNA in patients with metastatic breast cancer and analyzed 74 plasma DNA samples from 58 patients. We used a microfluidic device to analyze size distribution of the plasma DNA and whole genome sequencing (plasma-Seq) to identify copy number changes in the plasma. Highly variable AFs of mutant fragments were detected through analyses of 74 plasma samples from 58 patients, and did not reflect the tumor burden in all cases. To show that tumor burden is not necessarily reflected by mutated AFs we analyzed an index patient in detail. This index patient had more than 100,000 circulating tumor cells (CTCs) in three serial blood analyses and we comprehensively analyzed the primary tumor, metastatic deposits, single and pools of CTCs, and ctDNA using whole-genome, exome, or targeted deep sequencing. Accurate evaluation of the allele fraction (AFs) of mutated DNA fragments was performed using targeted deep-sequencing. Sequencing of four different regions of the primary tumor and three metastatic lymph node regions revealed genetically homogeneous cancer disease. Subsequently, detailed analyses of 551 CTCs demonstrated a high degree of similarities to the primary tumor and metastases and verified the genetically homogeneous cancer. However, despite the extraordinarily high CTC number ctDNA analyses detected a very low mutant AF of only 2-3% in each of the serial samples, which did not reflect the tumor burden or the dynamics of this progressive disease. These data suggest a highly variable range of ctDNA in patients with metastatic breast cancer, which may have an impact on the use of ctDNA as a predictive and prognostic biomarker.

## 139

**Extrachromosomal driver mutations in glioblastoma and low grade glioma.** *S.I. Nikolaev[1], F. Santoni[1,2], M. Garieri[1], P. Makrythanasis[1], E. Falconnet[1], M. Guipponi[2], A. Vannier[2], I. Radovanovic[3,4], F. Bena[2], K. Schaller[3,4], V. Dutoit[5], V. Clement-Schatlo[3,4], P-Y. Dietrich[5], S.E. Antonarakis[1,6].* 1) GeDev, University of Geneva, Geneva, Switzerland; 2) Geneva University Hospitals - HUG. Service of Genetic Medicine. 4 Rue Gabrielle-Perret-Gentil, 1211 Geneva 4, Switzerland; 3) Department of Clinical Neuroscience. University of Geneva Medical School. 1 rue Michel Servet, 1211 Geneva 4, Switzerland; 4) Department of Neurosurgery, Geneva University Hospitals - HUG. 4 Rue Gabrielle-Perret-Gentil, 1211, Geneva 4, Switzerland; 5) Center of Oncology. Geneva University Hospitals - HUG. 4 Rue Gabrielle-Perret-Gentil, 1211, Geneva 4, Switzerland; 6) IGE3 institute of Genetics and Genomics of Geneva. 1 rue Michel Servet, 1211.

Alteration of the number of copies of Double Minutes (DMs) with oncogenic EGFR mutations in response to tyrosine kinase inhibitors (TKIs) is a novel adaptive mechanism of glioblastoma. In this study we provide evidence that such mutations in DMs, called here Amplification Linked Extrachromosomal Mutations (ALEMs), originate extrachromosomally and could therefore be completely eliminated from the cancer cells. By exome sequencing of 7 glioblastoma patients we revealed ALEMs in EGFR, PDGFRA and other genes. These mutations together with DMs were lost by cancer cells in culture. We confirmed the extrachromosomal origin of such mutations by showing that wild type and mutated DMs may coexist in the same tumor. Analysis of 4198 tumors suggested the presence of ALEMs across different tumor types with the highest prevalence in glioblastomas and low grade gliomas. The extrachromosomal nature of ALEMs explains the observed drastic changes in the amounts of mutated oncogenes (like EGFR or PDGFRA) in glioblastoma in response to environmental changes.

## 140

**Automated tumor phylogeny reconstruction using multi-sample deep sequencing somatic variants.** *V. Popic[1], R. Salari[1], D. Kashef-Haghighi[1], D. Newburger[2], R. West[3], S. Batzoglou[1].* 1) Department of Computer Science, Stanford University, Stanford, CA; 2) Biomedical Informatics Training Program, Stanford University, Stanford, CA; 3) Department of Pathology, Stanford University School of Medicine, Stanford, CA.

Numerous studies have shown tumors to be highly heterogeneous, consisting of cell subpopulations with distinct somatic mutational profiles. Tumor heterogeneity is often studied by comparison of multiple tumor samples that are extracted from a single patient either at different points in time during cancer development or from different regions of the same tumor or its metastases. Most existing multi-sample studies infer phylogenetic cancer cell lineage trees either manually or with classical species phylogenetic approaches that do not model sample heterogeneity. Here we present SMutH, *S*omatic *Mut*ation *H*ierarchies, a novel computational method that automates the phylogenetic inference of cancer progression from multiple somatic samples. Our method avoids the common assumption of clonal homogeneity of samples and is able to reconstruct the lineage relationships even when each sample is a heterogeneous mixture of cells. SMutH uses variant allele frequencies (VAFs) of somatic SNVs obtained by deep sequencing to reconstruct multi-sample cell lineage trees and infer the subclonal composition of the samples. SMutH clusters SNVs based on their VAFs and presence patterns across samples and incorporates the resulting clusters into an evolutionary constraint network, which encodes all possible precedence relationships among SNV clusters. In order to trace cell lineage trees and identify sample subclones, the constraint network is searched for phylogenetically valid spanning trees. We evaluated SMutH on two published datasets of clear cell renal cell carcinoma (ccRCC) (Gerlinger et al 2014) and high-grade serous ovarian cancer (HGSC) (Bashashati et al 2013), as well as on simulated data. We found that our method is highly effective in reconstructing the underlying cell lineage phylogenies in real data and simulations. The trees generated by SMutH were nearly identical topologically to the published ccRCC trees. For the HGSC dataset, SMutH produced trees with better support from the data (as confirmed by manual inspection). SMutH also revealed additional heterogeneity in the samples of both studies. In particular, SMutH identified subclones in one more sample of the ccRCC study (in addition to the reported six samples) and three samples of the HGSC study, all supported by the data, demonstrating the need for phylogeny inference methods specialized for heterogeneous cancer datasets.

## 141

**Development and validation of a ultra-high depth FFPE targeted exome sequencing platform for routine cancer patient care.** *K. Chen, F. Meric-Bernstam, H. Zhao, Q. Zhang, N. Ezzeddine, L. Tang, P. Song, Y. Qi, Y. Mao, T. Chen, Z. Chong, W. Zhou, X. Zheng, A. Johnson, S. Kopetz, M. Davies, J. DeGroot, S. Moulder, K. Aldape, M. Routbort, R. Luthra, K. Shaw, J. Mendelsohn, G. Mills, A. Eterovic.* The University of Texas MD Anderson cancer center, Houston, TX.

Although recent studies have indicated the potential of revolutionizing cancer patient care based on routine genomic sequencing, the community is still at an early stage of developing an optimized, practical approach that can deliver highly sensitive detection of actionable genomic aberrations at a reasonable cost. We established an ultra-deep targeted sequencing platform for 201 cancer genes to identify DNA alterations that would inform on disease status or could be potentially actionable in clinical tumor samples. Here we emphasize the significance and the direct correlation between depth of sequencing, input DNA and the ability to call rare mutational events with confidence and accuracy based on results in test samples. We then assayed 515 tumor samples and matched germline (blood) from 12 disease sites in 475 patients. With a mean haploid coverage of over 900x, we identified 4793 non-synonymous mutations, at a false discovery rate < 2.2%. Our results were highly concordant with test results obtained from a CLIA compliant hotspot panel, but identified alterations in potentially clinically actionable genes in twice as many patients. About 15.2% mutations are in low (< 10%) allele frequency with a landscape similar to that of high frequency mutations, which indicated their potential relevance for clinical decision making. In conclusion, our results indicated that an ultra-deep targeted sequencing approach can potentially impact both research and patient care through robust identification of low-level mutations that are potentially relevant for clinical decision making and that could be missed by other sequencing approaches that do not allow high depth of sequencing.

## 142

**Novel insights regarding the pathogenesis and treatment of Pseudoxanthoma Elasticum.** *S.G. Ziegler[1,2], C.R. Ferreira[3], A.B. Pinkerton[4], J.L. Millan[4], W.A. Gahl[3], H.C. Dietz[1,2].* 1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) HHMI, Chevy Chase, MD; 3) MGB, NHGRI, NIH, Bethesda, MD; 4) Sanford-Burnham Medical Research Institute, La Jolla, CA.

Biallelic mutations in *ABCC6*, an ATP-dependent transporter whose ligand remains unknown, classically cause pseudoxanthoma elasticum (PXE) characterized by adult-onset elastic fiber calcification in the eyes, skin, and vasculature. Patients with *ABCC6* mutations can also develop a more severe phenotype with myocardial infarction and stroke by 3 months of age that is indistinguishable from generalized arterial calcification of infancy (GACI). GACI is more commonly caused by biallelic mutations in *ENPP1* that encodes an extracellular enzyme that degrades ATP into AMP and pyrophosphate. We reasoned that elucidation of the mechanism for this striking locus heterogeneity might inform the pathogenesis of *ABCC6*-associated PXE. A cross between Abcc6 knockout (KO) mice and Enpp1-targeted mice revealed strong evidence for genetic interaction; Abcc6 KO mice with one mutated Enpp1 allele showed acceleration and worsening of the calcification phenotype, as assayed by microCT and echocardiogram. In contrast, Abcc6 KO mice with two targeted Enpp1 alleles were indistinguishable from Enpp1 homozygotes, suggesting that Abcc6 acts upstream of Enpp1. The strong expression of Abcc6 in the liver led to the prevailing view that peripheral tissue calcification in PXE reflects failed liver secretion of an inhibitor of calcification that acts predominantly in an endocrine manner. Contrary to this view, we found that fibroblasts from patients with biallelic mutations in either *ABCC6* or *ENPP1* had increased tissue non-specific alkaline phosphatase activity (TNAP) and showed a strong tendency for induced calcification *in vitro* that was prevented by a TNAP inhibitor. Breakdown products of ATP closely regulate TNAP activity suggesting that PXE, like GACI, is caused by defects in extracellular ATP metabolism. Our generation and use of conditional Abcc6-targeted mice revealed that liver-specific deletion using an Albumin-Cre driver failed to recapitulate the PXE calcification phenotype that was observed upon constitutive Abcc6 deletion using CMV-Cre. While a similar strategy was used to exclude the independent relevance of endothelial, vascular smooth muscle and marrow-derived cells, and pericytes, additional studies will explore the candidacy of other lineages. Taken together, these data suggest that PXE is caused by a reversible liver-independent and perhaps cell-autonomous mechanism that induces defects in local extracellular ATP metabolism upstream of Enpp1.

## 143

**Decipher Mitochondrial Disorders using Exome Sequencing.** *R. Kopajtich[1,2], T. Haack[1,2], L. Kremer[1,2], C. Biagosch[1,2], B. Haberberger[1,2], T. Wieland[1,2], T. Schwarzmayr[1,2], P. Freisinger[3], T. Klopstock[4], J. Mayr[5], W. Sperl[5], M. Minczuk[6], T.M. Strom[1,2], T. Meitinger[1,2], H. Prokisch[1,2].* 1) Institute of Human Genetics, Helmholtz Zentrum München, Munich / Neuherberg, Germany; 2) Institute of Human Genetics, Technische Universität München, Munich, Germany; 3) Department of Pediatrics, Community Hospital Reutlingen, Germany; 4) Department of Neurology, Ludwig-Maximilians Universität München, Munich, Germany; 5) Department of Pediatrics, Paracelsus Medical University Salzburg, Salzburg, Austria; 6) MRC Mitochondrial Biology Unit, Cambridge, United Kingdom.

Mitochondrial disorders are a genetically and clinically highly heterogeneous group of diseases characterized by defective oxidative phosphorylation. Despite good progress in the field, most disease causing mutations still have to be identified. During the course of three years, we applied whole exome sequencing to investigate more than 400 unrelated individuals with a suspected mitochondrial disorder and provided molecular diagnoses to 200 patients with mutations affecting almost 100 distinct genes. In a quarter of patients, we identified mutations in known disease genes. In another quarter of patients, we identified mutations in genes previously not associated with mitochondrial disorders. Mutations in the majority of genes are rare and could be identified due to loss-of-function alleles in evolutionary conserved genes such as *MGME1*, an exonuclease involved in mitochondrial replication. Mutations in other genes are more frequent, with *ACAD9* being the most common finding with more than 15 cases, providing statistical evidence for the association with isolated respiratory chain complex I deficiency. Diagnostic challenges are patients with recessive mutations in more than one gene resulting in a compound clinical phenotype. Evolving topics are tRNA modifying enzymes (*ELAC2*, *MTO1* and *GTPBP3*) and tRNA synthetases, both involved in the translation of mitochondrial proteins as well as cofactor metabolism defects. For more than 30 patients (out of 200), the molecular diagnosis offered rational therapeutic options. In summary, the genetically heterogeneous group of mitochondrial disorders is an example par excellence for the application of genome wide sequencing which is underway to be implemented at an early stage in the routine diagnostic work-up of pediatric patients suffering from genetically unclear metabolic conditions.

## 144

**Distinct clinical phenotypes in two unrelated patients with mutations in the *TRNT1* gene encoding tRNA nucleotidyl transferase.** *F. Sasarman[1, 2], I. Thiffault[2, 3], W. Weraarpachai[1], S. Salomon[1], C. Maftei[2], J. Gauthier[2], N. Webb[1, 2], O. Elpeleg[4], C. Brunel-Guitton[2], G. Mitchell[2], E.A. Shoubridge[1].* 1) Molecular Neurogenetics, Montreal Neurological Institute and McGill University, Montreal, Quebec, Canada; 2) Medical Genetics, Pediatrics, CHU Sainte-Justine, Université de Montréal, Montreal, Quebec, Canada; 3) Center for Pediatric Genomic Medicine, Children's Mercy Hospital, Kansas City, MO, USA; 4) Genetic and Metabolic Diseases, Hadassah Medical Center, Jerusalem, Israel.

Addition of the trinucleotide CCA to the 3'end of transfer RNAs (tRNAs) is required for amino acid attachment, tRNA positioning on the ribosome and translation termination. The enzyme responsible is TRNT1 (tRNA nucleotidyl transferase), active both in the cytoplasm and mitochondria. We describe the first identification of mutations in the *TRNT1* gene [MIM 612907], using exome sequencing in two unrelated patients with largely non-overlapping clinical phenotypes. Patient 1, a girl born to consanguineous parents, presented at 3 weeks of age with a crisis of lactic acidosis, evolved to severe developmental delay, hypotonia, microcephaly, seizures, progressive cortical atrophy, neurosensorial deafness, sideroblastic anemia, renal Fanconi syndrome and nephrocalcinosis and died at 21 months during a febrile episode with severe lactic acidosis and hepatic failure. Patient 2, a boy with a milder systemic phenotype, presented at 3.5 years with gait ataxia, dysarthria, gross motor regression, hypotonia, ptosis, horizontal ophthalmoplegia, and had abnormal signals in brainstem and dentate nuclei. He never had seizures or acidotic crises. Exome sequencing of both patients revealed mutations in the *TRNT1* gene at evolutionary conserved positions: a homozygous c.C443T; p.A148V mutation in Patient 1, and combined heterozygous mutations c.383A>G; p.D128G and c.518A>T; p.Y173F in Patient 2. Levels of mutant TRNT1 protein in fibroblasts from Patient 1 were 10% of control, while they were normal in fibroblasts from Patient 2, suggesting a possible genotype-phenotype correlation. Muscle from Patient 1 showed a generalized decrease in the enzymatic activity of all mitochondrial respiratory chain complexes, while mitochondrial translation, respiratory chain assembly and function were normal in fibroblasts despite reduced levels of TRNT1 protein. Knockdown of TRNT1 to immunologically-undetectable levels abolished mitochondrial translation in patient fibroblasts, and had a differential effect on the steady-state levels of individual tRNAs: mitochondrial tRNASer(AGY), which has the most non-canonical structure, was undetectable, while cytoplasmic tRNALys(UUU) and tRNAMet were 30% of control. In control muscle, TRNT1 levels are 10% of the levels in fibroblasts, suggesting that muscle may be particularly vulnerable to *TRNT1* mutations.

## 145

**Mutation in the tRNA-modification enzyme GTPBP3 causes hypertrophic cardiomyopathy with abnormal respiratory chain assembly.** *M. METODIEV[1], Z. ASSOULINE[2], M. RIO[2], F. FEILLET[3], B. MOUSSON de CAMARET[4], D. CHRETIEN[1], A. MUNNICH[1,2], A. RÖTIG[1].* 1) INSERM U1163, Université Paris Descartes-Sorbonne Paris Cité, Institut Imagine, 24 Boulevard du Montparnasse, 75015 Paris, France; 2) Departments of Pediatrics and Genetics, Hôpital Necker-Enfants Malades, 149 Rue de Sèvres, 75015 Paris, France; 3) Service de médecine infantile, Hôpital d'enfants de Brabois, CHU de Nancy, Rue du Morvan, 54511 Vandoeuvre-les Nancy, France; 4) des Maladies Héréditaires du Métabolisme, CHU de Lyon, 59 bd Pine, 69677 Bron.

Hypertrophic cardiomyopathy caused by oxidative phosphorylation deficiency has been hitherto ascribed to either altered respiratory chain assembly or impaired translation of proteins encoded by the mitochondrial genome. Exome sequencing in two sibs, with severe hypertrophic cardiomyopathy and combined respiratory chain defect allowed to identify a homozygous frameshift mutation (c.32_33delinsGTG) in GTPBP3, a protein involved in the post-transcriptional modification of several mitochondrial transfer RNAs. This frameshift mutation caused a premature stop codon, abolished the GTPBP3 protein and resulted in an impairment of fully assembled complex I. An abnormal accumulation of an assembly intermediate of complex I with a mild accumulation of the F1F0 subcomplex of complex V was also observed The defect in respiratory chain assembly could be reversed by overexpression of the two wild-type GTPBP3 isoform cDNAs, namely GTPBP3ins8 and GTPBP3del8 in fibroblasts of the affected child. In yeast, the homolog of human GTPBP3, Mss1p, forms a heterodimeric complex with Mto1p and catalyzes the 5-carboxymethylaminomethylation of the wobble uridine base in three mitochondrial tRNAs, namely tRNAGln, tRNAGlu, and tRNALys. Interestingly, mutations in MTO1 have been also shown to cause severe hypertrophic cardiomyopathy in human. In conclusion, we report here that mutation in either subunit of the GTPBP3-MTO1 complex causes severe hypertrophic cardiomyopathy in human.

## 146

**Application of cellular O-linked glycomics analysis for the diagnosis of protein glycosylation disorders.** *M. He[1,2], X. Li[1,2], M. Raihan[1,2], L. Tan[1,2], M. Bennett[1,2], W. Gahl[3], M. Davids[3], M. Kane[3], C.F. Boerkoel[3].* 1) Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA; 2) Palmieri Metabolic Disease Laboratory, Children's Hospital of Philadelphia, Philadelphia, PA; 3) Undiagnosed Disease Program, NHGRI, NIH, Bethesda, MD.

The glycosylation of protein is an important post-translational modification in many biological systems. Our laboratory has developed the plasma O-glycan assay for diagnosis of the multiple glycosylation disorder, congenital disorder of glycosylation (CDG) type II, using mass spectrometry technology. However, O-linked glycoproteins in peripheral blood only represent a small fraction of O-linked glycoproteome in humans and O-glycan species released from total glycoproteins in plasma are very limited with minimum amount of the core 2 species. Furthermore, it is known that alteration in glycan structure often shortens the half-life of glycoproteins in circulation and the liver plays an active role in removing truncated glycans from the peripheral blood through glycoprotein scavenge pathway. Thus O-linked glycan analysis in plasma only is likely inadequate for detecting defects in O-linked protein glycosylation. In this study, we describe cellular O-glycomics analysis in cultured fibroblast and its application to detect O-linked glycosylation disorders. We found that a board spectra of O-linked GalNAc glycosylation were detected after they were released from fibroblast total glycoproteins using the beta elimination method. The quantification of both mucin core 1 and core 2 species allows measurement of defects in multiple steps in O-linked glycan biosynthesis pathway in fibroblast lines from known CDG patients including GALNT3-CDG, COG7-CDG as well as deficiencies in nucleotide sugar synthesis and transport. We have measured fibroblast O-glycomics in 33 fibroblast lines from patients with borderline changes of O-linked protein glycosylation in plasma. 4 of them (12%) showed profound deficiency in O-linked protein glycosylation that would not be possible to pinpoint in plasma based analysis. Among 43 cell lines from patients with other indications of possible CDG based on unusual N-linked glycan or free glycan profiles of body fluids or clinical findings, 13 of them (30%) showed O-glycomics changes that were significantly deviated from the control group. Our study demonstrates that cellular O-linked glycomics analysis is more informative and sensitive for detecting O-linked or multiple glycosylation disorders comparing to plasma O-glycan analysis alone. Therefore it is an important diagnostic tool for this largely under-recognized group of CDGs.

## 147

**Metabolic diversion towards non-toxic metabolites for therapy of primary hyperoxaluria type 1.** *R. Castello[1], R. Borzone[1], P. Annunziata[1], P. Piccolo[1], N. Brunetti-Pierri[1,2].* 1) Telethon Institute of Genetics and Medicine, Naples, Italy; 2) Department of Translational Medicine, Federico II University of Naples, Italy.

Primary hyperoxaluria type 1 (PH1) is an inborn error of liver metabolism due to deficiency of peroxisomal enzyme alanine:glyoxylate-aminotransferase (AGT) which catalyzes the conversion of glyoxylate to glycine. In PH1 patients, glyoxylate cannot be efficiently converted into glycine and is instead oxidized to oxalate, leading to hyperoxalemia and hyperoxaluria. In turn, this causes the deposition of insoluble calcium oxalate in the kidney and in other tissues, leading to nephrolithiasis, nephrocalcinosis, kidney failure, and systemic tissue damage. Combined liver/kidney transplantation is the only therapeutic strategy available to prevent disease progression. The role of glyoxylate reductase/hydroxypyruvate reductase (GRHPR) in glyoxalate oxidation and oxalate detoxification has been controversial. We hypothesize that GRHPR overexpression results in significant long-term reduction of hyperoxaluria in PH1. To test this hypothesis, we injected $Agxt^{-/-}$ mice with an helper-dependent adenoviral vector expressing murine GRHPR (HDAd-GRHPR) in hepatocytes. The injection of HDAd-GRHPR resulted in significant reduction of hyperoxaluria and concomitant increase of serum glycolate that was not associated with evidence of toxicity. Glutamate-pyruvate transaminase (GPT) in the cytosol efficiently transaminate glyoxylate using L-glutamate and L-alanine as amino-group donors. We hypothesize that GPT overexpression will steer more glyoxylate towards transamination to diminish oxalate production. To test this hypothesis, we injected $Agxt^{-/-}$ mice with a helper-dependent adenoviral vector expressing murine GPT (HDAd-GPT) in hepatocytes. The injection of HDAd-GPT also resulted in significant long-term reduction of hyperoxaluria. In summary, the results of this study show that metabolic diversion towards non-toxic metabolites have potential for treatment of hyperoxaluria. Besides gene transfer, such diversion could be obtained with small molecules increasing GRHPR and/or GPT expression. Moreover, vector-mediated GRHPR or GPT overexpression may be an alternative or adjunctive strategy to ehnance efficiency of gene replacement therapy for PH1. In addition, this approach may be valuable for patients harboring null mutations in the gene encoding AGT which are at increased risk for an immune reaction against the transgene product encoded by vector-transduced cells.

## 148

**Transcriptome and microRNA profiling reveals deregulated microRNAs and mRNAs in the brain of neuronopathic Gaucher disease mice.** *Y. Sun[1,2], N. Dasgupta[1], Y. Xu[1,2], B. Liou[1], R. Li[1,2], Y. Peng[1], M. Pandey[1,2], S. Tinch[1], V. Inskeep[1], G.A. Grabowski[3].* 1) Divison of Human Genetics, Cincinnati Children's Hospital, Cincinnati, OH; 2) Department of Pediatrics, University of Cincinnati College of Medicine Cincinnati, OH 45229; 3) Synageva BioPharma Corp, Lexington, MA 02421.

Gaucher disease is caused by deficiency of lysosomal acid β-glucosidase (GCase) leading to accumulation of glucosylceramide (GC) and glucosylsphingosine (GS) in the viscera and CNS. CNS pathogenesis results from neuronal degeneration propagated by the toxic effects of GS and GC. To understand the pathogenic mechanisms in neuronopathic Gaucher disease (nGD), global profiles of differentially expressed mRNAs (DEGs) and microRNA (DEmiRs) were analysed using a viable nGD mice (4L;C*). This model develops neurological deficits analogous to sub-acute human nGD. The brain of 4L;C* mice accumulates GC and GS with resultant inflammation and decreased mitochondrial function. DEGs and DEmiRs were analyzed using isolated RNA from cerebral cortex (CO), brain stem (BS), midbrain (MID) and cerebellum (CB) of age and strain matched 4L;C* and WT mice using RNASeq. These brain regions were also analyzed in 4L;C* mice treated with isofagomine, a pharmacologic chaperone for GCase. Analyses showed region specific and common DEGs and DEmiRs in CO, BS, MID or CB. The predicted DEmiR-target DEGs with inverse correlations of microRNA represented about 46%, 47%, 58% and 51% of total DEGs in CO, BS, MID and CB, respectively. The DEGs regional specific deregulation in 4L;C* brains was significantly altered after isofagomine treatment. Isofagomine treatment also normalized part of the abnormalities of DEmiRs and their target DEGs, but also induced additional changes in DEmiRs expression. Total altered DEGs from 4L;C* brain regions were classified to two major groups: inflammatory and non-inflammatory. The inflammatory DEGs were about 25% of total DEGs in each brain regions. IPA analyses of all brain regions showed that top functional pathways of inflammatory DEGs include the roles of macrophages, dendritic cell maturation, acute phase responses, and NK-kB signaling. The top functional groups of non-inflammatory DEGs were eIF2 signaling, axonal guidance signaling, mTOR signaling, neurological disease and mitochondrial system. These analyses demonstrate that the neurodegenerative phenotypes in 4L;C* mice were associated with regional brain transcriptional changes in mRNAs and microRNAs. These abnormalities were broadly related to neuronal functions, such as neuronal differentiation, synaptic plasticity, mitochondria function and inflammation. This study provides new insights into the pathological mechanisms of nGD and the molecular basis for development of novel therapeutic targets.

## 149

**A mouse model of cblA class isolated methylmalonic acidemia (MMA) displays reduced survival, growth failure, renal disease and secondary mitochondrial dysfunction.** *M.W. Epping[1], C.X. Wang[1], P.M. Zerfas[2], G. Elliot[1], L. Li[3], I. Manoli[1], C.P. Venditti[1].* 1) Genetics and Molecular Biology Branch, NHGRI, NIH, Bethesda, MD; 2) Office of Research Services, Division of Veterinary Resources, NIH, Bethesda, MD; 3) Division of Nephrology and Hypertension, Georgetown University Medical Center, Washington, DC.

The Methylmalonic Aciduria cbl¬A Type (MMAA) gene product is responsible for the gated transfer of adenosylcobalamin to methylmalonyl-CoA mutase (MUT) and the protection of MUT from oxidative inactivation. Mutations in the MMAA gene cause the cblA complementation class of isolated MMA that typically has a milder clinical course than MUT deficiency, although is still associated with chronic renal failure, basal ganglia stroke, and optic nerve atrophy (ONA). Knock out murine models of Mut display immediate neonatal lethality and, while valuable for the testing of gene therapy, a need to model an ameliorated form of MMA to discern disease pathophysiology exists. Mmaa-/- mice were generated on a C57B6/Sv129 background and harbor a reporter-tagged deletion in Mmaa that removes exons 3-5 and replaces them with a LacZ-IRES-neo cassette. Mmaa-/- mice were born in Mendelian proportions. The homozygous mutants exhibited decreased survival, with 10% reaching two months of age, and were significantly growth retarded, even when maintained on high fat and carbohydrate diet ($P<0.05$). Immunoreactive MMAA was not detected in liver extracts from Mmaa-/- mice indicating the knock out allele is a null. The concentrations of plasma methylmalonic acid were significantly higher in Mmaa-/- mice, ranging between 63.88-1641.91 μM, compared to the heterozygous range of 5.36-13.56 μM ($P < 0.0001$). Stable isotope studies indicated that the mutant mice exhibit a diminished capacity to oxidize 1-C-13 propionate ($P < 0.0001$). Additionally, the glomerular filtration rate, measured with FITC-sinistrin, showed that Mmaa-/- mice have kidney function diminished to 45% of their heterozygous littermates. Consistent with the organ-specific mitochondrial dysfunction of MMA, electron microscopy confirmed the presence of megamitochondria, with aberrant cristae, and abnormal structure in the hepatocytes, proximal renal tubular epithelial cells, and brown fat. Regular subcutaneous injection of vitamin B12 (OHcbl) improved survival, weight gain, and propionate oxidation in Mmaa-/- males, which maintain shorter survival periods than females. Mmaa-/- mice represent a promising model for examining the pathophysiology and disease manifestations of vitamin-B12-responsive MMA, including hepatorenal disease, metabolic stroke, and ONA, and will allow the development and testing of gene and small molecule therapies for the disease.

## 150

**A genotype likelihood based phasing and imputation method for massive sample sizes of low-coverage sequencing data.** *W. Kretzschmar[1], J. Marchini[1,2], The Haplotype Reference Consortium.* 1) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 2) Department of Statistics, University of Oxford, Oxford, United Kingdom.

Calling genotypes from low-coverage sequencing data is a computationally challenging task. Applying existing methods to cohorts of a few thousand samples typically takes many weeks on large-scale compute clusters. Such methods will not scale to calling genotypes for the first release of the Haplotype Reference Consortium (HRC) (http://www.haplotype-reference-consortium.org/), which will consist of ~31,500 samples. We have developed methods that substantially cut the running time of calling genotypes in the HRC. Our method is an adaptive MCMC algorithm for genotype calling and haplotype imputation, derived from SNPTools (Wang et al. Genome Res 2013), that learns local haplotype clustering as the MCMC chain progresses, and acts to guide the proposal distribution for haplotype sharing between samples. This adaptive scheme is very flexible and can naturally accommodate any existing haplotypes estimates. Our method also supports phasing from reference panels and the inclusion of knowledge about family structure. In addition, we have implemented this approach on a GPU resulting in a dramatic increase in speed. To illustrate the improvements in speed we applied several methods to a single chunk of 1,024 sites from the HRC pilot project consisting of genotype likelihoods on 12,753 samples. Beagle (v3.3.2) took 1016m (averaged over 12 regions). Our new method took 65m, and the GPU implementation of our method took 3m at comparable accuracy. These new methods provide a computational solution for calling genotypes in next-generation sequencing studies of tens to hundreds of thousands of samples. We plan to provide a public software implementation of our GPU method that can be run via cloud computing.

## 151

**Imputation server: next generation genotype imputation service.** *C. Fuchsberger[1], L. Forer[2], S. Schönherr[2], F. Kronenberg[2], G. Abecasis[1].* 1) Ctr Statistical Genetics, Univ Michigan, Ann Arbor, MI; 2) Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria.

Genotype imputation is a key step in the analysis of genome-wide association studies (GWAS). However, imputation into large GWA studies requires expertise and substantial computational resources. Moreover, although upcoming mega reference panels, such as the 30k panel from the Haplotype Consortium, will improve imputation of rare and less common variants, but cannot always be shared broadly, due to consent and privacy restrictions on the original samples. To keep imputation broadly accessible, we designed a web service (called Imputation server) that imputes GWAS data following the MapReduce paradigm without directly sharing the reference panel. To ensure a high level of data sensitivity we have implemented several strategies. All interactions with the server are encrypted. After the imputation process is completed, results are encrypted with a one-time password and are only kept for a few days on our server. Input data is deleted, as soon it is no longer needed. We protect non-public available reference panels by preventing direct access and pseudo imputations geared to unveil the identity of reference panel members. To make this service highly scalable, we have re-engineered the core algorithms in our imputation engine, resulting in a speed-up of ~20x compared to our previous implementation. Our imputation server accepts phased and unphased GWAS genotypes and performs several quality checks, such as strand orientation, variant coding, file integrity, minor allele frequency, and sample and variant missingness. Our service can handle imputation of ~1,500 unphased genomes per day and can easily be scaled up. The service can be accessed for free at https://imputationserver.sph.umich.edu.

## 152

**Improved haplotype phasing using identity by descent.** *B.L. Browning[1,2,3], S.R. Browning[2].* 1) Medicine, Division of Medical Genetics, University of Washington, Seattle, WA; 2) Biostatistics, University of Washington, Seattle, WA; 3) Genome Sciences, University of Washington, Seattle, WA.

We present a new haplotype phasing method that achieves higher accuracy than existing methods. The method is based on the Beagle haplotype frequency model, but unlike the original Beagle phasing method, the new method incorporates genetic recombination, genotype error, and segments of identity by descent.

We compared the new haplotype phasing method to Beagle (r1230) and to SHAPEIT version 2 (r778) using Illumina Human 1M SNP data for chromosome 20. We phased 44 HapMap3 CEU trio offspring together with subsets of Wellcome Trust Case Control Consortium 2 controls (n=650, 1300, 2600, 5200). Phase error was measured at trio offspring genotypes on chromosome 20 that have phase determined by parental genotypes. The SHAPEIT "states" parameter was set at 6400 in order to increase its phasing accuracy.

The new haplotype phasing method produced haplotype switch error rates that were 20-25% lower than the error rates for the existing Beagle method and 1-7% lower than the error rates for SHAPEIT. The difference in switch error rates between the new method and SHAPEIT increased with increasing sample size.

The new haplotype phasing method will be incorporated into version 4 of the Beagle software package (http://faculty.washington.edu/browning/beagle/beagle.html).

## 153

**Reducing pervasive false positive identical-by-descent segments detected by large-scale pedigree analysis.** *E.Y. Durand, N. Eriksson, C.Y. McLean.* 23andMe, Inc., Mountain View, CA.

Analysis of genomic segments shared identical-by-descent (IBD) between individuals is fundamental to many genetic applications, from demographic inference to estimating the heritability of diseases. A large number of methods to detect IBD segments have been developed recently. However, IBD detection accuracy in non-simulated data is largely unknown. In principle, it can be evaluated using known pedigrees, as IBD segments are by definition inherited without recombination down a family tree. We extracted 25,432 genotyped European individuals containing 2,952 father-mother-child trios from the 23andMe, Inc. dataset. We then used GERMLINE, a widely used IBD detection method, to detect IBD segments within this cohort. Exploiting known familial relationships, we identified a false positive rate over 67% for 2-4 centiMorgan (cM) segments, in sharp contrast with accuracies reported in simulated data at these sizes. We show that nearly all false positives arise due to allowing switch errors between haplotypes when detecting IBD, a necessity for retrieving long (> 6 cM) segments in the presence of imperfect phasing. We introduce HaploScore, a novel, computationally efficient metric that enables detection and filtering of false positive IBD segments on population-scale datasets. HaploScore scores IBD segments proportional to the number of switch errors they contain. Thus, it enables filtering of spurious segments reported due to GERMLINE being overly permissive to imperfect phasing. We replicate the false IBD findings and demonstrate the generalizability of HaploScore to alternative genotyping arrays using an independent cohort of 555 European individuals from the 1000 Genomes project. HaploScore can be readily adapted to improve the accuracy of segments reported by any IBD detection method, provided that estimates of the genotyping error rate and switch error rate are available.

## 154

**Parente2: A Fast and Accurate Method for Detecting Identity by Descent.** *S. Bercovici, J. M. Rodriguez, L. Huang, S. Batzoglou.* Computer Science, Stanford, Stanford, CA., Select a Country.

Identity-by-descent (IBD) inference is the problem of establishing a direct and explicit genetic connection between two individuals through a genomic segment that is inherited by both individuals from a recent common ancestor. IBD inference is key to a variety of population genomic studies, ranging from demographic studies to linking genomic variation with phenotype and disease. The problem of both accurate and efficient IBD detection has become increasingly challenging with the availability of large collections of human genotypes and genomes: given a cohort's size, as quadratic number of pairwise genome comparisons must be performed, in principle. Therefore, computation time and the false discovery rate can also scale quadratically. To enable practical large-scale IBD detection, we developed Parente2, a novel method for detecting IBD segments. Parente2 is based on an embedded log-likelihood ratio and uses an ensemble windowing approach to model complex linkage disequilibrium in the underlying studied population. Parente2 is applied directly on genotype data without the need to phase data prior to IBD inference. Through extensive simulations using real data, we evaluate Parente2's performance. We show that Parente2 is superior to previous state-of-the-art methods, detecting pairs of related individuals sharing a 4 cM IBD segment with 99.9%; sensitivity at a 0.1%; false positive rate, and achieving 79.2%; sensitivity at a 1%; false positive rate for the more challenging case of pairs sharing a 2 cM IBD segment. Additionally, Parente2 is efficient, providing one to two orders of magnitude speedup compared to previous state of the art methods. Parente2 is freely available at http://parente.stanford.edu/.

## 155

**Underdog: A Fully-Supervised Phasing Algorithm that Learns from Hundreds of Thousands of Samples and Phases in Minutes.** *K. Noto, Y. Wang, M. Barber, J. Granka, J. Byrnes, R. Curtis, N. Myres, C. Ball, K. Chahine.* AncestryDNA, San Francisco, CA.

Algorithms that phase, i.e., that separate diploid genotypes into a pair of haplotype chromosomes, traditionally do so by phasing many samples together, comparing the genotypes and potential haplotypes to others in the input, and iteratively improving the phase. The larger the input set, the more accurate the phase. However, when the input contains hundreds of thousands of samples, these algorithms become intractable, forcing users to discard potentially useful data. Furthermore, the entire process must be repeated to phase new samples. We suggest that if a training set is large enough, it can be used to build haplotype models that can phase new samples quickly and accurately without requiring that the new samples be used to determine the models. We present a new approach called Underdog, which learns haplotype models from hundreds of thousands of haplotype samples and saves those models for later reuse, enabling the user to rapidly phase new samples. Our results on two experimental data sets show that Underdog phases new samples with 20%-60% fewer errors than current state-of-the-art approaches, and because Underdog takes advantage of parallelization, it can do so in minutes instead of hours (a 100-fold reduction in running time is typical).

## 156

**Fast PCA of very large samples in linear time.** *K.J. Galinsky[1], P. Loh[2,3], G. Bhatia[2], S. Georgiev[4], S. Mukherjee[5], N.J. Patterson[3], A.L. Price[1,2,3].* 1) Department of Biostatistics, Harvard School of Public Health, Boston, MA; 2) Department of Epidemiology, Harvard School of Public Health, Boston, MA; 3) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 4) Stanford University School of Medicine, Palo Alto, CA; 5) Department of Statistical Science, Duke Institute for Genome Sciences & Policy, Durham, NC.

Principal components analysis (PCA) is an effective tool for inferring population structure and correcting for population stratification in genetic data. Traditionally, PCA runs in $O(MN^2N^3)$ time, where $M$ is the number of variants and $N$ is the number of samples. Here, we describe a new algorithm, fastpca, for approximating the top $K$ PCs that runs in time $O(MNK)$, making use of recent advances in random low-rank matrix approximation algorithms (Rokhlin et al. 2009). fastpca avoids computing the GRM and associated computational and memory storage costs, enabling PCA of very large datasets on standard hardware. We estimated the top 10 PCs of the WTCCC dataset (16k samples, 101k variants) in roughly 7 minutes while consuming 1GB of RAM, compared to 1 hour and 2.5GB for PLINK2. The fastpca approximation was extremely accurate ($r^2$>99% between all fastpca and PLINK2 PCs). The improvement in running time becomes even larger at larger samples sizes; for example, fastpca estimated the top 10 PCs of a simulated data set with 100k samples and 300k variants in 135 minutes 8.5GB of RAM, vs. an estimated 350 hours and 85GB of RAM using PLINK2. A recently published $O(MN^2)$ time method, flashpca, did not complete on this data set due to exceeding 40GB memory requirement. All of these analyses were based on LD-pruning SNPs with $r^2$>0.2, which leads to much more accurate PCs in simulations as compared to retaining all SNPs; more complex LD-adjustment strategies provide only a small further improvement.

## 157

**Fast Detection of IBD Segments Associated With Quantitative Traits in Genome-wide Association Studies.** *Z. Wang[1], E. Kang[1], B. Han[2,3], S. Snir[4], E. Eskin[1].* 1) Computer Science Department, University of California, Los Angeles, CA 90095; 2) Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; 3) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 4) Institute of Evolution, Department of Evolutionary and Environmental Biology, Faculty of Natural Sciences, University of Haifa, Israel.

Recently, many methods have been developed to detect the identity-by-descent (IBD) segments between a pair of individuals. These methods are able to detect very small shared IBD segments between a pair of individuals up to 2 centimorgans in length. This IBD information can be used to identify recent rare mutations associated with phenotype of interest. Previous approaches for IBD association were applicable to case/control phenotypes. In this work, we propose a novel and natural statistic for the IBD association testing, which can be applied to quantitative traits. A drawback of the statistic is that it requires a large number of permutations to assess the significance of the association, which can be a great computational challenge. We make a connection between the proposed statistic and linear models so that it does not require permutations to assess the significance of an association. In addition, our method can control population structure by utilizing linear mixed models.

## 158

**Large-scale profiling of sequence variation affecting transcription factor occupancy in vivo.** *M.T Maurano, E. Haugen, R. Sandstrom, J. Vierstra, J.A. Stamatoyannopoulos.* Dept. of Genome Sciences, Univ. of Washington, Seattle, WA, USA.

Genome-wide association studies have identified thousands of disease- and trait-associated variants that systematically localize in non-coding regulatory elements. The mechanistic investigation of sequence variation in these elements has been impeded by the difficulty of experimentally modelling the highly cell-type selective regulatory activity and complex physical organization of native loci. To identify functional variation affecting regulatory elements in their native locus configuration and cellular environment, we analyzed allelically resolved genomic DNaseI footprinting and ChIP-seq data to identify variants with consistent effect across 121 cell and tissue types. We report the functional classification of 359,477 regulatory variants, of which 66,957 demonstrate significant imbalance in chromatin accessibility in vivo. Discovery of functional variation depends strongly on sequencing depth, which can be efficiently augmented using a targeted approach. In contrast to the characteristic cell-type selectivity of the chromatin landscape, we find that sequence variants affect occupancy across multiple cellular contexts. We show that functional variation delineates characteristic sensitivity profiles for several hundred transcription factor motifs representing 56 families of non-redundant sequence specificities, and including many of the key factors linked to the establishment of accessible chromatin. Nevertheless, silent variants are found repeatedly at every position within the protein-DNA recognition interface, and the majority of variation is buffered in a site-dependent manner in vivo. We account for these local context effects by developing TF-specific profiles of functional variation, and demonstrate their utility for the functional classification of novel regulatory variation by identifying 438,097 variants in dbSNP strongly predicted to affect binding. In summary, our characterization of regulatory variation affecting TF activity provides a foundation for the etiological investigation of non-coding genetic associations.

## 159

**Consider the geneset: Why the transcripts used for variant annotation matter.** *A. Frankish[1], JM. Mudge[1], R. Petryszak[2], GRS. Ritchie[3,4], A. Brazma[2], JL. Harrow[1], GENCODE Consortium.* 1) Human and Vertebrate Analysis and Annotation Group, Computation Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; 2) Functional Genomics Team, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK; 3) Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; 4) Vertebrate Genomics Team, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

McCarthy et al.[1] recently demonstrated the large differences in prediction of loss-of-function variation when RefSeq and Ensembl transcripts are used for annotation. Ensembl displays the GENCODE geneset, the reference human gene annotation for the ENCODE project. Although the GENCODE and RefSeq genesets contain similar numbers of protein-coding genes, there are significant differences between them, e.g. in the annotation of alternative splicing where GENCODE protein-coding loci have a mean of 7.6 alternatively-spliced transcripts while RefSeq only have 2.1. Similarly, the GENCODE geneset is enriched compared to RefSeq for the annotation of long non-coding RNAs and pseudogenes, genomic coverage of annotated exons, extent of manual curation, experimental validation, and functionally descriptive biotypes. By representing more transcriptional complexity, the GENCODE geneset allows the annotation of a greater number of potentially interesting variants; the more detailed functional annotation of transcripts also assists with consequence calling. We will discuss GENCODE's extension and refinement of the geneset with the integration of RNAseq, CAGE, polyAseq, ribosome profiling, mass spectrometry and epigenomic data, to identify novel loci, define 5' and 3' transcript boundaries, identify novel translation initiation sites and improve functional annotation e.g. by confirming the translation of putative protein-coding transcripts. While our deep representation of the transcriptome is beneficial for some aspects of variant annotation, it may prove a hinderance to others, e.g. where a variant is predicted to have conflicting effects on different transcripts from the same gene. To address this, we will describe the filtering options provided to allow the user to reduce complexity of the GENCODE gene set and explain our use of RNAseq data to investigate the abundance of GENCODE-annotated genes, transcripts and exons, to present a smaller, but biologically relevant, set of features e.g. by presenting the reduced set of genes expressed in a tissue of interest. In summary, the set of transcripts selected as a basis for annotating variants affects both the number of variants identified as genic and their predicted functional consequences. The GENCODE geneset captures transcriptional complexity and describes its functional potential while permitting filtering of features to facilitate accurate interpretation of variation.

[1]Genome Medicine 2014, 6:26.

## 160

**Multi-sample isoform quantification from RNA-seq.** *A.E. Byrnes[1,2], J.B. Maller[1,2], A.R. Sanders[3,4], J. Nemesh[2], T. Sullivan[2], H.H. Göring[5], J. Duan[3,4], W. Moy[3], E.I. Drigalenko[5], P.V. Gejman[3,4], B.M. Neale[1,2].* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Broad Institute of MIT and Harvard, Cambridge, MA; 3) Department of Psychiatry and Behavioral Sciences, NorthShore University Health System, Evanston, IL; 4) Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL; 5) Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX.

Alternative splicing is critical for the regulation and diversity of the majority of human genes. (Wang et. al., 2008) The ability of a single gene to give rise to several diverse transcripts, and subsequently proteins, has been implicated in a wide range of processes and disorders from brain function to cancer proliferation. (David et. al., 2010; Blencowe, 2005) The recent advancements and price reduction in RNA-seq technologies presents an unprecedented opportunity to investigate splicing on a transcriptome-wide scale across many samples. However, the relatively short read-lengths do not provide perfect information as to which transcript isoforms are present. Here we present a systematic comparison between several existing methods for transcript assembly and quantification from RNA-seq data, including Cufflinks (Roberts, et. al., 2011), RSEM (Li and Dewy, 2011) and PSGInfer (LeGault and Dewy, 2013). We also propose a 2-step, multi-sample method for discovery and quantification of transcript isoforms (both known and novel) from paired-end RNA-seq data, while making use of a reference genome and any available annotation. Our method aims, first, to maximize information about splicing behavior by combining information from all aligned RNA-seq samples in order to construct a graph representing all possible transcripts, similar to approaches taken by PSGInfer and Cufflinks, but on all samples pooled together. In graph-building we weight each of the possible junctions between exons by the number of junction reads observed across all samples. We represent each isoform as a possible path through the graph and the use the weight of each edge as the initial probability in the following step. After constructing all likely isoforms from the data, we use the expectation-maximization algorithm to estimate their relative abundance, in addition to any known isoforms, similar to the methods applied in RSEM and eXpress. This second step allows us to specifically characterize the isoforms present in any individual and quantify their respective transcription for each sample separately. We will discuss the details of this method as well as the relative performance of all the above methods in real and simulated data. Our results have clear implications for the analysis of future work in alternative splicing.

## 161

**Characterizing the genetic architecture of gene expression variation in wild baboons via RNA sequencing.** *X. Zhou[1,2], J. Tung[3], S. Alberts[4], J. Altmann[5], M. Stephens[1,2], Y. Gilad[1].* 1) Department of Human Genetics, University of Chicago, Chicago, IL; 2) Department of Statistics, University of Chicago, Chicago, IL; 3) Department of Evolutionary Anthropology, Duke University, Durham, NC; 4) Department of Biology, Duke University, Durham, NC; 5) Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ.

Gene expression variation is well documented in human populations and its genetic architecture has been extensively explored in recent large-scale studies. However, we still know little about the genetic architecture of gene expression variation in other species, particularly our closest living relatives, the nonhuman primates. To address this gap, we performed an RNA sequencing (RNA-seq)-based study in 63 wild baboons, members of the intensively studied Amboseli baboon population in Kenya. Our study design allowed us to measure gene expression levels and call genetic variants using the same data set, enabling us to subsequently map cis-eQTLs (expression quantitative loci). We validated our approach using an existing human HapMap RNA-seq data set. We detected more than one thousand variants affecting gene expression levels in baboons, which is approximately four times more eQTLs than detectable using the same approach on a HapMap human data set of comparable size. This increase in power appears to stem from a combination of increased genetic variation, enrichment of SNPs with high minor allele frequencies, and longer-range linkage disequilibrium in the baboon data set relative to the human data set. As observed in humans, baboon eQTLs are enriched inside genes and near transcription start sites and associated with allelic specific expression in heterozygotes. eQTL effect sizes in the baboons were negatively correlated with minor allele frequency, consistent with arguments that negative selection often acts on gene expression variation to reduce the impact of the regulatory variation. Further, genes with large effect eQTL in baboons overlapped significantly with genes with large effect eQTL in humans. This set of overlapping genes was significantly less conserved across vertebrates at the sequence level than genes with large effect eQTL in only one species, which were in turn less conserved than genes with no detectable eQTL in either species. Finally, using a Bayesian sparse linear mixed model, we estimate that the cis-regulatory variants in baboons together explain approximately half of the genetic variance for gene expression levels, which is comparable to the results obtained in the same tissue in a human population. Together, our comparative eQTL mapping study represents an important first step towards understanding the genetic architecture of gene expression variation in natural primate populations.

## 162

**Genetic control of chromatin in a Human population.** *O. Delaneau[1], S. Waszak[2], A. Gschwind[3], H. Kilpinen[1], S. Raghav[2], R. Witwicki[3], A. Orioli[3], M. Wiederkehr[3], M. Gutierrez-Arcelus[1], N. Panousis[1], A. Yurovsky[1], T. Lappalainen[1], L. Romano-Palumbo[1], A. Planchon[1], D. Bielser[1], I. Padioleau[1], G. Udin[2], S. Thurnheer[4], D. Hacker[4], N. Hernandez[3], A. Reymond[3], B. Deplancke[2], E. Dermitzakis[1].* 1) Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland; 2) Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland; 3) Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, Geneva, Switzerland; 4) Protein Expression Core Facility, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.

Non-coding regulatory DNA variants have been found to be associated with gene expression, yet the precise molecular basis by which they act remains elusive. We hypothesize that an integrated study of chromatin states coupled with personal genome information might enable in-depth characterization of these regulatory variants. We quantified gene expression (mRNA), genome-wide DNA binding of two regulatory proteins (polymerase II, PU.1) and three histone post-translational modifications marks, that pinpoint promoter, enhancer and active regions (H3K4me3, H3K4me1 and H3K27ac, respectively) in lymphoblastoid cell lines of 50 unrelated individuals that were whole genome-sequenced as part of the 1000 genomes project. This data allowed us mapping cis-acting QTLs for each molecular phenotype and assessing the degree of coordinated behaviors between distinct layers of gene regulation. We find that the various molecular phenotypes show abundant coordination in activity levels at enhancer and/or promoter elements, forming 'chromatin modules' that can extent over hundreds of kb and regroup hundreds of chromatin sites. We show that overall chromatin activity at these modules is tightly correlated with changes in expression levels at genes nearby, highlighting their central role in transcription. We also mapped thousands of cis-acting QTLs for gene expression (eQTLs), histone modification levels (hmQTLs) and TF binding (tfQTLs) at 10% FDR. We find that they are widespread across the genome and that they explain a substantial fraction of inter-individual variability in chromatin activity. In addition, we find large overlaps between the various QTLs (50% of eQTLs are also hmQTLs) which reflects the genetic signal propagation through the multiple phenotypic layers and show that the genetic perturbation of chromatin tend to be causal to changes in gene expression levels. Overall, this large-scale study that integrates DNA, RNA, regulatory proteins and histone modifications provides novel insights into the mechanisms underlying regulatory variation and their effects on transcription.

## 163

**Chromatin accessibility profiling of developing cerebellar granule neurons reveals novel neuronal enhancers and regulatory scheme for ZIC transcription factors.** *C.L. Frank[1,2], F. Liu[3], R. Wijayatunge[3], L. Song[2], C.M. Vockley[2,4], A. Safi[2], G.E. Crawford[2,5], A.E. West[2,3].* 1) Department of Molecular Genetics and Microbiology, Duke University, Durham, NC; 2) Institute for Genome Sciences and Policy, Duke University, Durham, NC; 3) Department of Neurobiology, Duke University, Durham, NC; 4) Department of Cell Biology, Duke University, Durham, NC; 5) Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, NC.

Chromatin regulation mediates the cell-type specific expression of genes by establishing differential accessibility to transcription factor binding sites across the genome in cells of distinct fate lineages. However, the dynamics of chromatin regulation over the time-course of cellular differentiation within a single fate lineage remain poorly understood. We set out to characterize the temporal relationship between chromatin regulation and gene expression in the developing mouse cerebellar cortex, which is dominantly comprised of a single type of neuron, cerebellar granule neurons (CGNs). We used DNase-seq to globally map chromatin accessibility of cis-regulatory elements and RNA-seq to profile transcript abundance at three key time-points in postnatal development. We observed widespread chromatin accessibility changes at 24,886 regulatory elements (FDR<.05). Equivalent profiling for three time-points of purified CGNs differentiating in culture further improves temporal resolution of early chromatin events. The majority of these dynamic elements appear to be developmental stage-specific neuronal enhancers, which is supported by these regions being (i) primarily located outside of proximal promoters, (ii) overlapping H3K27ac and H3K4me1 ChIP-seq peaks from adult cerebellum, (iii) enriched nearby developmentally-regulated genes, (iv) enriched for a publicly available set of confirmed hindbrain enhancer regions that function in embryonic mice, and (v) sufficient to drive reporter gene expression in cultured mouse neurons. Motif discovery in the differentially accessible elements revealed several transcription factor families, including zinc finger proteins of the cerebellum (ZIC), that likely bind these elements. We confirmed ZIC involvement by globally mapping ZIC binding sites in early and adult postnatal cerebellum by ChIP-seq. Despite consistent expression of Zic genes, binding patterns are highly dynamic across development. Knockdown of Zic1 and Zic2 resulted in loss of up-regulation of key CGN maturity marker genes, suggesting that ZIC chromatin-gated access to the DNA template is required for driving mature neuronal gene expression patterns, a previously unappreciated role for these factors. Together this study has revealed chromatin dynamics at thousands of novel enhancers that facilitate expression patterns necessary for neuronal differentiation and function.

## 164

**Identification and characterization of enhancer and target gene pairs in mammalian genomes.** *Y.-C. Hwang[1], C.-F. Lin[2,3], O. Valladares[2,3], J. Malamon[2,3], Q. Zheng[4,5], B. Gregory[1,4,5], L.-S. Wang[1,2,3,5].* 1) Genomics and Computational Biology Graduate Group, University of Pennsylvania Perelman School of Medicine; 2) Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine; 3) Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine; 4) Department of Biology, University of Pennsylvania, Philadelphia, PA; 5) Penn Genome Frontiers Institute, University of Pennsylvania, Philadelphia, PA.

Genome-wide association studies have shown the majority of disease- and trait-associated genetic variations lie within non-coding regions of the human genome. It has been hypothesized that many of these variants may affect non-coding regulatory elements. One class of the elements is enhancer elements, which regulate gene expression through long-range interactions with the promoters of protein-coding loci. However, the interactions between enhancer elements and their target genes can be linearly distal and orientation-independent, and probing all possible enhancer-target gene pairs in the genome is laborious and remains largely unsolved. To identify all enhancers and the genes they regulate, we reanalyzed Hi-C datasets of human cells (hESC, IMR90, GM06996, K562) and mouse cells (mESC and cortex). We first extracted restriction fragments (intervals between two adjacent restriction sites) with significantly higher Hi-C read counts than expected. These restriction fragments are referred to as Hi-C peaks. The Hi-C peaks are identified as enhancer elements with the following criteria: they have to (1) pair with a coding gene promoter region; (2) overlap with sites having known enhancer-associated histone modifications (H3K27ac, H3K4me1, etc); and (3) reside in DNase I hypersensitive sites. Using this analysis pipeline, we have identified between 2,540 and 13,867 enhancer-target gene pairs for human and mouse genomes. As expected, enhancers are more conserved and highly enriched with p300 binding activity, while their target promoters are ~20% more likely to be in RNA polymerase II binding sites and cell-type-specific. We found enhancers can act pleiotropically by regulating more than one gene while there is also redundancy in enhancer-target gene pairs to provide precise gene regulation. We also found that ~90% of the pairs are intra-chromosomal and the majority of the interactions are within 1Mbp of each other. By down sampling Hi-C reads, we found increased read coverage allows improved detection of longer distance interactions. These results suggest that long-range interactions are relatively transient in the cell. This comprehensive enhancer-target gene catalog will allow us to identify disease-linked polymorphisms that lie within enhancers, as well as their regulated genes as candidate disease genes. By comparing the enhancer-gene pairs between human and mouse embryonic stem cells, we can study the evolution of enhancer-mediated regulatory mechanisms.

## 165

**Domains of genome-wide gene expression dysregulation in Down syndrome.** *S.E. Antonarakis[1], A. Letourneau[1], F.A. Santoni[1], X. Bonilla[1], M.R. Sailani[1], D. Robyr[1], D. Gonzalez[2], J. Kind[3], C. Chevalier[4], R. Thurman[5], R.S. Sandstrom[5], Y. Hibaoui[6], M. Garieri[1], K. Popadin[1], E. Falconnet[1], M. Gagnebin[1], M. Gehrig[1], A. Vannier[1], M. Guipponi[1], E. Migliavacca[1], S. Deutsch[1], A. Feki[6], J. Stamatoyannopoulos[5], Y. Herault[4], B. van Steensel[3], R. Guigo[2], C. Borel[1].* 1) Genetic Medicine, University of Geneva, Geneva, Switzerland; 2) Center for Genomic Regulation, University Pompeu Fabra, Barcelona, Spain; 3) Division of Gene Regulation, Netherlands Cancer Institute, Amsterdam, The Netherlands; 4) Institut de Génétique Biologie Moléculaire et Cellulaire, Translational medicine and Neuroscience program, IGBMC, and Université de Strasbourg, France; 5) Department of Genome Sciences, University of Washington, Seattle, Washington, USA; 6) Stem Cell Research Laboratory, Department of Obstetrics and Gynecology, Geneva University Hospitals, Geneva, Switzerland.

Trisomy 21 is the most frequent genetic cause of cognitive impairment. To assess the perturbations of gene expression in trisomy 21, and to eliminate the noise of genomic variability, we studied the transcriptome of fetal fibroblasts from a pair of monozygotic twins discordant for trisomy 21. We have shown that the differential expression between the twins is organized in domains along all chromosomes that are either upregulated or downregulated (Nature 2014; PMID 24740065). These gene expression dysregulation domains (GEDDs) can be defined by the expression level of their gene content, and are well conserved in induced pluripotent stem cells derived from the twins' fibroblasts. Comparison of the transcriptome of the Ts65Dn mouse model of Down's syndrome and normal littermate mouse fibroblasts also showed GEDDs along the mouse chromosomes that were syntenic in human. The GEDDs correlate with the lamina-associated (LADs) and replication domains of mammalian cells. The overall position of LADs was not altered in trisomic cells; however, the H3K4me3 profile of the trisomic fibroblasts was modified and accurately followed the GEDD pattern. These results indicate that the nuclear compartments of trisomic cells undergo modifications of the chromatin environment influencing the overall transcriptome, and GEDDs may therefore contribute to some trisomy 21 phenotypes. GEDDs could be the result of genes on chromosome 21, or to the extra chromosomal material. To distinguish between the two possibilities, we use i/ a series of mouse models of human trisomy 21 with different partial trisomies and monosomies; ii/ targeted disruption of one allele of candidate genes in a trisomy background; iii/ fibroblasts from mosaic trisomies 13 and 18.

## 166

**Investigation of Synthetic Association in GWAS using PheWAS and Exome Sequencing.** *L. Bastarache, J. Bochenek, T. Edwards, Y. Xu, J. Pulley, E. Bowton, H. Mo, W. Wei, L. Wiley, D. Roden, J. Denny.* Vanderbilt University, Nashville, TN.

Genome-wide association studies (GWAS) have identified thousands of common variants associated with human diseases or traits. The biological effects on traits of common GWAS SNP associations are often not apparent by functional annotations. The synthetic association (SA) hypothesis proposes that one or more rare causal alleles on a haplotype containing a common, neutral allele, may explain some GWAS findings. To test the plausibility of this hypothesis, we investigated rare variants located near known GWAS findings. In this study, we evaluated 29,722 European-ancestry individuals genotyped on the Illumina HumanExome array with traits defined by electronic medical records in a phenome-wide association scan (PheWAS). For results reported in the GWAS catalog that reach genome-wide significance ($p<5\times10\text{-}8$), we mapped the disease or trait to a PheWAS phenotype. We then evaluated all available missense or nonsense SNPs with minor allele frequency (MAF) ≤5% in the GWAS-reported genes for that phenotype. We tested 104 unique phenotypes and 1,743 unique phenotype-gene pairs using single SNP tests of association adjusted for age, sex, and principal components. There were 84 associations for rare variants with $p<0.01$ and 59 that had an odds ratio (OR) greater than 2 or less than 0.5 for 38 unique disease phenotypes. There were 22 associations that exceeded a Bonferroni correction for tested phenotype-gene pairs. For example, with primary biliary cirrhosis, we observed associations with rs200568391 CLEC16A (MAF 0.1%, OR=11.19, $p=3.6\times10\text{-}8$) and with IKZF3 rs149299224 (MAF 0.3%, OR=6.24, $p=8.7\times10\text{-}7$); the ORs in the GWAS catalog were 1.29 and 1.44, respectively. Similarly, rare variants in genes associated with schizophrenia (ARL6IP4), Crohn's disease (NOD2, ATG16L2, C7orf72), hypothyroidism (C6orf15), and ulcerative colitis (C6orf223) had rare variants with p<10-4 and OR=1.82-12.53. Associations with Sjögren's syndrome (COL11A2), esophageal cancer (CORO2B), and basal cell carcinoma (PADI4) were statistically significant. Conditional analysis is ongoing to investigate SA in cases where we genotype the GWAS identified SNP. This study demonstrates that rare variation may contribute to the genetic risk of a wide variety of phenotypes, and that large EMR cohorts and the exome array may accelerate investigation of SA and related mechanisms by which rare variants contribute to risk in heritable traits.

## 167

**Beware of circularity: A critical assessment of the state of the art in deleteriousness prediction of missense variants.** *C.A. Azencott[1,2,3,4], D. Grimm[4,5], J.W. Smoller[6,7,8], L. Duncan[10,9,6], K. Borgwardt[4,5].* 1) MINES ParisTech, PSL Research University, Centre for computational biology, 77300 Fontainebleau, France; 2) Institut Curie, 75248 Paris Cedex 05, France; 3) INSERM, U900, 75248 Paris Cedex 05, France; 4) Machine Learning and Computational Biology research group, Max Planck Institute for Intelligent Systems and Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany; 5) Zentrum fuer Bioinformatik, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany; 6) Broad Institute of MIT and Harvard, Cambridge, MA; 7) Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA; 8) Harvard Medical School, Department of Psychiatry, Boston, MA; 9) Harvard Medical School, Department of Medicine, Boston, MA; 10) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA.

Discrimination between disease-causing missense mutations and neutral polymorphisms is a key challenge in current sequencing studies. It is therefore critical to be able to evaluate fairly and without bias the performance of the many *in silico* predictors of deleteriousness. However, current analyses of such tools and their combinations are liable to suffer from the effects of *circularity*, which occurs when predictors are evaluated on data that are not independent from those that were used to build them, and may lead to overly optimistic results. Circularity can first stem from the overlap between training and evaluation datasets, which may result in the well-studied phenomenon of overfitting: a tool that is too tailored to a given dataset will be more likely than others to perform well on that set, but incurs the risk of failing more heavily at classifying novel variants. Second, we find that circularity may result from an investigation bias in the way mutation databases are populated: in most cases, all the variants of the same protein are annotated with the same (neutral or pathogenic) status. Furthermore, proteins containing only deleterious SNVs comprise many more labeled variants than their counterparts containing only neutral SNVs. Ignoring this, we find that assigning a variant the same status as that of its closest variant on the genomic sequence outperforms all state-of-the-art tools. Given these barriers to valid assessment of the performance of deleteriousness prediction tools, we employ approaches that avoid circularity, and hence provide independent evaluation of ten state-of-the-art tools and their combinations. Our detailed analysis provides scientists with critical insights to guide their choice of tool as well as the future development of new methods for deleteriousness prediction. In particular, we demonstrate that the performance of FatHMM-W relies mostly on the knowledge of the labels of neighboring variants, which may hinder its ability to annotate variants in the less explored regions of the genome. We also find that PolyPhen2 performs as well or better than all other tools at discriminating between cases and controls in a novel autism-relevant dataset. Based on our findings about the mutation databases available for training deleteriousness prediction tools, we predict that retraining PolyPhen2 features on the Varibench dataset will yield even better performance, and we show that this is true for the autism-relevant dataset.

## 168

**Application of Clinical Text Data for Phenome-Wide Association Studies (PheWASs).** *S.J. Hebbring, M. Rastegar-Mojarad, Z. Ye, J. Mayer, C. Jacobson, S. Lin.* Marshfield Clinic Research Foundation, Marshfield, WI.

Genome-Wide Association Studies (GWAS) have proven effective in describing the genetic complexities of common diseases. Phenome-Wide Association Studies (PheWASs) using diagnostic codes embedded in electronic medical record (EMR) systems have proven effective as an alternative/complementary approach to GWAS. The PheWAS technique has the capacity to identify novel gene-disease associations and link multiple conditions to a common genetic etiology. The majority of PheWASs published to date have utilized ICD9 diagnostic codes to define cases and controls, but it has been shown that ICD9 codes have limited utility. ICD9 codes are primarily used for billing, can have limited phenotypic granularity, and often do not allow for other clinically relevant information to be used for PheWAS interpretation.

As an alternative to ICD9 coding, a text-based phenome was defined from 1,564,831 clinical notes from 4,204 patients containing 423,537,905 words linked to Marshfield Clinic's EMR system. Clinical text data were cross referenced with the UMLS Medical Dictionary of disease terms and drug names to enrich for 23,384 clinically relevant word strings that defined the text-based phenome. Five SNPs known to be associated with different phenotypes were genotyped on the 4,204 patients and associated across the text-based phenome. All five SNPs had expected word strings associated with SNP genotype ($p<0.02$) with most at or near the top of their respective PheWAS ranking. For example, SNP rs1061170, a SNP in CFH that is known to be associated with age related macular degeneration (AMD), was strongly associated with AMD related word strings including "macular degeneration" (p=1.8E-8), "nonexudative" (p=2.3E-7), "exudative" (p=1.4E-6), and "visudyne," a drug commonly prescribed to treat AMD (p=3.9E-7). When comparing results from the text-based PheWAS and an ICD9-based PheWAS, the text-based PheWAS performed equivalently to the ICD9-based PheWAS with three of the five SNPs having stronger p-values.

In conclusion, this study demonstrates for the first time that raw text data from clinical notes in an EMR system can be used effectively to define a phenome. This study also validates that clinical text data, including drug data, can be applied to a PheWAS as an alternative and complementary approach to a GWAS or ICD9-based PheWAS.

## 169

**The Warped Linear Mixed Model: finding optimal phenotype transformations yields a substantial increase in signal in genetic analyses.** *N. Fusi[1], C. Lippert[1], N. Lawrence[2], O. Stegle[3].* 1) eScience group, Microsoft Research, Los Angeles, USA; 2) Department of Computer Science, University of Sheffield, Sheffield, UK; 3) European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK.

Linear mixed models are a core statistical approach used in several key areas of genetics. In particular, they provide state-of-the-art solutions for genome-wide association studies, heritability estimation and phenotype prediction. However, one of the fundamental assumptions of these models—that the noise is Gaussian distributed—rarely holds in practice. We show that as a result, standard approaches yield sub-optimal performance, resulting in significant losses in power for GWAS, increased bias in heritability estimation, and reduced accuracy for phenotype predictions. One way to mitigate this problem is to apply an appropriate transformation (e.g., log transform) as a preprocessing step of the phenotypic data. However, choosing the "right" transformation is challenging because of the need to manually define a set of transformations, and choose one over another, without a clear objective function that could be used to guide this decision. Thus, the problem has only been partially, and unsatisfactory solved. Here, we comprehensively address this important problem in genetics by introducing a robust and statistically principled method, the "Warped Linear Mixed Model". Our approach automatically learns a suitable phenotype transformation from the observed data (both phenotypic and genotypic). This data-driven approach enables an infinite set of transformations to be automatically searched through, using the principles of statistical inference to determine which transformation is most suitable. In extensive synthetic and real experiments, we find up to twofold increases in GWAS power, reduced bias in heritability estimation of up to 30%, and significantly increased accuracy in phenotype prediction. Importantly, our warped linear mixed model is general and can be used in place of standard linear mixed models in a wide range of applications in genetics.

## 170

**PhenomeCentral: An Integrated Portal for Sharing Patient Phenotype and Genotype Data for Rare Genetic Disorders.** *M. Brudno[1], M. Girdea[1], O.J. Buske[1], S. Dumitriu[1], H. Trang[1], T. Hartley[2], D. Smedley[3], S. Kohler[4], P.N. Robinson[4], T.E. Dudding[5], H. Lochmuller[6], C.F. Boerkoel[7], W.A. Gahl[7], K. Boycott[2], Canadian CARE for RARE, NIH Undiagnosed Diseases Program, RD-Connect, Care for Rare Australia.* 1) Centre for Computational Medicne, Hospital for Sick Children & University of Toronto, Toronto, ON, Canada; 2) Children's Hospital of Eastern Ontario, Ottawa, ON Canada; 3) Sanger Institute, Hinxton, UK; 4) Institute for Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin, Germany; 5) Hunter Genetics and University of Newcastle, Newcastle, NSW, Australia; 6) Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK; 7) National Institutes of Health Undiagnosed Diseases Program, NIH, Bethesda USA.

The availability of low-cost genome sequencing has allowed for the identification of the molecular cause of hundreds of rare genetic disorders. Solved disorders, however, represent only the "tip of the iceberg". Because the discovery of disease-causing variants typically requires confirmation of the mutation or gene in multiple unrelated individuals, a larger number of genetic disorders remain unsolved due to difficulty identifying second families. As many groups are now tackling these remaining undiagnosed disorders, which may be present in only a handful of individuals seen at different hospitals and sequenced by different centers, there is a pressing need for tools enabling global collaboration.

To help clinicians and rare disorder scientists identify additional families with a specific rare disease through sharing of genetic and phenotypic data, we have developed PhenomeCentral (http://phenomecentral.org). Each patient record within PhenomeCentral consists of a thorough phenotypic description capturing observed abnormalities and relevant absent manifestations expressed using Human Phenotype Ontology terms, as well as the observed genetic variants, specified either as a whole exome/genome or a list of candidate genes. New patients can be added either using the PhenoTips User Interface, built into PhenomeCentral, or uploaded in bulk. For a given patient record the contributor is shown information about other phenotypically similar patients, together with potential genetic causes, prioritized by the Exomiser algorithm. The phenotypic and genetic features are presented without revealing additional patient information or identifying the contributors, while enabling direct communication for any subsequent data sharing. PhenomeCentral currently incorporates phenotype data for >450 patients with rare genetic disorders lacking molecular diagnosis, including 260 with available whole exome/genome data. PhenomeCentral has been used to identify the likely causative mutation in the *EFTUD2* gene in two patients with atypical presentations, as well as to generate leads by identifying novel phenotypic clusters with mutations in several genes. Participants in PhenomeCentral include the Canadian CARE for RARE project, the NIH Undiagnosed Diseases Program, RD-Connect, and Care for Rare Australia. Membership is open to all clinicians and rare disorder scientists affiliated with academic or not-for-profit research centers.

## 171

**Facilitating the interpretation of rare pathogenic variation in a clinical setting with DECIPHER.** *G.J. Swaminathan[1], E. Bragin[1], E.A. Chatzimichali[1], S. Brent[1], A.P. Bevan[1], H.V. Firth[1,2], M.E. Hurles[1].* 1) Wellcome Trust Sanger Institute, Hinxton, Cambridge, Cambridgeshire, United Kingdom; 2) Cambridge University Department of Medical Genetics, Addenbrooke's Hospital, Cambridge CB2 2QQ, United Kingdom.

DECIPHER (Est. 2004) (https://decipher.sanger.ac.uk) is a web-based tool and database that aids the discovery and interpretation of pathogenic genetic variation (sequence and/or copy-number) in rare disorders obtained in a clinical setting or as part of an approved research study. Over 250 academic departments in genetic medicine contribute phenotype-linked variation data to the database for analysis and interpretation. Following informed consent, shared anonymised patient data enables the identification of clusters of patients with similar phenotype-linked genomic findings and encourages collaboration and contact between members. DECIPHER also facilitates contact between external users and consortium members making it an invaluable collaborative resource for genomic research and clinical diagnosis. Since 2010 alone, over 600 peer-reviewed research publications have benefited from such collaborations. The database contains more than 30000 patient records of which over 14000 have been consented for anonymous sharing with the wider community.

DECIPHER accepts the deposition of both sequence (point mutations, inDels) and copy-number variants with associated phenotypes and other relevant information (pathogenicity, inheritance, genotype etc.). Data are analysed in real-time and displayed in the context of affected gene/s scored by different prioritization criteria (OMIM, haploinsufficiency etc.) along with positional overlaps with known diseases/syndromes and patients with shared phenotypes. A purpose-built highly customizable genome browser (Genoverse: http://genoverse.org) provides visualization of the deposited variant/s against data in DECIPHER and other data sources including ClinVar, LSDB, ISCA etc. Different levels of data sharing are provided for, including sharing of selected data within a group (individuals, specific communities, projects) as well as fully anonymised public access. DECIPHER also provides programmatic access to public anonymous patient data as part of the Genomic Matchmaker Exchange project of the Global Alliance for Genomic Health (GA4GH) initiative.

We will present features of the DECIPHER database and illustrate the critical role it plays as a clinical and research utility for the interpretation of rare genetic disorders.

## 172

**GeneMatcher: A Matching Tool for Identification of Individuals with Mutations in the Same Gene.** *A. Hamosh[1], N. Sobreira[1], F. Schiettecatte[2], D. Valle[1].* 1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) FS Consulting, Salem, MA.

In the last few years, whole exome sequencing (WES) has been the main method used to search for Mendelian disease genes. Identifying the pathogenic mutation from among thousands to millions of genomic variants in typical WES is a challenge. In more than half of the individuals who have a clinical WES the responsible gene and variants cannot be determined. Reasons for this relatively low yield include phenotypic variation; the current, small number of known disease genes (~3300 or only 15% of the total protein coding genes); uncertainty regarding functional consequences of identified variants; and, limited connections between clinicians with clinical WES data and basic scientists. Structured, comprehensive phenotypic data and its sharing as well as improvements in searching for patients or model organisms with mutations in specific candidate genes are important for the success of clinical and research WES analysis. Here we describe GeneMatcher (www.genematcher.org), a freely accessible web-based tool designed to enable connections between clinicians and researchers from around the world who share an interest in the same gene or genes. No identifiable data are collected. The site allows investigators to post a gene(s) (by gene symbol, base pair position, Entrez- or Ensembl-Gene ID) of interest and will connect investigators who post the same gene. The match is done automatically, submitters will automatically receive email notification and follow-up is at the discretion of the submitters. There is no way to search the database. Submitters have access to their own data and may edit it or delete it at will. There is also an option to provide diagnosis based upon OMIM® number and match on that, but this is not required. If a match is not identified at the time of submission the genes of interest will continue to be queried by new entries. An API has been developed that allows the submitter to expand queries within GeneMatcher to other available matching sites, upon request. Presently, the connection is between GeneMatcher and Phenome Central, but other databases will soon be connected to GeneMatcher. In the future, we also plan to enable matching based upon phenotypic features, with or without candidate genes to enhance interpretation of clinical and research exome sequence data. As of 4 June 2014, GeneMatcher has 593 genes submitted by 101 submitters from 18 countries and 14 matches have occurred enabling collaborations between clinicians and researchers.

## 173

**Findings from the Critical Assessment of Genome Interpretation, a community experiment to evaluate phenotype prediction.** *S.E. Brenner[1], J. Moult[2], CAGI Participants.* 1) University of California, Berkeley, CA; 2) IBBR, University of Maryland, Rockville, MD.

The Critical Assessment of Genome Interpretation (CAGI, 'kā-jē) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In the experiment, participants are provided genetic variants and make predictions of resulting phenotype, for ten challenges. These predictions are evaluated against experimental characterizations by independent assessors.

For example, in a challenge to predict Crohn's disease from exomes, several groups performed remarkably well, with one group achieving a ROC AUC of 0.94. The experiment also revealed important population structure to Crohn's disease in Germany. In another challenge, two groups were able to successfully map a significant number of Personal Genome Project complete genomes to corresponding trait profiles.

Other challenges were to predict which variants of BRCA1, BRCA2, and the MRN complex are associated with increased risk of breast cancer; to associate exomes, variants, and disease in lipid diseases; to predict how variants in p53 gene exons affect mRNA splicing; to predict how well variants of p16 tumor suppressor protein inhibit cell proliferation; and to identify potential causative SNPs in disease-associated loci.

Overall, CAGI revealed that the phenotype prediction methods embody a rich representation of biological knowledge, making statistically significant predictions. However, the accuracy of prediction on the phenotypic impact of any specific variant was unsatisfactory and of questionable clinical utility. The most effective predictions came from methods honed to the precise challenge. Prediction methods are clearly growing in sophistication, yet there are extensive opportunities for further progress.

Complete information about CAGI may be found at http://genomeinterpretation.org.

## 174

**The UG2G initiative: A study of disease susceptibility in 7000 individuals from Uganda using whole genome sequencing and genotyping approaches.** *D. Gurdasani[1], T. Carstensen[1], S. Fatumo[1], C.S. Franklin[1], E. Wheeler[1], I. Tachmazidou[1], J. Huang[1], A. Karabarinde[2], G. Asiki[2].* 1) Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) Medical Research Council/Uganda Virus Research Institute, Uganda.

Given the genetic diversity in Africa, and the current lack of understanding of genetic determinants of disease susceptibility in the region, novel efforts to advance this understanding by large-scale genome wide association studies (GWAS) are essential. The UG2G (Uganda 2000 Genomes) project builds on previous initiatives to examine African genetic diversity (African Genome Variation Project), representing a substantial advance in large-scale African genomics. Its primary objectives are to 1. Examine the heritability and genetic determinants of a wide range of biomedical traits to identify novel susceptibility loci among Africans; 2. Provide a global resource for researchers, by generating the largest African reference panel for imputation to date, and a repository of information relating to genetic variation in Africa, relevant to population genetics and fine-mapping efforts; 3. Strengthen research capacity, training, and collaboration across the region; 4. Provide novel insights into population demographic history in the region; and 5. Characterise in detail de novo mutation rates and recombination patterns in Africa. This initiative comprises whole genome sequence (WGS) and dense genotype data on 2000 and 5000 individuals, respectively, from the General Population Cohort (GPC), Uganda. The GPC is an annual survey that collects detailed information on over 50 traits, including cardio-metabolic, anthropometric, haematological, liver, renal function and infectious disease traits from ~20,000 individuals across 25 villages in south-western Uganda. We carried out low coverage 4x WGS (Illumina HiSeq 2000) of 2000 individuals, representing multiple ethno-linguistic groups, in addition to dense genotyping (Illumina Omni 2.5M chip array) of 5000 individuals. For in-depth characterisation of variation, we additionally sequenced trios at high coverage (30x). Based on preliminary data, we identify a novel locus associated with HbA1C levels in the HbA2 gene on chr16 (P=6.9e-19). This 3.8kb deletion is common among Africans (MAF=0.25) and rare among Europeans (MAF=0.005), and has been previously shown to be strongly associated with several haematological traits among African-Americans—consistent with our findings from the GPC. Our findings provide proof-of-concept that in addition to providing an invaluable resource to researchers worldwide, such large-scale efforts in Africa can identify novel susceptibility loci associated with complex disease traits.

## 175

**Long-range Haplotype Mapping in Hispanic/Latinos Reveals Loci for Short Stature.** *G. Belbin[1,5,6], D. Ruderfer[2,4,3], K. Slivinski[5], MC. Yee[8], J. Jeff[5], O. Gottesman[5], EA. Stahl[2,3,5,6], RJF. Loos[5,7], EP. Bottinger[5], EE. Kenny[1,4,5,6].* 1) Department of Genetics and Genomics, Icahn School of Medicine at Mt Sinai, New York, NY; 2) Broad Institute, Cambridge, MA; 3) Division of Psychiatric Genomics, Icahn School of Medicine at Mt Sinai, New York, NY; 4) Center for Statistical Genetics, Icahn School of Medicine at Mt Sinai, New York, NY; 5) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mt Sinai, New York, NY; 6) Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mt Sinai, New York, NY; 7) The Mindich Child Health and Development Institute, Icahn School of Medicine at Mt Sinai, New York, NY; 8) Carnegie Institution for Science, Dept. of Plant Biology, Stanford, CA.

The Hispanic/Latino (HL) population of Northern Manhattan represents a diverse recent diaspora population, with 95% of the individuals reporting having grandparents born outside of the United States. Of these 43% report grandparents born in Puerto Rico, 23% the Dominican Republic, 13% Central America, and 5%, 4%, and 2% from Mexico, South America, and Europe respectively. Despite complex patterns of migration, admixture, and diversity, strong signatures of cryptic relatedness persist amongst HLs. We have detected long-range genomic tract sharing (>3cM), or identity-by-descent (IBD), across 5,194 HL in the Mount Sinai BioMe Biobank. We observed an average population level IBD sharing of 0.0025 in HL, which is 2.5- and 5-fold higher than that observed in BioMe European- and African-American populations, respectively. We hypothesize that these patterns of recent migration and genetic drift may drive some otherwise rare functional alleles to detectable frequency. We clustered groups of homologous IBD tracts (n= 112,250) segregating in this HL population. We observed that IBD clusters represent a class of low frequency alleles (median minor allele frequency = 0.0077, s.d.=0.0015). We performed a genome-wide association of the IBD clusters, or 'population-based linkage', to detect loci implicated in height, a highly heritable polygenic trait. 15 independent loci surpassed our empirically derived genome-wide significance threshold of <4.4710-4, 11 of which replicated in an independent cohort of BioMe HLs. Strikingly, two regions confer strong recessive effects. In the case of the top hit on 9q32 (MAF< 0.005; p<8×10-6), homozygous non-referent individuals were shorter by 6" or 10", for men or women, respectively, compared to the population mean (5' 7" and 5' 2" for men and women, respectively). In addition, IBD haplotypes in the 9q32 cluster harbored a significant enrichment of Native American ancestry (p<1×10-16). Finally, this interval contains a number of biologically compelling candidate genes, including COL27A1 and PALM2. This study demonstrates that rich population structure, rather than being a confounding factor in biomedical discovery efforts, may be leveraged to reveal novel genetic associations with complex human traits.

## 176

**A Haplotype Reference Panel of over 31,000 individuals and Next Generation Imputation Methods.** *S. Das on behalf of Haplotype Reference Consortium.* University of Michigan, Ann Arbor, Michigan.

Genotype imputation is now a key tool in the analysis of human genetic studies, enabling array-based genetic association studies to examine the millions of variants that are being discovered by advances in whole genome sequencing. Examining these variants increases power and resolution of genetic association studies and makes it easier to compare the results of studies conducted using different arrays. Genotype imputation improves in accuracy with increasing numbers of sequenced samples, particularly for low frequency variants. The goal of the Haplotype Reference Consortium is to combine haplotype information from ongoing whole genome sequencing studies to create a large imputation resource. To date, we have collected information on >31,500 sequenced whole genomes, aggregated over 20 studies of predominantly European ancestry, to create a very large reference panel of human haplotypes where ~50M genetic variants are observed 5 or more times. These haplotypes can be used to guide genotype imputation and haplotype estimation. In preliminary empirical evaluations, our panel provides substantial increases in accuracy relative to the 1000 Genomes Project Phase 1 reference panel and other smaller panels, particularly for variants with frequency <5%. I will describe our evaluation of strategies for merging haplotypes and variant lists across studies and advances in methods for genotype likelihood-based haplotype estimation that can be applied to 10,000s of samples. I will also summarize new methods for next generation imputation that perform faster and require less memory than contemporary methods while attaining similar levels of imputation accuracy. Our full resource is available to the community through imputation servers that enable scientists to impute missing variants in any study and respect the privacy of subjects contributing to the studies that constitute the Haplotype Reference Consortium. The majority of haplotypes will also be deposited in the European Genotype Archive.

## 177

**A rare variant local haplotype sharing method with application to admixed populations.** *S. Hooker[1,2], G.T. Wang[1,2], B. Li[1,2], Y. Guan[2,3], S.M. Leal[1,2].* 1) Center for Statistical Genetics, Baylor College of Medicine, Houston, TX., USA; 2) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX., USA; 3) Department of Pediatrics, Baylor College of Medicine, Houston, TX., USA.

With the advent of next generation sequencing there is great interest in studying the involvement of rare variants in complex trait etiology. For many complex traits sequence data is being generated on DNA samples from African Americans and Hispanics to elucidate rare variant associations. Analyses of admixed populations present special challenges due to spurious associations which can occur because of confounding. However using information on admixture and local ancestry can also be highly beneficial and increase the power to detect associations in these populations. Here a local haplotype sharing (LHS) method (Xu and Guan 2014) was extended to test for rare variant (RV) associations in admixed populations. Previously the Weighted Haplotype and Imputation-based Test (WHAIT) (Li et al. 2010) was proposed to test for rare variant associations using haplotype data. The RV-LHS method unlike WHAIT, does not require reconstruction of haplotypes which can be both computationally intensive and error prone. Additionally the RV-LHS uses information on local ancestry which is particularly advantageous when analyzing admixed populations. Results will be shown from simulation studies performed for rare variant data from an admixed population. Both Type I and II errors are evaluated for the RV-LHS method. Additionally the power of the RV-LHS method is compared to WHAIT as well as several other non-haplotype-based rare variant association methods including the combined multivariate collapsing (CMC) (Li and Leal, 2008), Variable Threshold (VT) (Price et al. 2010) and Sequence Kernel Association Test (SKAT) (Wu et al. 2010). Several heart, lung and blood phenotypes were analyzed using sequence data on African-Americans from the NHLBI-Exome Sequencing Project to better evaluate the performance of the RV-LHS compared to other rare variant association methods.

## 178

**Rare mutations associating with serum creatinine and chronic kidney disease.** *G. Sveinbjornsson[1], E. Mikaelsdottir[1], R. Palsson[2,3], O.S. Indridason[3], H. Holm[1], A. Jonasdottir[1], A. Helgason[1,4], S. Sigurdsson[1], A. Jonasdottir[1], A. Sigurdsson[1], G.I. Eyjolfsson[5], O. Sigurdardottir[6], O.T. Magnusson[1], A. Kong[1,7], G. Masson[1], P. Sulem[1], I. Olafsson[8], U. Thorsteinsdottir[1,2], D.F. Gudbjartsson[1,7], K. Stefansson[1,2].* 1) DeCode Genetics, Reykjavik, Reykjavik, Iceland; 2) Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland; 3) Internal Medicine Services, Landspitali - The National University Hospital of Iceland, 101 Reykjavik, Iceland,; 4) Department of Anthropology, University of Iceland, 101 Reykjavik, Iceland; 5) Icelandic Medical Center (Laeknasetrid) Laboratory in Mjodd (RAM), 109 Reykjavik, Iceland; 6) Department of Clinical Biochemistry, Akureyri Hospital, 600 Akureyri, Iceland,; 7) School of Engineering and Natural Sciences, University of Iceland, 101 Reykjavik, Iceland,; 8) Department of Clinical Biochemistry, The National University Hospital of Iceland, 101 Reykjavík, Iceland.

Chronic kidney disease (CKD) is a term applied to irreversible loss of kidney function and is a serious public health problem. It is diagnosed through the measurement of serum creatinine (SCr) which reflects glomerular filtration rate. Here, we impute sequence variants identified by sequencing the whole genomes of 2,230 Icelanders into 81,656 chip-typed individuals and 112,630 relatives of chip-typed individuals with SCr measurements. In addition to replicating established loci, we discovered missense and loss of function variants associating with SCr in three solute carriers (SLC6A19, SLC25A45, SLC47A1) and two E3 ubiquitin ligases (RNF186, RNF128). With SCr effects of four variants between 0.085 and 0.129 standard deviations these rare variants have a greater impact on SCr than the previously reported common variants with a maximum effect of 0.069. We tested the variants associating with SCr for association with CKD in a sample of 15,594 Icelandic cases and 291,420 controls and found significant associations at three of them. Of note were four mutations in SLC6A19 that associate with reduced SCr, three of which have been shown to cause Hartnup disease.

## 179

**Rare coding variants in collagen genes increase risk of adolescent idiopathic scoliosis.** *G. Haller[1], D. Alvarado[1], J. Buchan[1], K. McCall[1], P. Yang[1], C. Cruchaga[3], M. Harms[2], A. Goate[3], M. Willing[4], E. Baschal[5], N. Miller[5], C. Wise[6,7,8,9], M. Dobbs[1,10], C. Gurnett[1,2,4].* 1) Department of Orthopaedic Surgery, Washington University School of Medicine, Saint Louis, MO; 2) Department of Neurology, Washington University School of Medicine, Saint Louis, MO; 3) Department of Psychiatry, Washington University School of Medicine, Saint Louis, MO; 4) Department of Pediatrics, Washington University School of Medicine, Saint Louis, MO; 5) Department of Orthopaedic Surgery, University of Colorado, Denver, CO; 6) Department of Orthopaedic Surgery, Texas Southwestern Medical Center at Dallas, Dallas, TX; 7) Department of Pediatrics, Texas Southwestern Medical Center at Dallas, Dallas, TX; 8) McDermott Center for Human Growth and Development, Texas Southwestern Medical Center at Dallas, Dallas, TX; 9) Seay Center for Musculoskeletal Research, Texas Scottish Rite Hospital for Children, Dallas, TX; 10) Shriners Hospital for Children, Saint Louis, MO.

Adolescent idiopathic scoliosis (AIS) is the most common form of childhood spinal deformity, affecting ~3% of the population. Unfortunately, little is known of its etiology or genetic basis. To understand the role of rare variants in AIS, exome sequence data was generated for 154 unrelated AIS cases. We observed novel heterozygous mutations in musculoskeletal collagen genes in 32% (49/154) of unrelated affected individuals with AIS compared with 12% (508/4300) of controls ($P = 2.8 \times 10^{-8}$, Fisher's exact test), a finding which was strengthened by addition of Sanger sequenced family members ($P = 2 \times 10^{-14}$, FamSKAT). Additionally, we observed significantly more rare musculoskeletal collagen gene coding variants per person among AIS cases compared to controls (1.9 vs. 1.2 variants, t-test $p = 2.3 \times 10^{-8}$) which remained significant after removal of individuals harboring novel variants (1.8 vs. 1.1 variants, t-test $p = 3.5 \times 10^{-7}$), suggesting a polygenic burden of collagen rare variants. We validated these findings by resequencing a subset of six musculoskeletal collagen genes in a second cohort of AIS patients, and again observed significantly more novel variants per gene as well as a significantly higher burden of rare missense variants per individual in AIS cases compared with controls. Examination of the functional effects of two novel AIS-associated variants (*COL1A1* His48Glu and *COL1A2* Pro1016His) *in vitro* revealed increased intracellular type I procollagen compared to WT, suggesting a collagen processing defect. Overall, our analysis suggests that as much as 10% of AIS can be explained by collagen gene variation. These results establish alterations in a subset of collagen genes as a major genetic risk factor for AIS.

## 180

**Rare Coding Variants Are Associated with Osteoporotic Fracture: A Large-scale Exome-Chip Analysis of 44,130 Adult Caucasian Men and Women in CHARGE and GEFOS Consortia.** *Y. Hsu[1,2,3], K. Estrada[2,4], P. Leo[5], A. Teumer[6], C. Liu[7], C. Medina-Gomez[8], H. Zheng[9], R. Minster[10], LP. Lyytikäinen[11], R. Pengelly[12], R. Cruz Guerrero[13], L. Oei[8], M. Claussnitzer[1], M. Kahonen[14], C. Cooper[15], A. Hannemann[16], D. Karasik[1], A. Uitterlinden[8], LA. Cupples[7], JA. Riancho Moral[17], J. Holloway[12], E. Duncan[18], T. Lehtimäki[11], T. Harris[19], H. Wallaschofski[16], B. Richards[9], F. Rivadeneira[8], M. Brown[20], D. Chasman[21], D. Kiel[1].* 1) HSL Institute for Aging Research, Harvard Medical School, Boston, MA; 2) BROAD Institute of MIT and Harvard, Cambridge, MA; 3) Molecular and Integrative Physiological Sciences, Harvard School of Public Health, Boston, MA; 4) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 5) University of Queensland Diamantina Institute, Brisbane, Australia; 6) Interfaculty Institute for Genetics and Functional Genomics, University of Greifswald, Greifswald, Germany; 7) Biostatistics Dept. Boston University, Boston, MA; 8) Erasmus Medical Center, Rotterdam, Netherlands; 9) Departments of Medicine, Human Genetics, Epidemiology and Biostatistics, McGill University, Montreal, Canada; 10) Department of Human Genetics and Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA; 11) Department of Clinical Chemistry, Fimlab Laboratories, Tampere, Finland; 12) Human Genetics and Genomic Medicine, University of Southampton Faculty of Medicine, Southampton, UK; 13) University of Santiago de Compostela, Santiago de Compostela, Spain; 14) Department of Clinical Physiology, University of Tampere School of Medicine, Tampere, Finland; 15) University of Southampton, Southampton, UK; 16) Institute of Clinical Chemistry and Laboratory Medicine, Institute for Community Medicine, University Medicine Greifswald, University of Greifswald, Greifswald, Germany; 17) Hospital U.M. Valdecilla-IFIMAV, University of Cantabria, Santander, Spain; 18) Royal Brisbane and Women's Hospital, Brisbane, Queensland, Australia; 19) Intramural Research Program, National Institute on Aging, Bethesda, MD; 20) Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Brisbane, Australia; 21) Brigham and Women's Hospital and Harvard Medical School, Boston, MA.

Bone mineral density (BMD) is the most commonly studied skeletal phenotype, yet osteoporotic fracture (FX) is the most important clinical consequence of low BMD. Importantly, FX has a BMD-adjusted heritability of 35%~69% in Caucasians, suggesting its genetic architecture may not be the same as low BMD. Identifying associated and causal variants is a necessary step to study the underlying biology of FX risk. The availability of the exome genotyping chips with 236K coding variants (non-synonymous, splice sites and stop-altering SNPs selected from ~18,000 genes) provides a feasible way to identify potentially causal variants in the exome. We conducted an exome-chip meta-analysis to identify novel functional coding variants that are associated with FX risk. The exome-chip was genotyped in 10 cohort studies comprising 44,130 old adult participants (8,781 FX cases and 35,349 controls; 86% female). FX (excluding fingers, toes, skull, pathological fractures and those resulting from high trauma) phenotypes were obtained by interview and confirmed in most studies through either clinical or X-ray confirmation. Since observed MAF of most genotyped variants (78%) was ≤ 1%, we performed gene-based collapsing tests (allele-count and SKAT tests) in each study to identify rare coding variants (MAF ≤ 1%) associated with FX. Covariates adjusted in the models included age, age2, sex, weight, height, ancestral genetic background, site for multisite studies and estrogen use. For family-based studies, a kinship matrix was incorporated into test statistics. An inverse-variance fixed effect meta-analysis (seqMeta package) was used to combine results. The most significant association was found in the *PPM1J* gene ($p=7.6\times10^{-12}$). Other novel associations that achieved exome chip -wide significance ($p < 4.2\times10^{-6}$, Bonferroni correction) were found in *WAC*, *DAZL*, *MRPS23* and *SMPDL3B* genes, which were not reported to be associated with BMD before. We also performed single variant association analysis for variants with MAF > 1%. The most significant, novel association was found for a missense SNP K450E in the *SLFN14* gene ($p=7\times10^{-6}$). Other strong associations ($p < 5\times10^{-5}$) were found for previously reported missense variants V667M and A1330V in *LRP5* and a common variant in *SLC25A13*. In summary, our analysis identified novel, rare, missense variants associated with FX. A large scale de-novo genotyping on selected variants in ~35,000 additional samples is underway to replicate these findings.

## 181

**Exploring the role of rare and low-frequency coding variants in adult height using an ExomeChip.** *M. Graff[1], K. Sin lo[2], K. Stirrups[3], C. Medina-Gomez[4], T. Esko[5,6], N.L. Heard-Costa[7], A.E. Justice[1], T.W. Winkler[8], L. Southam[3,9], C. Shurmann[10], J. Czajkowski[11], Y. Lu[10], K.L. Young[1], T.L. Edwards[12], A. Giri[12], C. Lindgren[13, 9], I.B. Borecki[11], K.E. North[1, 14], M. McCarthy[15,9], J.N. Hirschhorn[4, 13,16], P. Deloukas[3,17], F. Rivadeneira[4], T.M. Frayling[18], R.J.F. Loos[10], G. Lettre[19,20] For the BBMRI, the GOT2D, the CHARGE, and the GIANT Consortia.* 1) Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Montreal Heart Institute, Montreal, Quebec, Canada; 3) The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; 4) Departments of Epidemiology and Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; 5) Divisions of Endocrinology and Genetics and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA; 6) Estonian Genome Center, University of Tartu, Tartu, Estonia; 7) National Heart, Lung, and Blood Institute, the Framingham Heart Study, Framingham MA, USA; 8) Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany; 9) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 10) The Genetics of Obesity and Related Metabolic Traits Program, The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA;; 11) Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA; 12) Center for Human Genetics Research, Division of Epidemiology, Department of Medicine, Vanderbilt University, Nashville TN, USA; 13) Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA; 14) Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; 15) Oxford Centre for Diabetes, Endocrinology & Metabolism,University of Oxford, Oxford, UK; 16) Department of Genetics, Harvard Medical School, Boston, MA, USA; 17) William Harvey Research Institute, Barts and The London School of Medicine and Dentistry Queen Mary University of London Charterhouse Square, London, UK; 18) Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK; 19) Montreal Heart Institute,Université de Montréal, Montréal, Québec , Canada; 20) Faculty of Medicine, Université de Montréal, Montreal, Québec, Canada.

Adult height is a model complex phenotype being highly heritable, polygenic and accurately measured in large numbers of individuals. Genome wide association studies (GWAS) of 250,000 individuals have identified 697 common genetic variants in 424 loci at p<5×10-8 that explain ~20% of heritability. All common variants captured by GWAS are likely to explain ~60% of heritability, leaving 40% unexplained. To investigate the contribution of rare (<1% MAF) and low frequency (>1<5% MAF) coding variants to adult height we examined exome array genotype data containing approximately 215,000 protein coding variants in >230,000 individuals from 86 studies. We identified 55 variants at MAF <5% (8<1%, 47 >1<5%) associated with height at $p<5\times10^{-8}$. Forty-three of these variants occurred in 26 known height loci, including rare missense variants in *IHH* (MAF 0.3% $p = 4.7e^{-08}$) and *FBN2* (MAF 0.7% $p = 2.0e^{-21}$), two genes in which mutations cause monogenic growth defects. Twelve variants (2 rare, 10 low frequency) in 9 genes did not occur in known GWAS loci (>500kb) and strongly implicate the genes *PDE5A, DLG5, AMOTL1, SERPINA1, ZNF646, ZBTB7B, LAMB2, DUSP1* and *IL11* in human growth for the first time. The largest effect size was with a missense variant in *PDE5A* where heterozygous carriers were on ~2 cm taller/shorter ($p=9.5e^{-12}$). In addition, we identified a previously undetected common insertion deletion (MAF 10%, $p=1\times10^{-20}$, ~0.4 cm change in height) in a known locus that results in a frameshift in exon 1 in *CPNE1*. We conclude that exome chip sample sizes of similar magnitude to GWAS will likely identify low frequency variants of larger effect. However, given that the exome array only covers ~1% of the genome we will expect fewer loci. A large fraction of low frequency coding variant associations overlapped with known loci, and further studies will be needed to dissect whether these are due to linkage disequilibrium with known signals or represent independent functional variants.

## 182

**Increased Frequency of De novo Copy Number Variations in Congenital Heart Disease by Integrative Analysis of SNP Array and Exome Sequence Data.** *J.T. Glessner[1,2], A.G. Bick[3], K. Ito[3], J. Homsy[3], L. Rodriguez-Murillo[4,5], M. Fromer[5,6,7], E. Mazaika[3], B. Vardarajan[8], M. Italia[9], J. Leipzig[9], S.R. DePalma[3], R. Golhar[1], S.J. Sanders[10,11], B. Yamrom[12], M. Ronemus[12], I. Iossifov[12], A.J. Willsey[10,11], M.W. State[10,11], J.R. Kaltman[13], P.S. White[9], Y. Shen[8], D. Warburton[14], M. Brueckner[15], C. Seidman[3], E. Goldmuntz[16], B.D. Gelb[4,5], R. Lifton[10,17], J. Seidman[3], W.K. Chung[18], H. Hakonarson[1,2].* 1) Ctr Applied Genomics, Children's Hosp Philadelphia, Philadelphia, PA; 2) Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 3) Department of Genetics, Harvard Medical School, Boston, MA; 4) Mindich Child Health and Development Institute, Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY; 5) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 6) Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY; 7) Division of Psychiatric Genomics in the Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY; 8) Department of Systems Biology, Columbia University Medical Center, New York, NY; 9) Center for Biomedical Informatics, Children's Hospital of Philadelphia, Philadelphia, PA; 10) Department of Genetics, Yale University, New Haven, CT; 11) Department of Psychiatry, University of California, San Francisco, San Francisco, CA; 12) Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; 13) Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (NIH), Bethesda, MD; 14) Department of Genetics and Development (in Medicine), Columbia University Medical Center, New York, NY; 15) Department of Pediatrics, Yale University, New Haven, CT; 16) Division of Cardiology, Children's Hospital of Philadelphia; and Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA; 17) Department of Medicine, Yale University, New Haven, CT; 18) Departments of Pediatrics and Medicine, Columbia University Medical Center, New York, NY.

Congenital heart disease (CHD) is among the most common birth defects. Most cases are of unknown etiology. To determine the contribution of de novo copy number variants (CNVs) in the etiology of sporadic congenital heart disease (CHD), we studied 538 CHD trios using genome-wide dense single nucleotide polymorphism (SNP) arrays and/or whole exome sequencing (WES). Results were experimentally validated using digital droplet PCR. SNP and WES methods identified overlapping and unique CNVs; WES data enabled consistent detection of 1-10 kb CNVs, an order of magnitude smaller than previous reports. We compared validated CNVs in CHD cases to CNVs in 1,301 healthy control trios. The two complementary high-resolution technologies identified 65 validated de novo CNVs in 53 CHD cases, with median size of 110 kb. CNVs were identified as small as 2 kb (arrays) and 0.1 kb (WES). Two-thirds of CNVs contained five or fewer genes. A significant increase in CNV burden was observed when comparing CHD trios with healthy trios, using either SNP array (p=7×10-5, Odds Ratio (OR)=4.6) or WES data (p=6×10-4, OR=3.5). A significant increase in CNV burden remained, even after removing from these analyses the 16% of de novo CNV loci that were previously reported as pathogenic (p=0.02, OR=2.7). Among CHD cases, we observed recurrent de novo CNVs on 15q11.2 encompassing CYFIP1, NIPA1, and NIPA2 and single de novo CNVs encompassing DUSP1, JUN, JUP, MED15, MED9, PTPRE SREBF1, TOP2A, and ZEB2, genes that interact with established CHD proteins such as NKX2-5 and GATA4. Integrating de novo variants in WES and CNV data suggests that ETS1 is the pathogenic gene altered by 11q24.2-q25 deletions in Jacobsen syndrome and that CTBP2 is the pathogenic gene in 10q subtelomeric deletions. Collectively, these results demonstrate a significantly increased frequency of rare de novo CNVs in CHD patients compared with healthy controls and suggest several novel genetic loci for CHD.

## 183

**Context-specific eQTLs implicate differential genomic regulatory mechanisms in obese and lean Finns.** *A. Ko[1,2], R.C. Cantor[1], E. Nikkola[1], M. Alvarez[1], B. Pasaniuc[1,3,4], K.L. Mohlke[5], M. Boehnke[6], F.S. Collins[7], J. Kuusisto[8], M. Laakso[8], P. Pajukanta[1,2].* 1) Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA; 2) Molecular Biology Institute at UCLA, Los Angeles, CA; 3) Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles ,CA; 4) Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA; 5) Department of Genetics, University of North Carolina, Chapel Hill, NC; 6) Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI; 7) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 8) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.

Obesity is a worldwide health problem with an alarming prevalence of 35% in the U.S. Although obesity is highly heritable (up to 70%), the increase in its prevalence during the last 2 decades points to gene and environment (GxE) interactions in its etiology. However, small GxE effects are difficult to detect in human, especially given the uncontrollable environment in human studies. We hypothesized that in obesity, the cellular environment affects activation of genomic regulatory mechanisms. To this end, we investigated whether an intermediate mRNA expression phenotype differs between obese and lean individuals, utilizing RNA-sequencing of subcutaneous adipose, a highly obesity-relevant tissue. Our design identifies GxE interactions under either an obese or a lean physiological state by investigating expression quantitative trait loci (eQTL) within subgroups ascertained initially based on their body mass index (BMI). We will also investigate how the metabolic disturbances related to obesity differentiate these two groups in future studies. We performed eQTL mapping with 650,000 genotyped SNPs (MAF>1%) and 15,000 expressed genes in ~600 individuals from the Finnish METSIM cohort subdivided into two groups of ~300 based on the BMI median. All eQTLs that were only observed in the obese group but not in the lean or overall groups were considered obese-specific (OS), and vice versa for lean-specific (LS). After correcting for multiple testing using an FDR<0.05, we discovered 6,689 cis (1Mb) and 1,414 trans OS eQTLs; and 3,992 cis and 2,068 trans LS eQTLs, respectively. The frequencies of the context-specific OS and LS cis eQTL SNPs did not differ between these groups (P>0.1), suggesting that expression profiles change due to different genomic regulatory mechanisms in OS versus LS cis eQTLs. This was supported by our preliminary data demonstrating that the OS cis eQTLs were enriched for both obesity GWAS loci (P=0.003 Fisher's exact (FE)), and all active enhancers (P=0.01 FE), whereas LS cis eQTLs were enriched for non-synonymous (ns) variants ($P=2.5\times10^{-16}$ Binomial). For instance, a common ns variant (MAF 9%) in the gene, DAK, regulates the expression of FADS2 in the lean Finns exclusively. FADS2 encodes the fatty acid desaturase 2 enzyme, a key regulator of unsaturated fatty acids. Our results suggest that the distinct cellular environment drives specific genomic regulatory mechanisms depending on obese and lean physiological conditions in adipose tissue.

## 184

**Use of low read depth whole genome sequence data to examine the genomic architecture of commonly measured lipid sub-fractions: the UK10K study.** *J. Huang[1], J. Min[2], V. Iotchkova[1], M. Mangino[1], A. Gaye[1], M. Kleber[1], G. Malerba[1], M. Cocca[1], T. Michela[1], I. Tachmazidou[1], H. Chheda[1], A. Manning[1], A. Wood[1], R. Scott[1], T. Gaunt[1], W. Zhang[1], F. Rivadeneira[1], N. Soranzo[1], N. Timpson[2], UK10K Consortium Cohorts Group.* 1) Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK; 2) MRC Integrative Epidemiology Unit, University of Bristol, UK.

Lipid levels are highly heritable risk factors for coronary artery disease and vascular outcomes, and augmentation of their levels therapeutically has well characterized impacts on risk profile. Discovery of inherited variation influencing lipid levels has relevance for treatment of cardiovascular disease through the identification of novel therapeutic targets and the assessment of the causal impact of specific profiles. As part of the UK10K Consortium Cohorts arm we undertook low read depth whole genome sequencing in individuals from two deeply phenotyped British cohorts, the St Thomas' Twin Registry (TwinsUK) and the Avon Longitudinal Study of Parents and Children (ALSPAC). In this lipid focused initiative, we analyzed the cohorts dataset with the following three aims: (i) to discover additional variants of low and intermediate frequency that are poorly represented in genome-wide SNP arrays; (ii) to characterize the contribution of rare, deleterious variants to population level trait variance, and to rank human genes according to the presence of these variants; (iii) to systematically evaluate the contribution of low-frequency and rare variants in non-coding regions of the genome. We further combined the UK10K data with meta-analysis of 18 additional cohorts (N=~ 25,000) with SNP array data imputed to the combined UK10K+ Genomes Project haplotype reference panel. Overall, we describe 21 novel independent genome-wide significant loci (defined as $P < 5\times10^{-8}$). Of these, 18 map to known GWAS signals, including several low-frequency variants (MAF <5%) near *PCSK9*, *LPL*, *APOC3*|*APOA4*|*APOA1*, *PCSK7*, *CETP*, *LIPG*, *LDLR* and *APOE*. We further use Bayesian approaches to fine-map genetic association signals at known loci, and explore possible functional consequences of putative causative variant through integration with functional annotation maps of the human genome. Finally, we report association and replication results of rare variant tests (MAF <1%) from sequence kernel based association testing at both exome-wide and genome-wide scale. Overall, our results demonstrate the value for low-coverage whole genome sequencing and shed lights on the role of rare genetic variants on serum lipids levels.

## 185

**Population-specific imputations identify a deleterious coding variant in *ABCA6* associated with cholesterol levels: The Genome of the Netherlands.** *C.M. Van Duijn[1], E.M. van Leeuwen[1], M.A. Swertz[2,3], D.I. Boomsma[4], P.E. Slagboom[5], G.B. van Ommen[6], C. Wijmenga[3], P.I.W. de Bakker[7,8] on behalf of the CHARGE lipids WG and the Genome of the Netherlands consortium.* 1) Epidemiology, Erasmus MC, Rotterdam, Zuid Holland, Netherlands; 2) University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, the Netherlands; 3) University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands; 4) Department of Biological Psychology, VU University Amsterdam, Amsterdam, the Netherlands; 5) Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, the Netherlands; 6) Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands; 7) Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands; 8) Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.

Genome-wide association studies (GWAS) have identified a large number of loci associated with blood lipid levels and there is no evidence that this approach has reached its limits. Despite the fact that rare functional variants are known to play a major role in lipid metabolism for long, there has been a lack of success finding such variants in population-based studies. The power of searches for rare functional variants may improve by the use of reference sets based on next generation sequencing (NGS) specific to distinct populations. This allows a better quality imputation of rare variants in particular those of which the frequency is increased within a specific population. The Genome of the Netherlands (GoNL) project led to a reference for the Dutch population, based on NGS in 250 parent-offspring trios (13x coverage). Nine large Dutch epidemiological cohorts imputed with the GoNL reference panel and conducted a GWAS on high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, total cholesterol (TC) and triglyceride (TG) levels. Meta-analysis comprised around 35,000 samples and revealed both rare (minor allele frequency (MAF) < 0.05) and common variants (MAF >= 0.05) associated with HDL (N=60 variants), LDL (N=142 variants), TC (N=134 variants) and TG (N=16 variants) in known and novel loci . A comparison of allele frequencies shows that these variants are more rare than those identified by Teslovich *et al.* and Willer *et al.* In addition, we replicated the majority of the lipid loci described by Teslovich *et al.* and Willer *et al.* despite a sample size of about 20% of the other studies. We replicated 4 novel loci by association analysis in cohorts from the CHARGE consortium which used the 1000 Genomes reference panel (Phase 1 integrated release v3, April 2012). Among the novel replicated loci is rs77542162 for both LDL ($N_{total}$=57,593, *p*-value=1.33E-12) and TC ($N_{total}$=65,305, *p*-value=7.31E-11). This exonic variant, which is predicted to be damaging, is located on chromosome 17 within the *ABCA6* gene (ATP-binding cassette, sub-family A (ABC1), member 6). The frequency of this variant is 3.65 fold enriched in the GoNL reference panel as compared to the 1000 Genomes reference panel. The effect size of this variant (0.135 for LDL and 0.140 for TC) is very similar to those observed for other well-known lipid genes, such as *LDLR* and *CETP*. This suggest that next generation sequencing effort may yield clinically relevant findings.

## 186

**Null alleles at NPC1L1, the therapeutic target for the LDL-lowering drug ezetimibe, and protection from coronary heart disease.** *N. Stitziel[1], S. Kathiresan[2], Myocardial Infarction Genetics Consortium.* 1) Division of Cardiology, Department of Medicine and Division of Statistical Genomics, Washington University School of Medicine, St. Louis MO; 2) Center for Human Genetic Research and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA; Program in Medical and Population Genetics, Broad Institute, Cambridge, MA.

BACKGROUND: Ezetimibe lowers plasma low-density lipoprotein (LDL) cholesterol by inactivating the Niemann-Pick C1-Like 1 (NPC1L1) gene product. Though widely prescribed, it is unknown if ezetimibe therapy reduces risk for coronary heart disease; a phase III randomized controlled trial including >18,000 participants is ongoing to test this hypothesis. Human mutations that inactivate a gene encoding a drug target can mimic drug action and thus can be used to infer the potential efficacy of that drug. METHODS: We sequenced the exons of NPC1L1 in a total of 4,703 cases with early-onset coronary heart disease and 5,090 controls free of coronary heart disease, identified carriers of null alleles (nonsense, splice-site, or frameshift mutations), and tested the association of carrier status with early coronary heart disease. In replication analyses, we sequenced NPC1L1 in an additional 1,528 coronary heart disease cases and 3,820 controls and genotyped a specific null alleles - p.Arg406X - in 14,708 coronary heart disease cases and 24,628 controls. We also tested the association of carrier status with plasma lipids. RESULTS: In the discovery study, we identified four NPC1L1 null alleles carried by ten individuals. Carrier status was associated with an 89% reduced risk of early coronary heart disease (OR 0.11; 95% CI 0.01 to 0.88; P=0.01). The observation of reduced risk was replicated in independent samples (P=0.02). Combining discovery and replication, carriers had a 78% reduced risk for coronary heart disease (OR 0.22; 95% CI 0.08 to 0.63; P=9×10-4; 4 carriers among 20,939 cases and 36 carriers among 33,538 controls) as well as 15.5 mg/dl lower plasma LDL cholesterol (P=0.02). CONCLUSIONS: Naturally occurring mutations in humans that disrupt NPC1L1 function are associated with reduced risk for coronary heart disease and lower LDL cholesterol. .

## 187

**Trans-ethnic genome-wide association study identifies 15 new genetic loci influencing blood pressure traits, and implicates a role for DNA methylation: the International Genetics of Blood Pressure (iGEN-BP) Study.** *M. Loh[1], F. Takeuchi[2], N. Verweij[3], X. Wang[4], W. Zhang[1,5], International Genetics of Blood Pressure Study.* 1) Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom; 2) Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Tokyo, Japan; 3) University of Groningen, University Medical Center Groningen, Department of Cardiology, Groningen, The Netherlands; 4) Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore; 5) Ealing Hospital NHS Trust, Middlesex, UK.

High blood pressure (BP) is a major risk factor for myocardial infarction, stroke and chronic kidney disease. Genome-wide association (GWA) studies have identified 50 genetic loci influencing BP in predominantly European populations. However the mechanisms linking the identified loci to BP phenotypes remain largely unknown. We carried out a trans-ethnic GWA and replication study of 5 BP phenotypes (systolic BP, diastolic BP, pulse pressure, mean arterial pressure and hypertension) amongst up to 313,449 individuals of East Asian, European and South Asian ancestry, with imputation of 2.1M HapMap2 SNPs. We then studied DNA methylation at the identified loci to investigate potential DNA regulatory mechanisms. We identify and replicate 38 loci at P<10-*9*, of which 15 are novel ($P=7.5\times10^{-10}$ to $P=2.5\times10^{-15}$) There was little evidence for heterogeneity of effect between the ethnic groups in either the GWA or replication data. The sentinel SNPs are enriched for variants associated with adiposity, type 2 diabetes, coronary heart disease, and kidney function in published GWA studies ($P=2.5\times10^{-3}$ to $1.6\times10^{-10}$). A weighted genetic risk score also predicts increased left ventricular mass, circulating levels of NT-proBNP, cardiovascular and all-cause mortality (P=0.05 to $1.1\times10^{-20}$). Thirteen of the sentinel SNPs are strongly associated with methylation of nearby CpG sites ($P=10^{-6}$ to $10^{-300}$). Mendelian randomisation experiments implicate a role for methylation mediating the relationship between the DNA sequence variation and BP at the newly identified loci. At five loci the CpG sites are associated with expression of their nearest gene ($P=10^{-4}$ to $10^{-14}$). The sentinel SNPs and leading CpG sites point to genes involved in vascular (*IGFBP3*, *KCNK3*, *PDE3A* and *PRDM6*) and renal (*ARHGAP24*, *OSR1*, *SLC22A7*, *TBX2*) structure and function. Our results identify 15 novel loci influencing BP, and provide the first evidence for DNA methylation as a potential mediator of the relationship between DNA sequence variation and BP phenotypes. Our findings provide the rationale for functional studies to evaluate DNA methylation as a potential therapeutic target in cardiovascular risk reduction.

## 188

**Pathologically Different than Coronary Artery Disease, Myocardial Infarction has a Minimal Heritable Component.** *B. Horne[1,2], S. Knight[1,2].* 1) Intermountain Heart Institute, Intermountain Medical Center, Salt Lake City, UT; 2) Genetic Epidemiology Division, Department of Medicine, University of Utah, Salt Lake City, UT.

Background: Coronary artery disease (CAD) and its features, including lesion location and morphology, are heritable and dozens of loci for CAD have been validated. In contrast, only one myocardial infarction (MI) locus has been validated among MI cases and non-MI CAD controls (the ABO gene). MI occurs among patients with mild CAD and severe CAD at a similar rate. No study has tested whether MI is heritable independent of its required atherosclerotic substrate. This study evaluated whether MI is familial among patients with mild CAD and with clinically-significant 3-vessel CAD. Methods: The Intermountain Genealogy Registry includes >23 million individuals within extended family pedigrees, which includes ~700,000 patients seen since 1994 who have linked clinical data in the Intermountain Healthcare data warehouse. A genealogical index of familiality (GIF: 10,000 times the average pairwise kinship coefficient) was calculated for MI and CAD cases who had at least 3 generations and 12 ancestors in their pedigree (and for all patients in the registry as a control GIF). MI subjects were studied if they had only 1-vessel CAD (≥70% stenosis) or mild CAD (20-60% stenosis) with (I) MI when younger (n=194, females aged≤70 years, males≤65 years) or (II) MI at any age (n=489), or if they had 3-vessel CAD (≥70% stenosis) with (III) MI at any age (n=314) or with (IV) no MI ever (n=317, females aged>70 years, males>65 years). The phenotypic characteristics of high-risk pedigrees (n=10 cases per pedigree) were also examined. Results: The GIF was 0.448 for controls and GIF=0.395 in cases with younger MI and mild CAD (p>0.05), while GIF=0.440 in cases with MI at any age and mild CAD (p>0.05 vs. control). In contrast, among cases with MI at any age and 3-vessel CAD, the GIF was 0.564 (p<0.001 vs. control GIF=0.448) and, for older cases who never had an MI but had 3-vessel CAD, the GIF was similarly elevated at 0.645 (p<0.001 vs. control). Conclusion: While CAD was confirmed to be heritable regardless of MI events, the presence of MI without substantial CAD was not familial. This suggests that MI per se is only weakly heritable and that prior claims that MI aggregates in families were likely measuring the inheritance of atherosclerosis/CAD loci and not loci pre-disposing to MI per se. Further investigation of these findings in other cohorts and identification of all environmental triggers of MI are needed.

## 189

**Is type 2 diabetes a causal risk factor for coronary artery disease? Multivariate mendelian randomization to test causal relationships among cardiometabolic traits.** *R. Do[1, 2], M. Daly[2,3], B. Neale[2,3], S. Kathiresan[1,2].* 1) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA; 2) Broad Institute, Cambridge, MA, USA; 3) Analytic and Translational Genetic Unit, Massachusetts General Hospital, Boston, MA, USA.

Observational epidemiological studies have established correlations among cardiometabolic traits such as type 2 diabetes (T2D), body mass index (BMI) and coronary artery disease (CAD); however, causal inference based on these correlations is challenging. We have developed and implemented a human genetics approach, multivariate mendelian randomization (MMR), that leverages genetic effect sizes of common SNPs to dissect causal influences amongst a set of correlated traits. We first examined the epidemiological association between all pairs of ten cardiometabolic traits, including plasma lipids (low density lipoprotein cholesterol (LDL-C), high density lipoprotein cholesterol (HDL-C), triglycerides), BMI, T2D, systolic blood pressure, fasting insulin, fasting glucose, C-reactive protein and CAD. In 9,610 European Americans from the Atherosclerosis Risk in Communities study, we observed widespread correlations between the ten traits (76 out of 90 comparisons are significantly correlated with $P<0.05$). Using MMR, we were able to infer causal relationships among the ten traits. Since our approach requires a set of SNPs where effects on the biomarkers of interest are precisely measured, we leveraged estimates of effects from published large-scale genome-wide association studies for each trait. From this analysis, we highlight three sets of observations. First, we observed that, across 32 SNPs associated with BMI at genome-wide significance, the BMI-increasing allele was correlated with effects on five other traits: lower HDL-C (beta=-0.33, $P=8.6\times10^{-6}$); higher triglycerides (beta=0.23, $P=5.9\times10^{-9}$), higher risk for T2D (beta=0.85, $P=8.8\times10^{-9}$), higher fasting insulin (beta=0.17, $P=4.4\times10^{-7}$) and higher C-reactive protein (beta=0.37, $P=6.7\times10^{-8}$). Second, across 19 SNPs associated with fasting insulin at genome-wide significance, the fasting-insulin-raising allele was correlated with effects on two traits: lower HDL-C (beta=-1.28, $P=1.1\times10^{-5}$) and higher triglycerides (beta=0.87, $P=2.9\times10^{-5}$). Finally, in a single model including effect sizes of all nine risk factors and CAD, we observed only four significant predictors for CAD including the strength of a SNP's effect on LDL-C (beta=0.45, $P=1.9\times10^{-19}$), on triglycerides (beta=0.31, $P=4.1\times10^{-6}$), on blood pressure (beta=0.35, $P=0.003$), and T2D (beta=0.13, $P=1.1\times10^{-4}$). These results suggest that as few as four factors causally relate to CAD and that MMR can be used to infer causal relationships between correlated traits.

## 190

**Vertical Transmission of Autism Spectrum Disorder.** *N. Risch[1,2], L. Shen[2], Y. Qian[3], M. Massolo[3], L. Croen[2,3].* 1) Institute for Human Genetics, UCSF, San Francisco, CA., USA; 2) Research Program on Genes, Environment and Health, Kaiser Permanente Division of Research, Oakland, CA, USA; 3) Autism Research Program, Kaiser Permanente Division of Research, Oakland, CA, USA.

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with a complex pattern of inheritance. Twin and family studies have consistently provided evidence of a genetic contribution through an elevated concordance in MZ versus DZ twins, and higher recurrence risks for full sibs compared to maternal half sibs. However, an elevated recurrence risk to maternal versus paternal half-sibs, a significant birth order effect, and increasing risk with short inter-birth interval also suggest maternal effects. It has generally been assumed that individuals with ASD rarely if ever reproduce. Thus, no studies of risk to the children of parents with ASD have been reported. Here we report results from the first such study. It is based on the Kaiser Permanente Northern California (KPNC) Family Linkage Database (KPFLD). All KPNC members age 26 or less born prior to 2009 were linked to state of California birth certificates to identify parents. Parent information was then matched to the KPNC adult membership to identify nuclear families. In total, approximately 500,000 families were identified to form the KPFLD. The KPFLD was then linked to the KPNC ASD registry, which maintains records on all individuals in KPNC with a presumptive diagnosis of ASD. The linkage of the KPFLD with the ASD registry identified 225 nuclear families with an ASD affected mother, 172 families with an ASD affected father, and 37 families with both an ASD affected mother and father (dual matings). The risk to children of affected mothers was observed to be 31.7% (155 affected out of 489 total children); the risk to children of affected fathers was somewhat lower at 25.1% (80 affected out of 319 total children). The risk to children of dual matings was 46.8% (37 affected out of 79 total children). We also noted a birth order effect in children of affected mothers. In the children of affected mothers and unaffected fathers, the risks were 36.4%, 29.1% and 25.5% to first-born, second-born and later-born children, respectively. For dual matings, the risks were 67.6%, 34.5% and 15.4% to first-born, second-born and later-born children, respectively. By contrast, the recurrence risk to full sibs of ASD children was 11.4% (370 affected out of 3243). The remarkably high risk to children of affected parents demonstrates the vertical transmissibility of ASD, and the birth order effects again emphasize the potential role of maternal environment in modulating genetic risks.

## 191

**Epidemiological and genomic studies suggest a significant effect of comorbidity of intellectual disability towards estimates of autism prevalence.** *S. Girirajan[1], J.A. Rosenfeld[2], A. Polyak[1].* 1) Biochem & Molecular Biol, Pennsylvania State University, University Park, PA. 16802; 2) Signature Genomics Laboratories, PerkinElmer, Inc., Spokane, WA 99207.

Current estimates of autism prevalence fail to take into account the effect of comorbidity of related neurodevelopmental phenotypes. We analyzed 11 years (2000-2010) of longitudinal data on approximately 6 million children per year from special education enrollment and 5,894 children referred for clinical microarray testing due to autistic features. We measured changes in autism prevalence and frequency, age-specific rates, and copy number variation (CNV) burden of comorbid features in autism. We found a 331% increase in the prevalence of autism from 2000-2010 based on special education enrollment. This prevalence increase corresponded to a concomitant decrease in the prevalence of intellectual disability (ID). The combined prevalence of ID and autism showed a 15% increase from 2000-2010 (22-fold less than autism prevalence alone). The decrease in ID prevalence equaled 61.5% of the increase in autism prevalence suggesting recategorization of diagnosis of ID to autism. Further, the frequency of phenotypes comorbid with autism was influenced by ascertainment, CNV burden, and gender. Comorbidity rates were higher for children in the clinical testing population compared to recent epidemiological estimates (40.3% vs 55.5%, p<0.0001, OR=1.84) and even higher among those with genomic disorders (64% vs 55.5%, p<0.0001, OR=2.51). The frequency of comorbid features varied across autism related genomic disorders such as 16p11.2 deletion (50%), 16p11.2 duplication (100%), 15q11.2 deletion (77%), and 1q21.1 duplication (56%). In the clinical testing population females showed a higher frequency of comorbid features (60% vs 46%, $p=8.2\times10^{-5}$, OR=1.27) and were more likely to manifest epilepsy comorbid with autism compared to males (p=0.02, OR=1.63). Among 1,588 autistic individuals carrying rare CNVs, females showed a higher large CNV burden for autism comorbid with ID (p=0.006, OR=2.14). Additional evidence suggests related genetic factors for varying comorbid features: individuals with epilepsy and autism were more likely to have ID compared to those with autism only ($p=7.72\times10^{-13}$, OR=1.97), and similar CNV burdens were observed for both autistic individuals with epilepsy and those with ID (p=0.44). These observations support a global neurodevelopmental model where causal genes are shared between autism and related neurodevelopmental phenotypes, and phenotypes can be modified by genetic background to affect the age of onset, severity and variability of the disorder.

## 192

**Partial deletion of the Monoamine Oxidase A (*MAOA*) gene in a 3-generation family with two severely affected intellectually disabled males and a healthy female carrier.** *N. de Leeuw[1,4], M.I. Schouten[1], R. van Beek[1], R. Pfundt[1], M.M. Verbeek[2,3,4], H.G. Brunner[1].* 1) Department of Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands; 2) Department of Neurology, Radboud University Medical Center, Nijmegen, the Netherlands; 3) Department of Laboratory Medicine, Radboud University Medical Center, Nijmegen, the Netherlands; 4) Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands.

Genome wide high resolution array analysis in a 12-year-old boy with severe intellectual disability (ID), absent speech, autism, stereotyped behaviour, cheerful, pulmonary stenosis and dysmorphic features revealed a 30 kb interstitial deletion in Xp11.3 which resulted in the loss of the last six exons of the *MAOA* gene (MIM 309850). The deletion was also detected in the healthy mother with normal intelligence and in a 79-year-old maternal grand-uncle with severe intellectual disability. Monoamine oxidase A (MAOA) is the major enzyme catalyzing the oxidative deamination of monoamine neurotransmitters, such as serotonin (5-hydroxytryptamine (5-HT)) and (nor)adrenalin, and plays a critically important role in brain development and functions. Abnormal MAOA activity has been reported to be associated with borderline ID (Brunner syndrome [MIM 300615]), depression, abnormal behaviour and autism spectrum disorder, but the molecular basis for these disease processes is unclear. Therefore, subsequent metabolic testing using UPLC was performed in this family and showed that serotonin levels in serum were dramatically increased in the index patient (3,892 nmol/l) and to a lesser extent in his mother (1,635 nmol/l), whereas 5-hydroxyindole-3-acetic acid (5-HIAA) was not detectable in serum of either two. The urinary concentration of 5-HIAA as well as homovanillic acid (HVA) and vanillylmandelic acid (MVA) were very low in the index patient, which is in agreement with the serum levels and indicates MAOA deficiency. Unexpected low levels of (nor)adrenalin and dopamine were measured in the patient's urine. Urinary values for 5-HIAA, HVA, VMA, (nor)adrenalin and dopamine were all normal in the mother. Although few single nucleotide variants in the *MAOA* gene and multiple gene deletions involving MAOA have been reported, this is the first report of a deletion restricted to the *MAOA* gene, without the concomitant loss of any other gene in Xp11.3. While the genetic findings show that only exons 10 through 15 of *MAOA* are deleted, we note that the severity of ID and of the biochemical changes suggest that MAOB activity might be compromised through a position effect on the neighbouring *MAOB* gene (MIM 309860). The clinical, genetic and metabolic findings in these three family members with the same deletion (two affected males and one healthy female) provide new insights into the molecular basis of neuropsychiatric disorders associated with MAOA dysfunction.

## 193

**A *Drosophila* model for 16p11.2 deletion shows differential sensitivity to gene dosage.** *J. Iyer, L. Pizzo, T. Le, P. Patel, L. Thomas, K. Vadodaria, S. Girirajan.* Departments of Biochemistry and Molecular Biology and Anthropology, The Pennsylvania State University, University Park, PA 16802.

Recent studies have suggested a significant role for rare copy number variants (CNVs) towards a wide range of phenotypes, such as intellectual disability, autism, and schizophrenia. This has posed challenges in the diagnosis and management of affected individuals carrying these CNVs; lack of functional candidate genes within the region has impaired our understanding of the mechanisms of disease. We tested dosage sensitivity of *Drosophila melanogaster* orthologs of human genes within the autism-associated 16p11.2 region. Taking advantage of the tissue-specific expression system conferred by the UAS-Gal4 system, we used RNA interference (RNAi) to achieve eye-specific (*GMR*-Gal4) and neuron-specific (*Elav*-Gal4) knockdown of 18 fly lines representing eight fly orthologs of human genes within 16p11.2. We developed novel quantitative methods to accurately assess the severity of neurodevelopmental phenotypes and to identify candidates that are sensitive to gene-dosage alteration. Combining data from gene expression using qPCR with phenotypic measurements of fly RNAi lines enabled us to correlate the effect of gene dosage alterations to severity. Seven out of eight genes within 16p11.2 showed dosage-dependent change in severity including *C16ORF53*, *CDIPT*, *KCTD13*, *FAM57B*, *ALDOA*, *PP419C*, and *MAPK3*. In aggregate, male flies with reduced dosage of 16p11.2 genes were more severe (p=0.0015) than female flies suggesting a potential biological component to the observed male bias in neurodevelopmental disorders. Neuron-specific reduction of gene expression resulted in lethality for *ALDOA* and *PP419C*. Further, to assess the impact of gene dosage on fly head size, we measured the distance between conserved bristle landmarks for six genes. Three genes, *KCTD13* (5.4 sd units), *MAPK3* (6.6 sd units) and *C16ORF53* (5.6 sd units) showed significant (p<0.001) macrocephaly at 50% expression levels and reflect the macrocephaly phenotypes observed in individuals with 16p11.2 deletions and autism. Interestingly, significant (p<0.001) microcephaly phenotypes were observed when the dosage of *CDIPT* (12.3 sd units) and *CORO1A* (11.2 sd units) was further reduced (expression <30%). We propose that conserved genes within rare CNVs associated with genomic disorders are involved in differential roles, sensitive to gene dosage alteration, and this complex interaction between *cis* and *trans* genetic variants contribute to the observed phenotypic variability.

## 194

**The discovery of integrated gene networks for autism.** *O. Penn[1], F. Hormozdiari[1], E. Borenstein[1,2,3], E.E. Eichler[1,4], SSC Sequencing Consortium.* 1) Department of Genome Sciences, University of Washington, Seattle, WA; 2) Department of Computer Science and Engineering, University of Washington, Seattle, WA; 3) Santa Fe Institute, Santa Fe, NM; 4) Howard Hughes Medical Institute, Seattle, WA.

Despite extensive genetic heterogeneity underlying disorders such as autism spectrum disorders (ASD) and intellectual disability (ID), there is compelling evidence that risk genes will map to a much smaller number of biologically functional modules. To discover modules enriched for *de novo* mutations in probands, we developed a novel computational method (MAGI) that simultaneously considers protein-protein interaction and RNAseq expression profiles during brain development. Applying the method to recent exome sequencing data from 1116 ASD and ID patients, we discovered two distinct significant modules ($p$<0.005) that differ in their properties and associated phenotypes. The first module consists of 80 genes associated with the Wnt and Notch signaling pathways, as well as with the SWI/SNF and NCOR complexes, and exhibits the highest expression early during embryonic development. Probands with truncating mutations in this module are enriched for micro and macrocephaly (KS test $p$=0.013). The second module consists of 24 genes associated with synaptic function, including long-term potentiation and calcium signaling, and shows higher levels of postnatal expression. Probands with *de novo* mutations in these modules are found to have lower IQ compared to probands with mutations outside these modules. In addition, missense mutations in both modules are predicted to be more deleterious using the C-scores measurement ($p$<$10^{-6}$), providing a useful approach for detecting potentially pathogenic missense. Applying the method independently to epilepsy and schizophrenia exome sequencing cohorts, we found marked overlap among modules suggesting shared common neurodevelopmental pathways. For example, *ZMNYND11*, *CUL3*, *SMARCC2*, and *GRIN2A* are part of both the ASD/ID and the schizophrenia modules and are indeed mutated in both cohorts. Adding mutations found in the full Simons Simplex Collection (SSC) to the analysis (2661 probands in total) resulted in the identification of more refined and biologically coherent modules. Five significant modules ($p$<0.05) were discovered, covering a total of 249 genes associated with the SWI/SNF complex, Wnt signaling, long term potentiation, proteasome, and ubiquitin mediated proteolysis pathways. Our approach provides a molecular framework for reducing the genetic heterogeneity of these diseases and a method for identifying *de novo* missense mutations important in ASD/ID etiology.

## 195

**Transcriptome Sequencing of Human Aging Brain Tissue Uncovers Widespread Genetic Effects on Splicing Alternations in Alzheimer's disease.** *T. Raj[1,2,3,4], J. Xu[1], C. McCabe[3], J.A. Schneider[5], N. Pochet[1,3,4], D.A. Bennett[5], P.L. De Jager[1,2,3,4].* 1) Department of Neurology, Brigham and Women's Hospital, Boston, MA; 2) Department of Medicine, Division of Genetics, Brigham and Women's Hospital, Boston, MA; 3) The Broad Institute, Cambridge, MA; 4) Harvard Medical School, Boston, MA; 5) RUSH Alzheimer's Disease Center, RUSH University Medical Center, Chicago, IL.

Recent studies suggest that alternations in RNA processing might be a key mechanism in Alzheimer's disease (AD) pathogenesis. Here, we used RNA-sequencing to survey genome-wide transcriptome profiles from dorsolateral prefrontal cortex in 541 subjects of a longitudinal age-related cognitive decline cohort. To explore the mechanisms that drive transcriptional variation associated with cognitive decline, we performed a systematic QTL study aimed at identifying local and distal genetic effects on exons, genes, and isoforms in the brain transcriptome. We identified hundreds of RNA splicing events that correlate with individual's trajectory of age-related cognitive decline. We observe widespread genetic effects on gene and isoform expression levels, with variants influencing transcript ratios (tr-QTLs) of 2,651 (FDR 0.10) genes in *cis* and 264 (bonferroni p< 0.05) genes in *trans*. Of these, ~15 % of tr-QTLs are retained-intron transcripts subject to non-sense mediated decay, suggesting an overabundance of unspliced transcripts in the aging brain that is consistent with recent reports. We also detect several AD susceptibility alleles influencing isoform ratios in *cis* including genes involved in immune processes (*PILRB*, *C4A*, and *C4B*) and others (*PICALM* and *ABCA7*). These results include a variant in the sortilin-related VPS10 protein (*SORCS3*) associated with isoform ratios of over 17 genes in *trans*, some of which are located within AD-associated loci including *SORL1*, *CLU*, and *PILRA*. Interestingly, we identified six *trans*-QTL effects on isoform ratios mediated through a *cis*-association with a known splicing factor hnRNP C: the SNP is associated with decreased expression of hnRNP in *cis* and splicing of the six target genes. At a less stringent threshold, we observe similar pattern of altered regulation of RNA targets mediated by other splicing factors including CUG-BP1 (*CELF1*) and SR proteins that are targets of *cis*-eQTL. We plan to pursue UV Cross-Linking and Immunoprecipitation (CLIP) experiments in the same tissue to validate the downstream RNA targets of the splicing factors. Overall, our comprehensive study of the aging brain transcriptome provides evidence that (1) genetic variants, including AD susceptibility variants, regulate the generation of mRNA isoforms through effects on RNA splicing factors and (2) the splicing machinery is altered by the pathophysiology of AD.

## 196

**Exome array analysis of 30,582 individuals confirms late-onset Alzheimer's disease (LOAD) risk from common variants and identifies novel rare LOAD susceptibility variants: the International Genomics of Alzheimer's Project (IGAP).** *A.C. Naj[1], S.J. van der Lee[2], M. Vronskaya[3], R. Sims[3], J. Jakobsdottir[4], C. van Duijn[2], L.-S. Wang[5], P. Amouyel[6], S. Seshadri[7], J. Williams[3], G. Schellenberg[5], International Genomics of Alzheimer's Project (IGAP).* 1) Biostatistics & Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; 2) Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, Netherlands; 3) Institute of Psychological Medicine and Clinical Neurosciences, Medical Research Council (MRC) Centre for Neuropsychiatric Genetics & Genomics, Cardiff University, Cardiff, Wales, UK; 4) Icelandic Heart Association, Kopavogur, Iceland; 5) Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA; 6) Institut Pasteur de Lille, Lille, France; 7) Department of Neurology, Boston University School of Medicine, Boston, MA.

Genomic studies of late-onset Alzheimer's disease (LOAD) have identified as many as 22 genes with common susceptibility variants and multiple genes with high-risk, rare coding variants (*APP*, *PSEN2*, *TREM2*, *PLD3*). While sequencing studies of thousands of individuals remain prohibitively costly, exome arrays, which capture nearly 250,000 putatively functional variants, provide a viable alternative for examining rare coding variants in large samples. Here we present findings from single variant association analyses on 15,788 AD cases and 19,795 controls of European ancestry from the Alzheimer's Disease Genetics Consortium (ADGC), Gene and Environmental Risk in Alzheimer's Disease (GERAD), and Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortia. The three AD genetics consortia genotyped samples on the Illumina v1.0 (~90%) and v1.1 (~10%) HumanExome arrays, and performed genotype calling and quality control (QC) independently. After QC, data were available for analysis on 30,582 individuals in ten datasets with 203,267 SNPs polymorphic in at least one dataset. We used a score test approach with adjustment for age, sex, and population substructure in each study, and then performed meta-analysis across studies using the R/seqMeta package. Of 16 variants present in more than one dataset and associated with LOAD at $P<5\times10^{-8}$, 12 mapped to the *APOE* region, including four variants which captured the *APOE* ε2 association and one mapped to the *PICALM* GWAS signal. The remaining four loci included rare (MAFs=0.002-0.005) missense variants, of which three demonstrated genome-wide significance (from $P=2.47\times10^{-9}$ to $P=2.24\times10^{-8}$) with large effect sizes (ORs=2.40-4.97). The fourth variant examined was excluded from further consideration, as it demonstrated associations in opposite directions in the two datasets in which genotyping was successful. Additionally, associations of $P<10^{-5}$ were observed at common variants at 11 known LOAD risk loci. Additional genotyping of these and other strongly associated variants ($P<10^{-5}$) is underway in a replication dataset of more than 10,000 samples to confirm these findings. Gene-based analyses are also being performed and will be presented.

## 197

**Low-frequency variant imputation identifies rare variant candidate loci in a genome-wide association study of late-onset Alzheimer disease.** *K.L. Hamilton[1], B.W. Kunkle[1], A.C. Naj[2], W.R. Perry[1], A. Partch[2], O. Valladares[2], L.S. Wang[2], G. Jun[3], J. Chung[3], M.A. Schmidt[1], G.W. Beecham[1], E.R. Martin[1], R.P. Mayeux[4], J.L. Haines[5], L.A. Farrer[3], G.D. Schellenberg[2], The Alzheimer's Disease Genetics Consortium.* 1) Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, FL; 2) Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 3) School of Medicine, Boston University, Boston, MA; 4) Taub Institute of Research on Alzheimer's Disease, Columbia University, New York, NY; 5) Department of Epidemiology & Biostatistics, Case Western Reserve University, Cleveland, OH.

Recent Genome-wide Association Studies (GWAS) have identified 19 susceptibility loci, in addition to APOE, for Late-onset Alzheimer Disease (LOAD). The Alzheimer Disease Genetics Consortium (ADGC) conducted a rare variant focused GWAS using 31 datasets (including 16 new datasets), increasing our sample to 14,459 cases and 14,556 controls. Two datasets are family-based and thus potentially enriched for rare variants. Using IMPUTE2, all data were imputed using a 1000 Genomes 2012 reference set of over 37 million variants, many of which are low-frequency variants and indels. Association analysis was conducted adjusting for age, gender and population substructure. Individual datasets were analyzed with SNPTest using the score test (recommended for low-frequency variants), and within-study results were meta-analyzed in METAL. The imputation increased the number of high quality analyzable variants by 54.2% over the previous analyses done in collaboration with the International Genomics of Alzheimer Project (IGAP) (~15,500,000 variants vs ~7,100,000 variants). Approximately 9 million and 6.5 million of the total analyzable variants had a minor allele frequency (MAF) <0.05 and <0.01, respectively. 102 low-frequency variants (defined as MAF≤0.02) demonstrated genome-wide significance (P≤5×10-8), with 65 of these variants being located in the APOE region. While other previously associated common loci produced significant signals, none showed significance for low-frequency variants. 37 variants with MAF≤0.02 in 9 novel loci demonstrated genome-wide significance (5 are family dataset specific), with one of the loci including the previously reported TREM2 gene (P=1.64×10-8; MAF=0.002). Other significant loci include the actin-related gene, ACTN1 (P=3.92×10-11; MAF=0.001), which has been previously associated with LOAD related changes in hippocampal gene expression, and LRFN5 (P=4.40×10-13; MAF=0.002), a cell adhesion molecule that promotes neurite outgrowth and synapse maturation in hippocampal neurons. An additional 37 variants at 17 loci were nominally associated (P=3.92×10-6), including a variant in LUZP2, a gene previously implicated in LOAD in Amish through linkage and association analyses. Using an imputation set with a large amount of rare variation, we identified several novel rare variant candidate loci for LOAD, giving support to the hypothesis that rare and low-frequency variant imputation can identify novel associations with disease.

## 198

**Leveraging genetic variation from over 55,000 exomes to explore patterns of functional constraint on human protein-coding genes.** *K. Samocha[1,2,3], M. Lek[1,2], D. MacArthur[1,2], M. Daly[1,2,3], Exome Aggregation Consortium.* 1) Massachusetts General Hospital, Boston, MA; 2) Broad Institute of Harvard and MIT, Cambridge, MA; 3) Harvard Medical School, Boston, MA.

A critical challenge in human disease genetics is distinguishing disease-causing variants from the thousands of rare, potentially functional variants identified in any human genome. While many methods focus on predicting the deleteriousness of an individual variant, a complementary approach to improve the power for causal variant discovery is to focus on variants found in genes that typically show unusually low levels of variation in healthy individuals; these genes must be subject to a high level of functional constraint, increasing the probability that novel variants observed in them will have a deleterious phenotypic impact. The medical relevance of such genes has already been established (see, for example, Epi4K-Consortium 2013).

We extended earlier work and used a model of mutation to predict the expected amount of rare (minor allele frequency <0.001) variants for each gene in a cohort of over 55,000 reference individuals jointly called as a part of the Exome Aggregation Consortium (see abstract by M. Lek et al). This model accurately predicts the number of observed synonymous (and putatively neutral) variants per gene (Pearson's correlation = 0.94). It can also be used to define a metric of constraint for both missense and loss-of-function (LoF) variation. With over 55,000 individuals, this model has unprecedented power to confidently identify genes that are depleted for LoF variants, and to provide direct estimates of the human-specific selective constraint for each gene. The magnitude of empirical variation data in this analysis enables several powerful analyses for the first time. Genome-wide estimates have suggested 20-30% of missense variants may be equivalent to LoF variants. Here we evaluate this on a gene-by-gene basis and find strikingly that genes with equivalent strong selection against LoF variation show deficits of missense variation suggesting a wide range - from close to 0% to more than 50% - with which missense variants show equivalent deleterious impact. This provides a critical and heretofore missing parameter in estimating the pathogenic probability for a novel missense variant. Missense constraint can also be used to highlight specific coding regions within each gene that are intolerant of mutational changes. We explored various approaches to evaluate missense constraint for segments of genes, and aligned *de novo* variants from patients with autism to these locations for further confirmation of disease-relevance.

## 199

**Unveiling the Genetics Architectures of Rare Coding Variants in Blood Lipids Levels via Large Scale Meta-analysis.** *D. Liu on behalf of the Global Lipids Genetics Consortium.* Department of Public Health Sciences, Pennsylvania State University, Hershey, PA.

In order to understand the impact of rare coding variants on plasma lipids levels, we are performing large scale meta-analysis of exome-array data in the Global Lipids Genetics Consortium and developing novel statistical methods that summarize and describe the underlying genetic architecture. Association statistics were aggregated across 289,500 individuals from 94 studies with plasma lipids levels and exome-array genotypes. Meta-analysis of single variant and gene-level association tests were performed centrally using RAREMETAL for HDL, LDL, triglyceride and total cholesterol levels. After quality control, a total of 235,526 variants segregated in the aggregate dataset. Of these, 89.5% and 69.9% have minor allele frequency <1% and <0.1%, respectively. Among the coding variants, 204,618 are nonsynonymous (NS) and 7,438 are loss-of-function (LOF) variants. Using this data, we identified known and novel lipids genes, e.g. LDLR, APOC3, NPC1L1, APOH, ABCA6, as well as novel variants within known lipids loci. The impact of these new genes on MI are being investigated. In addition to identifying novel loci, large datasets provide us a unique opportunity to learn the genetic architecture of rare coding variants. To achieve this goal, we develop a novel empirical Bayesian model, which estimates genetic effect size distributions and posterior probability of associations for different classes of rare variants. The method only requires summary level data as input, allows the presence of multiple causative variants in the same gene, and accurately models the dependence between variants induced by linkage disequilibrium and relatedness. To apply the model to whole exome data, we also developed an efficient hybrid Expectation-Maximization and MCMC algorithm. Fitting the model to whole-exome data and adjusting for winner's curse, we showed that there is significant enrichment of causative variants among LOF alleles relative to NS variants (p<1e-5). The genetic effect sizes for LOF variants are ~3X larger on average than for NS variants. A fraction of >33% of the GWAS signals can be explained in full or in partial by rare coding variants. Jointly modeling effects of common and rare variants increases estimated heritability at GWAS loci by 20%. Taken together, our method and analysis provide one of the first comprehensive investigations of genetic architectures in >280,000 individuals, and establish the significance of studying rare coding variants for lipids traits.

## 200

**Efficient Bayesian mixed model analysis increases association power in large cohorts.** *P. Loh[1,2], G. Tucker[3], B. Bulik-Sullivan[2,4], B.J. Vilhjalmsson[1,2], H.K. Finucane[3], K. Galinsky[5], D.I. Chasman[6], B.M. Neale[2,4], B. Berger[3], N. Patterson[2], A.L. Price[1,2,5].* 1) Department of Epidemiology, Harvard School of Public Health, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 3) Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA; 4) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 5) Department of Biostatistics, Harvard School of Public Health, Boston, MA; 6) Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA.

Linear mixed models (LMM) are a powerful statistical tool for identifying genetic associations and avoiding confounding. However, mixed model analysis is computationally demanding, and is becoming infeasible as study sizes approach 100,000 samples. All existing methods rely on spectral analysis of a genetic relationship matrix (GRM) at time cost $O(MN^2)$ (where N = #samples and M = #SNPs). In addition, these methods implicitly assume an infinitesimal genetic architecture in which effect sizes are normally distributed, which can limit power (Yang et al. 2014 Nat Genet). Here, we present a far more efficient mixed model association method, BOLT-LMM, which requires only a small number of O(MN) iterations and increases power by modeling more realistic, non-infinitesimal genetic architectures via a Bayesian mixture prior on marker effect sizes. In the special case of the infinitesimal model, BOLT-LMM achieves results equivalent to existing methods at dramatically reduced time and memory cost. Algorithmically, BOLT-LMM performs O(MN)-time conjugate gradient and variational iterations to operate directly on raw genotypes stored compactly in memory, computing a retrospective score statistic that is robust to confounding while circumventing the GRM entirely. For a simulated data set of 100,000 samples typed at 300,000 SNPs, BOLT-LMM required <1 day and <8GB RAM, vs. >1 month and >150GB RAM required by existing mixed model methods; the fold-reduction in time and memory cost increases with sample size. In BOLT-LMM analysis of lipid traits in 23,294 samples from the Women's Genome Health Study (WGHS), the Bayesian non-infinitesimal model achieved up to a 7% (s.e. 1%) increase in chi-squared test statistics across known associated loci compared to standard mixed model analysis and an 8% increase compared to standard marginal analysis, consistent with simulations. In larger cohorts, theory and simulations show that the boost in chi-squared statistics - equivalent to a commensurate increase in effective sample size - increases with cohort size toward an asymptote of $1/(1-h^2_g)$, where $h^2_g$ is heritability explained by genotyped SNPs, leading to even larger increases in power. BOLT-LMM software is available at http://www.hsph.harvard.edu/alkes-price/software/.

## 201

**Recent demography and natural selection hamper power of rare variant association tests.** *L.H. Uricchio[1], J.S. Witte[2,3], R.D. Hernandez[3,4,5].* 1) Joint Bioengineering Graduate Group, UCSF and UC Berkeley, San Francisco, CA; 2) Department of Epidemiology and Biostatistics, UCSF, San Francisco, CA; 3) Institute for Human Genetics, UCSF, San Francisco, CA; 4) Institute for Quantitative Biosciences, UCSF, San Francisco, CA; 5) Bioengineering and Therapeutic Sciences, UCSF, San Francisco, CA.

Current findings suggest that common genetic variants may explain only a fraction of the heritability of complex human diseases, and recent research focus has turned toward rare variants. However, rare variants can only explain a large proportion of the additive genetic variance of complex traits when they have dramatically larger effect sizes than common variants. The simplest explanation for an inverse relationship between allele frequency and effect size is that natural selection prevents trait altering alleles from increasing in frequency. Unfortunately, most statistical tests for association between rare alleles and complex traits do not explicitly model natural selection. Moreover, recent demographic effects--such as the explosive growth experienced across many human populations--can also influence the genetic architecture of complex traits. To evaluate the impact of natural selection and demography on rare variant association tests, we jointly simulate DNA sequences and complex phenotypes under recently inferred models of human selection and multi-population demographic effects. We find that the statistical power of state-of-the-art rare variant tests (e.g., SKAT-O) is a decreasing function of the correlation between selection strength and effect size. This counterintuitive result means that rare variant-based statistical tests perform worst when rare variants explain a large proportion of the additive variance, even for large sample sizes in the thousands. We investigate the sensitivity of this conclusion to the length of the causal locus, the total number of causal loci, the shape of the distribution of selection coefficients and effect sizes, and sample size. Finally, we show that population genetic predictions of the relationship between allele frequency and the fraction of phenotypic variance explained could be used to characterize the distribution of effect sizes for complex human phenotypes.

## 202

**A statistical framework to leverage broad metabolite data in elucidating the associations between genetics and disease.** *C. Churchhouse, Slim Initiative in Genomic Medicine for the Americas (SIGMA) Type 2 Diabetes Consortium.* Broad Institute of Harvard and MIT, Cambridge, MA.

Genome-wide association studies have been applied to a broad spectrum of complex diseases, for which large consortial efforts have increased power to find true associations. Despite success in reproducibly identifying risk variants, GWAS have, in general, fallen short of elucidating pathophysiology. In combining genomic data with traits that may be biomarkers of risk or intermediary to a disease end point, we may shed more light on the underlying etiology. The advent of systematic metabolite profiling has enabled the quantification of thousands of metabolites in vivo, rendering the variation within the human metabolome accessible by analytical approaches, much like the genome. Large studies have been published in which broad panels of metabolites were analyzed as traditional GWAS or examined for links to disease risk, but there remain many challenges surrounding the use of metabolite data, some of which are analogous to those met in the field of quantitative genetics. This abstract describes progress on statistical methods developed to address these considerations. Specifically, these challenges include identifying and accounting for technical confounding, such as batch effects, which can introduce cryptic structure in the metabolite data. We have found, for example, that patterns of missing values relate to the order in which samples were profiled, requiring statistical quality control (QC) methods to avoid bias and false positives in downstream analyses. A further consideration is the highly correlated structure of metabolites resulting from the underlying molecular pathways through which they are related. Our approach leverages existing knowledge of these metabolic networks to both inform QC techniques and to reduce the dimensionality of the data and thus the penalty incurred for multiple hypothesis testing. Another challenge we address is potential confounding due to population structure and admixture that will become more problematic as metabolomics is applied to larger cohorts and a wider range of ethnicities. We will illustrate these methods on an empirical data set that includes ~12,000 metabolites measured in 865 serum samples collected at baseline in the longitudinal Mexico City Diabetes Study. Additionally, we have OMNI array, exome chip and exome sequence genotypes through which to investigate the application of these methods to understanding the associations between the triad of genetic variation, metabolism, and type 2 diabetes risk.

## 203

**Prioritizing Functional Variants in Genetic Association Studies.** *S. Sengupta, X. Wen, G. Abecasis.* Biostatistics, University of Michigan, Ann Arbor, MI.

Genome-wide association studies, which examine millions of genetic variants in thousands of individuals, have identified many complex trait associated loci. Most of these loci include many strongly associated variants and, often, variants which are not tested for association but are known to be in strong linkage disequilibrium with the variants exhibiting strongest evidence of association. The large number of variants actually or potentially showing evidence for association in each locus makes it challenging to prioritize likely functional variants at each locus. We reasoned that causal variants for each trait might share certain genomic features. For example, causal variants for lipid traits might preferentially overlap transcription factor binding sites active in liver, where important steps in lipid metabolism take place. More generally, causal variants for many traits might be non-synonymous variants that alter protein coding sequences. We develop a hierarchical model that identifies genomic features enriched among many associated loci and uses this information to prioritize likely functional variants in each locus. Our models can be fitted to summary statistics from individual studies, making it convenient to incorporate in ongoing genome-wide association study meta-analysis that can include 100,000s of individuals distributed across dozens of studies. We evaluate our method using simulations and application to genome-wide association study data for a variety of metabolic traits.

## 204

**A practical guide to study design, sample size requirement and statistical analyses methods for rare variant disease association studies.** *S.M. Leal, G.T. Wang, D. Zhang, Z. He, H. Dai, B. Li.* Center for Statistical Genetics, Department of , Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030.

We evaluated rare variant association (RVA) study designs and methods using real world data from NHLBI-Exome Sequencing Project (ESP) as well as exome sequence data which was simulated using state-of-the-art demographic models with purifying selection modeled after the empirical distribution of functional variants in ESP. Our simulated data is highly consistent with real world data distribution of singleton, doubleton and tripleton variants as well as the cumulative minor allele frequency (MAF) of variants for Europeans and Africans. Using resampled genotypes from ESP European American sequences, we evaluated relative power of 10 RVA methods by analyzing 16,568 genes across the genome, and we demonstrated that a method most powerful for one gene is not necessarily the most powerful for another, simply due to differences in the genomic sequence context, i.e., gene specific MAF spectrum and distribution of functional variants, rather than phenotypic model assumptions. Using simulated data of European samples we evaluated impact of phenotypic model, missing data, non-causal variants and choice of empirical MAF cutoff in RVA analysis. We found that the assumption of variable effects model favors variable threshold tests (e.g. VT) greatly, but the power gain of weighted burden tests (e.g. WSS) are marginal compare to the constant effect model. In the presence of strongly protective variants, SKAT-O/SKAT are consistently the most powerful tests, but they perform poorly when protective effect is mild compare with detrimental effect. The impact of non-causal variants and missing data are more significant than the choice of RVA methods, and the enrichment of functional variants is most crucial to the success of most RVA methods. The number of samples which need to be studied are highly dependent on gene size and the number of variant sites, for example under the assumption of moderate effect size for causal variants, i.e., odds ratio 2.0, for genes with short coding region lengths (~400bp), >90,000 samples are required to achieve a power of 80% to detect an association using an exome-wide significant level of $\alpha = 2.5\times10^{-6}$ while for average sized genes (~1,400bp), the required sample size is >50,000. We also showed that for exome chip design, the impact of exclusion of singletons and doubletons from observed samples are minimal compare to the impact of the large proportion of variants that were excluded from exome chip design due to inevitable sampling bias.

## 205

**A logistic mixed model approach to obtain a reduced model score for KBAC to adjust for population structure and relatedness between samples.** *G. Linse Peterson[1], J. Grover[1], B. Vilhjalmsson[2], G. Christensen[1], A. Scherer[1].* 1) Golden Helix, Inc, Bozeman, MT; 2) Harvard School of Public Health, Cambridge, MA.

Accounting for population structure, family structure, and inbreeding is a significant issue for burden and kernel association tests on rare variants from next generation DNA sequencing. Recent approaches such as the VC-score test and the methods outlined by Schaid have adjusted burden and kernel tests using linear regression mixed models and correcting for the population structure using a kinship matrix as a random effects matrix. However, these methods do not readily extend to a logistic regression framework; the method we present uses mixed-model logistic regression directly on a binary dependent variable to account for population structure and cryptic relatedness. We have implemented a solution that combines the power of a mixed model regression analysis with the ability to assess the rare variant burden using KBAC (Kernel-Based Adaptive Cluster method). While several optimizations are available for linear mixed model regression on a genome-wide scale, it is non-trivial to efficiently solve a logistic mixed model regression for every gene. Therefore, we have derived a transformed linear pseudo-model to solve the logistic mixed model equation optimized using EMMA (Efficient Mixed Model Algorithm), and we pre-compute and reuse the permutations for KBAC and the reduced models for those permutations. The result is an efficient logistic mixed model regression algorithm with a kinship random effects matrix for computing a modified score test for KBAC (MM-KBAC). In addition, the method for computing the kinship matrix can affect the power of the method to identify the gene(s) associated with the complex trait(s). Comparisons will be made between various methods for specifying the kinship matrix including IBS, IBD, and a pedigree-based matrix using GAW17 simulated data and 1000 Genomes data. We show that including a random effects matrix to account for population structure using a logistic model directly with KBAC results in an increased power to detect significant results and controls for Type I error when compared with family adjusting methods such as famSKAT or VC-score and methods assuming independence of samples including KBAC and SKAT-O for binary traits.

## 206

**A chromosome imbalance map of the human genome.** *M. Zarrei[1], J.R. MacDonald[1], R. Ziman[1], G. Pellecchia[1], D.J. Stavropoulos[2], D. Merico[1], S.W. Scherer[1,3].* 1) The Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, On, Canada; 2) Department of Pediatric Laboratory Medicine, Cytogenetics Laboratory, The Hospital for Sick Children, Toronto, On, Canada; 3) McLaughlin Centre, University of Toronto, Toronto, On, Canada.

Copy number differences between genomes are a major type of genetic variability. The chromosome imbalance map identifies genome regions that are prone to copy number variation (CNV) in apparently healthy individuals. We defined stringent quality and resolution requirements to select a “gold standard” subset of copy number variation studies from the Database of Genomic Variants. These variants, separately for deletions and duplications, were then clustered based on 50% reciprocal overlap to identify copy number variable regions (CNVRs) that were defined with the outmost coordinates of each cluster. Two chromosome imbalance maps were constructed, using different stringency levels. The inclusive map includes CNVRs supported by a minimum of two subjects, i.e. excluding all singleton variants, whereas the stringent map is composed of the regions that are supported by a minimum of two subjects and called in a minimum of two different studies. Approximately 9.5% of the human genome is variable according to the inclusive map, while 4.8% is variable according to the stringent map. The pericentric and subtelomeric regions of chromosomes show a particularly high rate of variation, and variability is correlated with presence of segmental duplications. We assessed the copy number variability of a comprehensive set of genomic features, with particular attention to exonic gene sequence. We found that the exons of all RefSeq genes are more variable compared to the entire genome for both gains and losses. However, higher copy number stability is observed for genes that are essential, or causally implicated in human disease (Mendelian disorders, cancer) or under negative selection for nonsynonymous variation. Exons of non-coding genes display greater variability, although highly conserved lincRNAs display a higher degree of stability. The enhancers and ultra-conserved elements are more stable than the genome whereas the proximal promoter regions are more variable than the genome. Functional category analysis revealed an enrichment in stable genes for macromolecular complexes (such as the proteasome), as well as pathways regulating cell cycle and organ development whereas olfactory receptors, xenobiotic metabolism and certain immune receptor families were found to be enriched in variable regions. Our chromosome imbalance map can be used as an effective tool for identifying variants within copy number variable regions from patient data in the clinical settings.

## 207

**Detection of known genomic regions and intragenic copy-number changes by an expanded exon-targeted array with comprehensive coverage of genes implicated in autism spectrum disorders (ASDs) and intellectual disability (ID).** *S.W. Cheung[1], P. Liu[1], T. Gambin[1], S. Gu[1], P. Hixson[1], C. Shaw[1], W. Bi[1], A. Breman[1], J. Smith[1], M. Haeri[1], A.N. Pursley[1], S. Lalani[1,2], C. Bacino[1,2], A.L. Beaudet[1,2], J.R. Lupski[1,2], P. Stankiewicz[1], A. Patel[1].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX,; 2) Department of Pediatrics, Baylor College of Medicine, Houston, TX.

With the rapid developments of whole- genome analysis techniques, an unprecedented opportunity to identify disease-causing genes for etiologically heterogenous conditions such as autism spectrum disorders (ASDs) and intellectual disability (ID) has become available. Multiple studies have now documented an excess of rare copy-number variants (CNVs) including exonic CNVs in patients with ASDs and ID. One strategy has been to focus on genes in shared pathways related to neuronal signaling and development, synapse function and chromatin regulation. As a result, we recently expanded our custom-designed oligo array to include >4800 disease or candidate genes with exon coverage and 60K SNP oligos. More specifically, coverage for all 50 genomic regions and 97% (120) of the 124 genes implicated in ASDs and ID [PMID: 24768552, table S6A and S6B], with an average of 85.65 oligonucleotide probes per gene, is included. 6499 patients have been evaluated by this array with indications provided for 4944 cases. 3016 patients (61%) studied had a clinical indication of developmental delay (DD), ID or autism. Overall, CNVs encompassing or disrupting a single gene were identified in 1536 cases (24%); 47% of these cases (726/1536) had an indication related with cognitive impairment. Examples of known pathogenic CNVs (both deletion and duplication) identified include the following genomic regions: 1q21 (16); DG (39); PWS/AS (21); 16p11.2 (24) and 15q13.3 (14). Examples of disease-causing exonic CNVs detected include: MECP2 (Rett syndrome and MECP2 Duplication syndrome); SHANK3 (Phelan-McDermid syndrome); and DMD (Duchenne muscular dystrophy). The more commonly observed CNVs in this cohort affecting known and candidate genes for ASDs/ID include RBFOX1 (27); NRXN1 (9); PARK2 (8); IMMP2L (25); AHI1 (4) and CNTN4 (4). Detection of recurring intragenic CNVs in this patient population will support reclassifying candidate genes implicated in neurodevelopmental disorders as disease-causing genes. For example, of the 10 cases with exonic CNVs in the candidate gene, ANKRD11, nine had indications of DD and ASDs. In summary, our approach not only enabled detection of known genetic conditions but has the potential to elucidate new disease genes. Additionally, the comprehensive coverage of genes implicated in ASDs and ID allows detection of exonic CNVs thus improving the diagnostic capability of this patient population through utilization of an already available clinical assay.

## 208

**Identification of pathogenic CNVs in a simplex autism cohort and measurement of effect size on cognitive, adaptive, and social function.** *A.E. Hare[1], D. Moreno De Luca[2], K.B. Boomer[3], S.J. Sanders[4,5], M.W. State[4,5], M. Benedetti[6], A.L. Beaudet[7], E.H. Cook[8], D.M. Martin[9], D.H. Ledbetter[1], C.L. Martin[1].* 1) Autism & Developmental Medicine Institute, Geisinger Health System, Danville, PA; 2) Department of Psychiatry, Yale University, New Haven, CT; 3) Department of Mathematics, Bucknell University, Lewisburg, PA; 4) Department of Genetics, Yale University, New Haven, CT; 5) Department of Psychiatry, UCSF, San Francisco, CA; 6) Simons Foundation, New York, NY; 7) Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX; 8) Institute for Juvenile Research, Department of Psychiatry, University of Illinois, Chicago, IL; 9) Departments of Pediatrics and Human Genetics, University of Michigan Medical Center, Ann Arbor, MI.

We conducted genome-wide copy number variant (CNV) analysis in 2,742 autism probands from the Simons Simplex Collection (SSC). CNVs ≥250 kb in size were classified as pathogenic based on clinical interpretation guidelines as determined by an expert panel of clinical cytogeneticists and geneticists. We analyzed *de novo* and inherited cases of recurrent (rec) CNVs (mediated by flanking segmental duplications) and *de novo* cases of non-recurrent (nrec) CNVs. This approach allowed us to identify rare, clinically significant pCNVs, which would not have reached statistical significance in a case/control study, by incorporating previously established evidence for pathogenicity. While statistically significant CNVs were previously identified in 1% of SSC cases, we identified pCNVs in 4.3% of cases, including 55 deletions (del), 56 duplications (dup), 2 unbalanced translocations and 6 sex chromosome aneuploidies (1 XXX, 2 XXY, and 3 XYY; removed from phenotypic analyses). pCNVs were identified in 13 rec and 41 nrec genomic regions. The most common rec pCNVs were dup 16p11.2 (n=12), dup 1q21.1 (n=11) and del 16p11.2 (n=9); all nrec pCNVs were unique. We measured the effect of carrying a pCNV on cognitive, adaptive, and social function by dividing the SSC probands into 5 categories: 1) no pCNV (nonCNV); 2) *de novo* del (dn-del); 3) *de novo* dup (dn-dup); 4) inherited del (inh-del); and 5) inherited dup (inh-dup). Dn-del probands had significantly lower cognitive ($p<0.001$) and adaptive ($p<0.001$) function than nonCNV probands. Cognitive ability of dn-del probands was influenced by pCNV size (-2.66 IQ points/Mb) and number of genes per pCNV (-0.43 IQ points/gene). Not surprisingly, dn-del probands did not differ from nonCNV probands in social function since the SSC cohort was ascertained by meeting full criteria for an autism diagnosis. However, we hypothesized that the presence of a pCNV may have a significant effect on a proband's Social Responsiveness Scale (SRS) raw score relative to their unaffected family members. We observed a 3.9 standard deviation (SD) shift between dn-del probands and their parents' SRS scores that was significantly higher ($p=0.033$) than the 3.4 SD shift observed in nonCNV probands. Dn-dup, inh-del, or inh-dup probands did not differ from nonCNV probands on any phenotypic measures, including familial SRS shift. These findings demonstrate that dn-del probands have a more deleterious phenotypic effect on cognitive, adaptive, and social function.

## 209

**Autism ten thousand genomes (AUT10K) project: a roadmap for the complete genetic landscape of autism spectrum disorder.** *S.W. Scherer[1,2], R.K.C. Yuen[1], H. Cao[3], X. Tong[3], D. Cao[3], Y. Sun[3], M. Li[3], W. Chen[3], X. Jin[3,4,5], J.L. Howe[1], C.R. Marshall[6], P. Szatmari[7], D. Merico[1], R.H. Ring[8].* 1) The Centre for Applied Genome, Peter Gilgan Centre for Research and Learning, Toronto, Ontario, Canada; 2) McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada; 3) BGI-Shenzhen, Bei Shan Road, Yantian, Shenzhen, China; 4) BGI@CHOP, Children's Hospital of Philadelphia, Philadelphia, USA; 5) School of Bioscience and Bioengineering, South China University of Technology, Guangzhou, China; 6) Molecular Genetics, Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Canada; 7) Centre for Addiction and Mental Health, University of Toronto, Toronto, Canada; 8) Autism Speaks, New York, USA.

Autism spectrum disorder (ASD) is a collection of neurodevelopmental conditions characterized by deficits of social interaction, communication and present of restricted and repetitive behaviors. The U.S. Centers for Disease Control and Prevention has recently reported that 1 in 68 children are diagnosed with ASD, making it one of the most common childhood disorders in the United States and worldwide. Over the past few years, large-scale genome-wide analyses (microarray and sequencing) have unveiled the important roles of *de novo* and rare inherited mutations in the etiology of ASD, which promises to enable early diagnosis and intervention. Initiated in 2012, the AUT10K project aims to establish the largest repository of ASD genomic sequence data by providing comprehensive whole genome sequence (WGS) and phenotype information of 10,000 individuals and families with ASD. Our pilot study performing WGS of 32 trios (ASD child and parents) showed that clinically-relevant genetic variants were found in ~50% of families. In a second stage, we have finished WGS of 200 ASD simplex trios with a depth of ~30X per genome using the Illumina HiSeq technology. Applying our newly developed variant detection pipeline, we found an average of 62.9 *de novo* single nucleotide variant (SNVs) and 19 *de novo* insertion/deletions (indels), data largely consistent with our previous findings. *SCN2A* remains to be the only gene with loss-of-function (LoF) mutations found in more than one family in this cohort. *De novo* LoF mutations were also detected in other known ASD-risk genes with high GC content (often difficult to assay in exome sequencing), such as *SHANK2* and *SHANK3*. *De novo* CNV detection remains to be a challenge to extract from WGS using existing tools, but we were able to find 8 putative *de novo* CNVs in coding region of the genome. We will present the advantages of WGS for resolving sequence variants residing in complex regions of the genome, leading to an improved clinical detection rate for ASD. We will also discuss our strategies on analyzing mutations beyond LoF mutations and coding regions, and share our experience with other aspects of the project such as big data storage and management.

## 210

**The identification of novel autism pathogenicity genes and their associated phenotypes.** *H.A.F. Stessman[1], B.J. O'Roak[2], E.A. Boyle[1], K.T. Witherspoon[1], B. Martin[1], C. Lee[1], L. Vives[1], C. Baker[1], J. Hiatt[1], D.A. Nickerson[1], R. Bernier[3], J. Shendure[1], E.E. Eichler[1,4].* 1) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA; 2) Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR; 3) Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA; 4) Howard Hughes Medical Institute, Seattle, WA.

Exome sequencing of families with cases of sporadic autism spectrum disorder (ASD) has suggested remarkable genetic heterogeneity with the identification of hundreds of candidate genes (>500) carrying sporadic disruptive mutations. Using modified molecular inversion probes (MIPs) and an established statistical framework for evaluating the likelihood of recurrent mutations at individual genes (O'Roak *et al.*, *Science*, 2012), we applied a genotype-first approach to identify loci with an excess of *de novo* mutation (Stessman *et al.*, *Cell*, 2014). We have now applied this approach to 99 candidate genes selected for their severity, recurrence, and strong biological network association. We increased overall capture efficiency as well as our ability to identify germline mosaicism by implementing small single-molecule tags into our MIP design (Hiatt *et al.*, *Genome Research*, 2013). In our most recent screen of 64 genes, over 3,700 probands and >2,500 unaffected siblings were scored for *de novo* mutations with >95% of the target regions uniquely captured 10 or more times. We have now identified 10 loci that show an excess ($p < 0.05$) of *de novo* mutation in ASD probands—*CHD8*, *PTEN*, *TBR1*, *GRIN2B*, *DYRK1A*, *ADNP*, *CHD2*, *SYNGAP1*, *TRIP12*, and *PAX5*—which may contribute to 1.5% of sporadic ASD. In addition, there are many genes where at least two recurrent *de novo* mutations have now been identified that have not yet reached significance, including *CTNNB1*, *SLC6A1*, *PPP2R5D*, *TCF7L2*, and *DSCAM*. Importantly, the total burden of *de novo* mutations was highly significant and skewed toward more severe events for probands in contrast to unaffected siblings. Recontact of patients with *CHD8*, *DYRK1A*, and *ADNP de novo* mutations has identified subtle phenotypic differences that suggest specific ASD subtypes associated with macrocephaly, intellectual disability, gastrointestinal dysfunction, epilepsy, and/or sleep dysfunction. This study reinforces the importance of *de novo* mutations in ASD and further highlights the strength of our genotype-first approach to identify new genetic subtypes of a highly heterogeneous disease.

## 211

**The 16p11.2 locus modulates brain structures common to autism, schizophrenia and obesity.** *S. Jacquemont[1], A.M. Maillard[1], A. Ruef[2], F. Pizzagalli[1,2], E. Migliavacca[3], L. Hippolyte[1], S. Adaszewski[2], J. Dukart[2], C. Ferrari[4], P. Conus[4], K. Männik[3], M. Zazhytska[3], V. Siffredi[1], P. Maeder[5], Z. Kutalik[6,7,8], F. Kherif[2], N. Hadjikhani[9,10], J.S. Beckmann[1,6,7], A. Reymond[3], B. Draganski[2,11], 16p11.2 European Consortium.* 1) Service of Medical Genetics, Centre hospitalier Universitaire Lausanne, Switzerland; 2) LREN- Department of clinical Neurosciences, Centre hospitalier Universitaire Lausanne, Switzerland; 3) Center for integrative Genomics, Univeristy of Lausanne, Switzerland; 4) Department of Psychiatry, CERY, Centre hospitalier Universitaire Lausanne, Switzerland; 5) Department of Radiology, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland; 6) Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland; 7) Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland; 8) Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland; 9) Brain Mind Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland Current affiliation: Guillberg Neuropsychiatry Centre, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; 10) Athinoula A. Martinos Center for Biomedical Imaging, Massachussetts General Hospital, Harvard Medical School, Charlestown, MA, USA; 11) Department of Neurology, Max-Planck Institute for Human Cognitive and Brain Science, Leipzig, Germany.

Anatomical structures and mechanisms linking genes to neuropsychiatric disorders are not deciphered. Reciprocal copy number variants at the 16p11.2 BP4-BP5 locus offer a unique opportunity to study intermediate phenotypes in carriers at high risk for autism spectrum disorder (ASD) or schizophrenia (SZ). We investigated variation in brain anatomy in 16p11.2 deletion and duplication carriers. Beyond gene dosage effects on global brain volume measures, we find robust patterns of alteration in reward, language and social cognition circuits. Changes in the reward system overlap with patterns of structural abnormalities common to ASD, SZ and obesity. Using measures of peripheral mRNA levels, we demonstrate that expression of 16p11.2 genes modulates brain structure, and confirm the findings related to the number of genomic copies. This combined molecular, neuroimaging and clinical approach, applied to larger datasets, will help interpret the relative contributions of genes to neuropsychiatric conditions by measuring their effect on local brain anatomy.

## 212

**Distinct properties of de novo mutations from whole genome sequencing of 50 patient-parent trios.** *M. Pinelli[1,2], B. Tan[3], M. van de Vorst[1], R. Leach[4], R. Klein[4], L.E.L. Vissers[1], H.G. Brunner[1,5], J.A. Veltman[1,5], A. Hoischen[1], C. Gilissen[1].* 1) Department of Human Genetics, Radboud University Medical Center, Nijmegen, Netherlands; 2) 3Dipartimento di Medicina Molecolare e Biotecnologie Mediche, Università Degli Studi di Napoli "Federico II"; 3) State Key Laboratory of Medical Genetics, Central South University, Changsha, China; 4) Complete Genomics Inc., Mountain View, CA, United States; 5) Department of Clinical Genetics, Maastricht University Medical Centre, Maastricht, Netherlands.

De novo germline mutations create the genetic variability that is the driving force of species evolution, but also the cause of sporadic genetic diseases. It is therefore of great interest to study their occurrence and associated risk factors, such as the increased parental age. Whole Genome Sequencing (WGS) of parent-offspring trios allows for the first time to observe the mutational processes in a single generation in full detail. Here we report on the rate and pattern of De Novo Mutations (DNM) based on WGS of 50 trios at high (80x median) coverage. The trios consisted of patients with severe intellectual disability and their unaffected parents (age range at day of birth 20-39 years, mothers, 21-44, fathers). We identified 2815 private DNM, corresponding to an average of 56 per trio (range 32-84). Systematic validation of coding DNM resulted in an 80% validation rate. The single factor mostly associated with the number of identified DNM was the paternal age ($p=10^{-6}$). By using the segregation of informative SNPs we determined the parental origin for 678 DNM (24%) and found a paternal/maternal ratio of 79% / 21%. Using the parental origin data we found a significant association between the paternal age and the number of paternal DNM ($p=0.004$, N= 536) as well as a suggestive association between maternal age and the number of maternal DNM ($p=0.05$, N=142). By comparison to simulated mutations, we find that DNM do not occur completely random in the genome, but have sequence signatures enriched for CpG (observed=20.6%, expected=2.6%, $p<10^{-16}$). Moreover a subset of the DNM were spatially clustered within individuals, lying within 10,000 nt proximity of eachother, (obs=1%, exp=0.0004%, $p<10^{-16}$). These shower DNM (sDNM) are depleted of CpG dinucleotides ($p<0.05$), suggesting a less prevalent role of methylation-based mutagenesis. Moreover sDNM occur in regions that are enriched for common SNP ($p<0.001$), suggesting that they occur in mutation-prone regions. We validated 7 of these showers by Sanger sequencing and determined 2 to be paternally inherited and one maternally. These results suggest single mutational events for sDNM. We are currently re-sequencing all trios to obtain haplotype resolved genomes. We believe that these results provide insight into the mutational mechanisms of de novo mutations.

## 213

**Human frontal cortex is enriched for somatic variations under physiological oxidative stress compared to the corpus callosum from same individuals.** *A. Mukhopadhyay[1,2], A. Sharma[1,2], R. Kumari[3,2], B. Mehani[1], A.H. Ansari[1], B. Varma[3], R. Rehman[4], B.K. Desiraju[4], U. Mabalirajan[4], A. Agrawal[4].* 1) Genomics & Molecular Medicine, CSIR-Institute of Genomics & Integrative Biology, Delhi, India; 2) Academy of Scientific and Innovative Research, New Delhi, India; 3) TRISUTRA Unit, CSIR-Institute of Genomics & Integrative Biology, Delhi, India; 4) Molecular Immunogenetics unit, CSIR-Institute of Genomics & Integrative Biology, Delhi, India.

Somatic variation is important in local genomic composition and outcome unexplained by Mendelian genetics. Using deep sequencing data for whole exomes we show that human brain harbors upto three folds higher proportion of somatic single base substitutions compared to other peripheral/circulatory tissue types. More than 70% of these somatic events were contributed by G:C>T:A transversion events and 48% of them were responsible for non-synonymous changes, which was twice the proportion of synonymous changes. Up to 98% of these transversions occurred in the frontal cortex (rich in neuronal cell bodies) of the brain when compared with the corpus callosum (lack of neuronal cell bodies) of the same individuals. We further show molecular evidence that these transversion events take place due to oxidative stress mediated modification of guanosine to 8-OH-dG primarily in the neuronal cells. We demonstrate significant higher amount of 8-OH-dG in the frontal cortex compared to the corpus callosum of the same individuals, which correlated with abundance of neurons in the frontal cortex. Interestingly, majority (upto 22%) of the G:C>T:A transversion events would result in Asp>Tyr or Glu>Stop changes in the frontal cortex samples (in heterozygous condition) which might provide functional advantage to the frontal cortex. A pathway based analysis of the genes where the G:C>T:A have taken place in each individual, revealed axon guidance pathway as significantly enriched implicating a favorable selection of such somatic variations. Our results provide first genetic evidence of contribution of somatic variation in normal human cortex with biological consequences.

## 214

**The Expansion of NIH's Genomic Data Sharing Policy.** *E. Luetkemeier, K. Langlais, R. Baker, C. Fomous, T. Paine, D. Paltoo.* Office of Science Policy, Office of the Director, NIH, Bethesda, MD.

Sharing research data supports the NIH mission and is essential to facilitate the translation of research results into knowledge, products, and procedures that improve human health. In 2014, NIH will expand its genomic data sharing policy to apply to all large-scale human and non-human genomic data generated from NIH-supported research, regardless of funding level or mechanism. The NIH Genomic Data Sharing (GDS) Policy, which extends the NIH Policy on Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies), will be effective in January 2015 and apply to research being funded in fiscal year 2016. The GDS Policy sets forth expectations that ensure the broad and responsible sharing of non-human and human genomic data. Its main provisions involve the responsibilities of investigators submitting data, responsibilities of investigators accessing and using data, and intellectual property. During the development of the policy, NIH sought input from a broad range of stakeholders through a public comment process. Comments were received from academic institutions, professional and scientific societies, disease and patient advocacy groups, tribal organizations, state public health agencies, health care providers, and the general public. Comments addressed the general role and value data sharing and as well as specific aspects of the draft Policy, e.g., scope and applicability, data sharing plans, timelines for data submission and release, and informed consent standards. This session will review the elements of the Policy and how they were shaped by public perspectives. Information about the implementation of the Policy will also be provided. The authors of this analysis are employed by the Office of Science Policy, Office of the Director, NIH.

## 215

**Data Sharing and dbGaP: A Survey of Practices and Opinions Among Human Geneticists.** *D. Kaufman, J. Bollinger, R. Dvoskin.* Genetics & Pub Policy Ctr, Johns Hopkins Univ, Washington, DC.

Background: Data sharing has become important in human genomic research; the NIH requires that GWAS data generated by federally funded research be deposited into the database of Genotypes and Phenotypes (dbGaP). However a number of challenges have caused frustration among researchers who wish to effectively share resources. Methods: An online survey was conducted to explore current practices and opinions related to data sharing and dbGaP. Participants were ascertained through the ASHG membership (n=159) and NCBI's lists of dbGaP data users (n=169) and contributors(n=29). Results: Three quarters of all respondents (n=357) share genetic and phenotypic data with other researchers. A total of 64% said the use of other researchers' samples and data has become more important to their work over the past five years, while 17% said sharing their own data conflicted with their career advancement. Two-thirds agreed that ensuring researchers have fair access to shared samples and data is a problem, while 57% said they thought data or samples are occasionally (44%) or frequently (13%) shared inappropriately in genetic research. Among ASHG members, 38% had contributed to dbGaP and 43% had used dbGaP. Among dbGaP users, 89% felt it had benefited their research; however, 55% said the dbGaP application process was more difficult than expected, and 27% said the data quality was lower than expected. Most (84%) believed that dbGaP contributors should follow a single standard for submitting phenotypic data. Despite these problems 84% supported requiring federally funded researchers to submit data to dbGaP. If it were up to them, though, 23% of researchers and 25% of dbGaP contributors would not submit data to dbGaP. When asked about needs and guidance related to data sharing, 80% said template consent language explaining data sharing through dbGaP would be helpful, and 74% said there is a need for guidance on the preparation of data sets for sharing. Conclusions: A genomic studies begin to require larger sample sizes, access to others' data and samples becomes critical. Issues of fair access for all researchers, and sharing while maintaining respect for and privacy of human subjects may escalate in importance. Researchers feel dbGaP can improve, but majorities support the role it plays and the requirements to deposit data, and feel that it has benefited their work. Standards for data submission may be among the most practical needs for researchers today.

## 216

**Experience with obtaining informed consent for genomic sequencing: Developing recommendations for best practices.** *B.A. Bernhardt[1], A.N. Tomlinson[1], D. Lautenbach[2], M.I. Roche[3], S.R. Scollon[4], D. Skinner[3].* 1) University of Pennsylvania, Philadelphia, PA; 2) Brigham and Women's Hospital, Boston, MA; 3) University of North Carolina, Chapel Hill, NC; 4) Baylor College of Medicine, Houston, TX.

**Background**: Whole exome and genome sequencing (WES/WGS) are offered increasingly in research and clinical settings, yet there are no guidelines for obtaining informed consent (IC) for this testing. **Methods:** To develop guidelines, we conducted semi-structured telephone interviews with 30 experienced genetic counselors and research coordinators obtaining IC for WES/ WGS in 11 NIH-funded Clinical Exploratory Sequencing Research (CSER) Consortium projects, as well as in 7 diverse clinical settings. Interviews focused on experiences with and challenges to IC; patients' common questions, concerns and misperceptions; and recommendations for IC. **Results:** Content analysis of transcribed interviews indicated that IC sessions varied between 15 and 70 minutes. Common reasons to decline testing include lack of insurance coverage (clinical settings), concerns about insurance discrimination and privacy, and feeling overwhelmed by the medical condition in the family. Nearly all opt for the return of all possible incidental findings. Patients'/participants' most common questions and concerns relate to practical details of the study or testing, the potential for insurance discrimination, data privacy protections, the implications of the results for their medical management and for relatives, and examples of conditions for which results are returned. Those obtaining IC concurred that sessions should focus on addressing areas of misperceptions and managing expectations. Specifically, participants/patients have unrealistically high expectations that they will learn a result that is diagnostic or immediately actionable and that they will personally benefit from testing. They tend to interpret negative results as excluding a genetic cause, do not anticipate uncertain results, and expect on-going analysis of stored data. With experience, most interviewees have learned to tailor the IC session content and process to patients' needs and interests, leading them to conduct sessions as an open dialogue rather than a review of the consent form. Interviewees believe that their experience with reviewing and returning results has allowed them to better prepare participants for the potential types of results that can be learned. **Conclusions:** These data suggest that the experiences across multiple centers and from diverse projects have converged on common elements of IC and guidance for the process of IC. Recommendations for IC for genomic sequencing will be presented.

## 217

**Developing a patient facing genome sequencing report: Results of Key Informant Interviews.** *J.L. Williams[1], A. Fan[1], H. Stuckey[2], D. Zallen[3], J. Green[1], M. Bonhag[1], L. Feldman[4], M. Segal[4], M.S. Williams[1].* 1) Genomic Medicine, Geisinger Health System, Danville, PA; 2) Department of Medicine, Penn State Hershey College of Medicine, Hershey, PA; 3) Department of Science and Technology in Society, Virginia Tech, Blacksburg, VA; 4) SimulConsult, Inc., Chestnut Hill, MA.

**Background:** Genome sequencing is emerging into clinical practice raising a number of potential issues for delivery systems. One example is that germline genomic data differs from other patient data in that it retains relevance for the individual's health over the entire lifespan. Given the current fragmented state of the delivery system this raises the question of how this information can be made available wherever the patient is receiving care. Indeed the individual is the only constant in the changing delivery landscape. This has led to exploration of care centered around the patient and their caregivers. The purpose of this research is to develop a patient facing genomic laboratory report with advanced functionality including point of care education and clinical decision support. Development will use providers and parents of affected patients to provide feedback on the desired elements for the provider and patient views and the usability of the report. **Methods:** A draft patient report was developed by the research team. The team includes a patient investigator and several experts in patient engagement and communication. The draft report was then presented to participants in a clinical research project exploring the use of genome sequencing for undiagnosed cognitive disability. Semi-structured interviews were used to elicit prior experience with genetic test result communication and feedback about the draft report. Interviews were transcribed and analyzed using the conceptual framework of existential phenomenology which favors the interpretation of meaning through subjective experiences. **Results:** Participants have endorsed the importance of having a report created for patients and family. In particular they noted that this allowed reading and re-reading of the report and to have a record of what was discussed. The draft report was found to be informative and written at an appropriate level. Different diagrams were presented and were judged to be helpful to understanding the report content. A consistent deficiency in the draft report from the participant perspective was a section on what to expect for the future. **Conclusions:** Participants value a report created for them. The results of the interviews will inform creation of a report that will be evaluated by a larger number of participants using focus groups. The final patient report will be used in the comparative effectiveness portion of the project.

## 218

**Use of My46 to return individual research results to families of children with Joubert syndrome.** *S.M. Jamal[1], A.G. Shankar[1], J. Dempsey[1], C. Isabella[1], J.H. Yu[1], J. Crouch[2], T.M. Harrell[1], M.J. Bamshad[1,3], D. Doherty[1], H.K. Tabor[1,2].* 1) Department of Pediatrics, University of Washington, Seattle, WA; 2) Treuman Katz Center for Pediatric Bioethics, Seattle Children's Research Institute, Seattle, WA; 3) Department of Genome Sciences, University of Washington, Seattle, WA.

The application of exome and whole genome sequencing in research has led to the increased identification of clinically important primary findings to be returned to participants and families. Yet, the requisite time, labor and expertise to recontact participants and return results remain important barriers. We offered primary research results to 47 families whose children were enrolled in a genetic research study of Joubert syndrome. Results were offered using My46 (http://www.my46.org), a web-based tool for return of genetic/genomic results that enables individuals to set their preferences for receiving results, read expert-curated, standardized information about conditions, and view results privately and at their convenience. Usability of, and satisfaction with, My46 were assessed using an online survey an average of 25 days after result receipt. Forty-three parents (91%) from 7 countries viewed their child(ren)'s result and most (68%) completed the post-results survey. Four parents (9%) did not view their child's result, despite multiple prompts and offers of assistance. The median time from notification of result availability to result viewing was 6 days. Most (72%) parents were receiving genetic results for their child for the first time. Almost all (91%) indicated satisfaction with receiving results and did not express regret with their decision to receive results. Many (41%) had already shared and 50% intended to share the result with their healthcare provider. Three parents reported feeling upset by the result, yet each reported learning useful health information and none were dissatisfied with the result return method. Most (78%) were satisfied receiving the result via My46 and only 2 (6%) said that they did not receive enough information to understand the result. Five parents (16%) contacted the genetic counselor for clarification of result interpretation and next steps. The majority (84%) of parents would use My46 again, and 81% would recommend it to someone else as a way to receive genetic research results. The mean scores for a modified Computer System Usability Questionnaire were positive, on a scale of 1 to 5 (5 being most favorable): usefulness=3.55, information quality=3.50, and interface quality=3.43, with an overall score of 3.52. These results demonstrate that My46 can effectively and efficiently return genetic research results with high usability and satisfaction, even for families receiving genetic test results for the first time.

## 219

**Patient Preferences for the Return of Individual Research Results Derived from Pediatric Biobank Samples.** *S. Savage[1], K. Christensen[2], N. Huntington[3,4], E. Weitzman[4,5,6], S. Ziniel[4,5,7], P. Bacon[8], C. Cacioppo[1], R. Green[2,9], I. Holm[1,4,10].* 1) Division of Genetics and Genomics, Children's Hospital Boston, Boston, MA; 2) Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School; 3) Division of Developmental Medicine, Boston Children's Hospital; 4) Department of Pediatrics, Harvard Medical School; 5) Division of Adolescent/Young Adult Medicine, Boston Children's Hospital; 6) Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Boston Children's Hospital; 7) Center for Patient Safety and Quality Research, Program for Patient Safety and Quality, Boston Children's Hospital; 8) Johns Hopkins University School of Medicine; 9) Partners HealthCare Center for Personalized Genetic Medicine; 10) Manton Center for Orphan Disease Research, Boston Children's Hospital.

Background: The discussion about handling results identified through research on biobank samples has included calls to incorporate participants' preferences when deciding what individual research results (IRRs) to disclose. However, little is known about participants' ability to establish preferences or how preference-setting would affect satisfaction with IRR disclosure. Methods: We invited parents of Boston Children's Hospital patients to complete a web-based survey exploring the impact of various policies for disclosing IRRs identified in biobanked samples from children. After an educational video, participants were randomized to 1 of 4 hypothetical scenarios for addressing IRRs, including a 'granular arm' where participants set preferences based on the preventability and severity of conditions. Additional choices allowed omission of IRRs about mental health, developmental, childhood degenerative, and adult-onset disorders. Participants viewed potential IRRs and rated their satisfaction of the content and process. Participants were then given the option to reset preferences. Results: Of 11,394 invited parents, 2,718 (23.9%) participated in the survey, including 1,026 randomized into the granular arm (mean age 44, 91% female, 86% white). Initially 62% wanted all 8 categories of IRR, 14% omitted only adult-onset, non-preventable, childhood degenerative, or mental health conditions, and 2% wanted no IRRs (mean # of categories=7.0). After viewing potential IRRs, 14% revised preferences, adding more categories on average (mean # of categories: 5.5 before reset vs. 6.5 after, $p<0.001$). Mean process and results satisfaction prior to the offer to reset preferences were 7.0 and 6.6, respectively, on 0-10 scales, with scores increasing among those who changed preferences (process: 5.1 before reset vs. 7.2 after, $p<0.001$; results: 4.7 before reset vs. 7.2 after, $p<0.001$). Conclusions: While the majority of participants wanted all potential IRRs and few wanted none, about 1/3 expressed more nuanced preferences. Many changed preferences after viewing potential IRRs, with a tendency to want more IRRs. Satisfaction was high for both the process and potential content of IRR return, with large improvements in satisfaction among those who reset preferences. Findings suggest that a high percentage of research participants would take advantage of an ability to set preferences, but that preferences may be difficult to establish without seeing examples of potential IRRs.

## 220

**Weapons, boxes, and credit reports: Metaphorical language in discussions of receiving exome and whole genome sequencing results.** *S.C. Nelson[1], J. Crouch[2], M.J. Bamshad[3,4], H.K. Tabor[2,3], J. Yu[3].* 1) Institute for Public Health Genetics, University of Washington, Seattle, WA; 2) Seattle Children's Research Institute, Seattle, WA; 3) Department of Pediatrics, University of Washington, Seattle, WA; 4) Department of Genome Sciences, University of Washington, Seattle, WA.

The rapid integration of genomics into clinical care, growing interest in offering genetic results to research participants, and consumer enthusiasm for personalized genomics requires greater understanding of how individuals conceptualize and communicate about genetic information, preferences for genetic information, and individual genetic results. At stake are the potential benefits and harms of accurate communication versus miscommunication. Metaphors such as the "genome as a blueprint" have been well studied in public discourse as a vehicle for conceptualizing genetics; however, the applicability of these metaphors to the translation of personal genetic information is largely unknown. We performed a qualitative analysis of metaphorical language in 40 interviews and 13 focus groups in which participants (n=109) were asked to discuss their preferences for and expectations about receiving genetic results from whole genome or exome sequencing. We identified several salient conceptual metaphors in which participants compared genetic information to physical objects such as tools, weapons, and goods in boxes. Dichotomies and tensions within these metaphors suggest the importance of agency and different loci of control. In some cases, genetic information empowered the individual to act (e.g., a weapon in their arsenal) while in other cases it overpowered them (e.g., they are bombarded by genetic information). Participants used metaphors such as storing results in a "lockbox" or unintentionally opening "Pandora's box" to describe the potential benefits and harms of incidental sequencing results. Metaphors comparing whole genome sequences to formal documents or reports (e.g., credit report) raised questions of authorship, ownership, and interpretability, and resonated with well-known genetics metaphors in the public domain (e.g., blueprint, recipe, book of life). These results have practical implications for understanding and perhaps influencing how research participants and patients perceive the desirability, utility, and actionability of genetic information. Our findings suggest that increased awareness of and attention to metaphorical language may serve to improve communication between stakeholders when discussing genetic information.

## 221

**Clinical Integration of Next Generation Sequencing: A Policy Analysis.** *A.L. McGuire[1], D.J. Kaufman[2], G.H Javitt[3], P.A. Deverka[4], D. Messner[4], R. Cook-Deegan[5], M.A. Curnutte[1], J. Bollinger[2], R. Dvoskin[2], S. Chandrasekharan[6], B.J. Evans[7].* 1) Ctr Med Ethics, Baylor Col Med, Houston, TX; 2) Genetics and Public Policy Center, Johns Hopkins University, Washington DC; 3) Berman Institute of Ethics, Johns Hopkins University, Baltimore MD; 4) Center for Medical Technology Policy, Baltimore, MD; 5) Center for Public Genomics and Sanford School of Public Policy, Duke University, Durham, NC; 6) Center for Public Genomics and Duke Global Health Institute, Duke University, Durham, NC; 7) The University of Houston Law Center, Houston, TX.

Introduction: Clinical next generation sequencing (NGS) technologies are challenging existing regulatory, reimbursement and intellectual property paradigms. The volume and breadth of data generated, the bioinformatics requirements for clinical interpretation, and the need to accurately interpret and communicate meaningful information to patients present a host of policy challenges for clinical integration and reimbursement of NGS testing. These are not areas that fit easily within existing policy frameworks . Methods: We recruited 60 stakeholders representing policy makers, payers, research funders, researchers, legal scholars, clinicians, industry representatives and patient advocates to participate in a 4-round Delphi process to both identify and prioritize key policy issues. We also conducted semi-structured interviews with 19industry leaders representing these stakeholder groups and performed an extensive literature and case law review. We prepared four white papers that summarize the major regulatory, reimbursement, intellectual property, and data access issues confronting clinical NGS as background material for the Delphi. Results: We will discuss major issues related to four key policy domains, including quality assurance, insurance coverage, intellectual property management, and data access, that must be addressed to ensure high quality clinical NGS. Findings from the Delphi process will be presented to show how individuals within these different groups understand and prioritize the major policy issues that have been identified. Discussion: New approaches will be needed to establish an oversight and incentive system that promotes appropriate, broad access to high-quality sequence data and valid reports while encouraging innovation. We advocate a coordinated policy approach, which first requires a comprehensive understanding of the existing regulatory and legal structures.

## 222

**Discovery and *in vivo* experimental validation of a novel, non-meiotic pathway governing production of spermatozoa and oocytes in human.** *A.S. Lee[1], N. Huang[1], Y. Yin[2], R.A. Hess[3], L. Ma[2], P.N. Schlegel[4], A.M. Lopes[5], D.T. Carrell[6], Z. Hu[7], D.F. Conrad[1].* 1) Department of Genetics, Washington University School of Medicine, St. Louis, MO; 2) Department of Dermatology, Washington University School of Medicine, St. Louis, MO; 3) Department of Comparative Biosciences, College of Veterinary Medicine, University of Illinois at Urbana-Champaign, Urbana, IL; 4) Department of Urology, Weill Cornell Medical College, New York-Presbyterian Hospital, New York, NY; 5) Institute of Molecular Pathology and Immunology, University of Porto, Portugal; 6) Department of Physiology, University of Utah School of Medicine, Salt Lake City, UT; 7) Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, China.

Here we report the first genome-wide association study (GWAS) for human gonadal function across males and females, and experimental validation of an exceptional result using mouse models. Although male and female infertility are typically researched as distinct diseases, some of the molecular machinery for gametogenesis is shared between both sexes. To search for a common genetic basis for sperm and egg production, we generated genome-wide CNV calls across 18,931 females (841 cases of primary ovarian insufficiency vs. 18,090 controls) and 5,197 males (1,705 cases of spermatogenic impairment vs. 3,492 controls), and used the resulting callsets to perform rare variant GWAS separately in males and females. Remarkably, the strongest association signal in both sexes was found at the same locus (*CSMD1*). *CSMD1* is located on 8p23, within the region of highest sequence divergence between human and chimpanzee outside the Y chromosome (on average 0.03 substitutions per site across 1Mb).

*CSMD1* is expressed in both the testes and ovaries and shares primary sequence homology with complement-interacting proteins, but little else is known of its biological function. Thus we turned to the mouse model for an in vivo reproductive model. We took two parallel approaches to perturb *Csmd1* in the mouse: (i) we delivered an shRNA targeting *Csmd1* directly into the testes via lentiviral vector, and (ii) we generated a colony of *Csmd1* gene knockout mice. We then performed extensive cellular and molecular phenotyping of the gonads, including histopathology; computer-assisted semen analysis; purification of germ cells via FACS; and RNAseq of whole testis, whole ovary, and FACS-purified male germ cells in wildtype and knockout.

*Csmd1*-deficient males show near-complete histological degeneration of the testes as early as 34 days of age (approx. onset of sexual maturity), with concomitant reduction in sperm count and motility. Remarkably, in both the post-meiotic (haploid) male germ cells *and* female ovaries, *Csmd1*-knockouts show deranged expression of ciliary and flagellar components of the axoneme (GO:0005930, GO:0044447, GO:0035082) which are typically associated with sperm motility. Taken together, these results suggest that at least a subset of overlapping processes govern both male and female germ cell development, and that *CSMD1* plays a crucial role in their maintenance.

## 223

**Complex dynamics of meiotic recombination initiation in laboratory mouse strains.** *K. Brick[1], F. Smagulova[2], R.D. Camerini-Otero[1], G. Petukhova[2].* 1) Genetics & Biochemistry Branch, NIDDK, National Institutes of Health, Bethesda, MD, USA; 2) Deparment of Biochemistry and Molecular Biology, Uniformed Services University of Health Sciences, Bethesda, MD, USA.

In mouse and in other mammals, the DNA double strand breaks (DSBs) that initiate meiotic recombination are directed to a subset of genomic loci called hotspots by the PRDM9 protein. The C-terminal zinc finger (ZnF) array of PRDM9 is responsible for directing sequence specific DNA binding of PRDM9, yet this ZnF-array is highly polymorphic with over 100 alleles described in mouse. We have previously developed the first method to directly map the sites of meiotic DSBs genome-wide and here, we exploit this technique to categorize the recombination initiation landscape in six common laboratory mouse strains with different PRDM9 alleles and in their F1 hybrids.

We find that DSB hotspots defined by the six different alleles of the PRDM9 protein occur mostly at mutually exclusive loci. Furthermore, even very similar PRDM9 alleles that differ by only a single zinc finger (C57Bl/6J and C3H/HeJ) define mostly different hotspots. In addition to clear dominance of some PRDM9 alleles over others, a striking finding in F1 mice is that of the appearance of "novel" DSB hotspots, not present in either parental genome. In some cases, such sites constitute >30% of all hotspots. By analyzing the DNA sequence pulled down from DMC1 ChIP-Seq in B6xCAST F1 mice we found that the vast majority of novel hotspots are formed by the PRDM9 allele from one parental strain on the chromosome of the other parental strain (non-self chromosome). A straightforward explanation is that novel hotspots arise due to the presence of new PRDM9 binding sites on the non-self chromosome, however we can only find evidence for this at ~50% of novel hotspots. Biased DSB initiation on the non-self chromosome is also observed at other hotspots and results in hotspots that are relatively stronger than in the parental genome. We will also discuss curiously frequent instances of hotspots that coincide with the sites where DSBs form in the absence of PRDM9 and will provide evidence for broad domains of DSB formation across mouse strains that agree with crossover data. Together, these findings elucidate some of the complex dynamics of DSB formation in mammalian meiosis.

## 224

**Bringing homologs together: Sex- and species-specific differences in synapsis.** *J. Gruhn[1], C. Rubio[2], P.A. Hunt[1], T. Hassold[1].* 1) School of Molecular Biosciences, Washington State University, Pullman, WA; 2) Instituto Valenciano de Infertilidad, Valencia, Spain.

Over the past decade, considerable attention has focused on how homologous chromosomes synapse in meiosis and whether this process is linked to crossover site designation. However, very little information is available on mammals and it is not known whether synapsis varies among species or between sexes. Accordingly, we initiated cytological analyses of synapsis in human spermatocytes and oocytes and, for comparison, in male and female mice. Our analyses of humans indicate surprising sex-specific differences. In the human male, there is typically a single distal synaptic initiation site (SIS) per chromosome arm and initiation at the centromere is never observed. In contrast, human females typically have multiple, interstitially located SISs that typically include the centromere; indeed, in many instances the centromere appears to be the first region to synapse. Interestingly, in mice, we observed few male:female differences, but some similarities with humans: e.g, like the human female, both male and female mice typically have multiple SISs per chromosome but, as in human males, centromeres appear to be refractory to synapsis. Our results provide evidence for at least three different "strategies" to synapsis among mammals - one involving the union of sequences in distal chromosome segments (human males); one involving centromeric interactions (human females), and one involving multiple points of synaptic initiation in non-centromeric locations (both sexes in mice). After identifying such diverse mechanisms for homolog synapsis, we were interested in determining if these differences had an impact on sex- and species-specific variation in crossovers (COs). Thus, we asked whether there was a correlation between SICs and COs, as has been reported for other non-mammalian species. Comparative mapping studies of SISs and COs failed to demonstrate an exact match in males or females of either humans or mice; therefore, the way in which COs are chosen appears to be different in mammalian species than in non-mammalian model organisms. Taken together, these studies demonstrate remarkable sex and species-specific differences in synaptic initiation, but these differences do not directly control variation in downstream recombination events in either humans or mice.

## 225

**Targeted resequencing identifies mutant selfish clones within the testis and unifies the concepts of somatic and germline mutation.** *G.J. Maher, E. Giannoulatou, S.J. McGowan, A. Goriely, A.O.M. Wilkie.* Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK.

**Introduction:** The selection and clonal expansion of pathogenic somatic mutations is associated with diseases such as cancer and tissue overgrowth; however such mutations are typically not inherited. In the special context of the testis, we and others have presented evidence that a similar process leads to elevated levels of specific mutations in sperm (up to $10^4$-fold higher than the background rate), a process that we term *selfish spermatogonial selection*. This mechanism is responsible for the high apparent birth rate of several spontaneous congenital disorders, and may have a wider role in cancer predisposition and neuropsychiatric conditions. Until now, however, it has not been possible to identify directly these mutant clonal expansion events in the cellular architecture of normal human testes. **Methods and Results:** In sections of formalin-fixed paraffin embedded (FFPE) normal testes from elderly men (aged 62-79 yr), we identified a subset of seminiferous tubules with a morphology and antigenic profile suggestive of mutant clonal expansion. Laser capture microdissection, whole genome amplification (WGA) and targeted resequencing (>100 candidate genes using Haloplex technology) of tubules from 8 testes identified previously defined selfish mutations encoding activating substitutions, including FGFR2 Y340C (corresponding to Pfeiffer syndrome when occurring as a germline mutation) and C342S (Crouzon syndrome), FGFR3 Y373C (thanatophoric dysplasia (TD) type I) and HRAS G13R (likely germline lethal). A subsequent targeted screen of 10 mutation hotspots in 6 genes using Ion Torrent PGM sequencing identified further tubules with mutations in FGFR3 (K650E (TD type II)) and HRAS (G13R). The mutations were specifically localised to the tubules with abnormal antigenic properties; surrounding normal tubules were mutation-negative. On histological examination, the mutant tubules had variably reduced spermatogenesis. To exclude confounding technical artefacts, all mutations were verified by Sanger sequencing of non-WGA material. **Conclusions:** For three genes (*FGFR2*, *FGFR3* and *HRAS*) our results demonstrate for the first time the occurrence of selfish spermatogonial selection at a cellular level in the testis. This opens up a new approach to analysing the germline profile of mutations in the testis and the signalling pathways affected, which will have implications for understanding the paternal age effect and the many diseases fuelled by a high rate of new mutations.

## 226

**Prevalence of pathogenic copy number variants (CNVs) for specific ultrasound detected structural abnormalities using prenatal chromosomal microarray (CMA) in a multi-center cohort.** *T. Leung[1], O. Chan[1], SW. Cheung[2], Y. Kwok[1], KW. Choy[1].* 1) Obstetrics & Gynaecology, The Chinese University of Hong Kong, Hong Kong, Hong Kong, Hong Kong; 2) Department of Molecular and Human Genetics, Baylor College of Medicine.

A multi-center study of 491 pregnancies with structural ultrasound abnormalities using prenatal CMA (customized 44K Fetal Chip V1.0) was conducted to determine the prevalence of pathogenic CNVs. Pregnancies with isolated abnormality were categorized into the specific system. Pregnancies with abnormality in more than one system were grouped under multiple abnormalities. Fetuses with abnormalities of soft markers only were not included in this cohort. The prevalence of pathogenic CNVs was analyzed for each category. Results: The overall prevalence of pathogenic CNVs in pregnancies with structural ultrasound abnormalities was 11.8%. Abnormal CNVs were identified in 30.9% cases with multiple anomalies, 11.5% cases with isolated CNS anomalies, 6.9% cases with isolated CVS anomalies and 11.5% cases with isolated intrauterine growth retardation (IUGR). Pathogenic CNVs were not identified among pregnancies with isolated facial, gastrointestinal or respiratory anomalies. CMA testing was supplemented by karyotype in 84.6% of pregnancies. Among pregnancies with a normal karyotype, abnormal CNVs were identified in 17.5% cases with multiple anomalies, 3.7% cases with isolated skeletal anomalies, 2.1% cases with isolated CNS anomalies, 2% cases with isolated CNS anomalies and 11.1% cases with isolated IUGR,. CMA did not detect any abnormal CNVs for pregnancies that had a normal karyotype and affected by other isolated structural abnormalities including facial, gastrointestinal, respiratory and urogenital system. Conclusion: Pathogenic CNVs were identified in 17.5% pregnancies with multiple structural abnormalities that had a normal karyotype. Pregnancies with isolated structural anomaly were only associated with 0-3.7% risk of pathogenic CNVs after a normal karyotype. Prenatal CMA should be recommended as first tier testing for pregnancies with multiple structural abnormalities. However, the additional benefit of CMA in pregnancies with isolated anomaly was questionable.

## 227

**Comprehensive genetic analysis of pregnancy loss by chromosomal microarrays: outcomes, benefits and challenges.** *T. Sahoo, M. Strecker, A. Mehta, N. Dzidic, R.W. Tyson, K. Hovanes.* Combimatrix, Irvine, CA.

INTRODUCTION: Chromosomal aneuploidies, polyploidies and unbalanced structural rearrangements account for >50% of all first trimester pregnancy losses. Determining the cause of pregnancy loss can assist in providing appropriate medical management and recurrence risk counseling. However, the limited success of classical cytogenetic analysis of products of conception (POC) has led to implementation of alternative methods of evaluation; in particular, chromosomal microarray analysis (CMA). Utilizing genomic DNA, CMA allows for direct cytogenomic analysis of both fresh POC tissue and archived formalin-fixed, paraffin-embedded (FFPE) samples. METHODS: At CombiMatrix we undertook a comprehensive, multi-year analysis of over 3400 POC specimens analyzed by various types of CMA: BAC-aCGH (N=678), oligo-aCGH (N=754), and single nucleotide polymorphism (SNP) array (N=1976). RESULTS: Of the 3408 consecutive specimens (fresh POC: 2712, FFPE: 564, other: 132) referred for analysis, 2345/2712 (86.5%) of fresh POC samples and 492/564 (87.2%) of FFPE samples provided a successful result. Common reasons for test failure included insufficient sample size and poor DNA quality. Of the 2930 samples resulted, 1533 (52%) were abnormal, 1330 (46%) were normal, and 67 (2%) revealed a variant of uncertain significance. As expected, CMA identified a broad spectrum of abnormalities: 1056 single or multiple trisomies, 187 triploidies, 161 monosomy X, and 129 with other gross genomic imbalances. Of the three platforms, the SNP array had the highest abnormality detection rate (53.8%), and in addition to chromosomal imbalances, also identified cases with whole genome or multiple regions of allelic homozygosity. CONCLUSIONS: The exacting nature of POC specimens precludes successful culture and cytogenetic analysis in 20-40% of cases, which underscores the value of CMA in successfully enabling analysis of a great majority of these specimens. In particular, SNP-based CMA allows for the identification of polyploidy, aneuploidy, allelic homozygosity, genomic imbalances, and maternal cell contamination in both fresh and FFPE tissue samples. This comprehensive study confirms and significantly extends the value of CMA in POC analysis, and highlights the spectrum of abnormalities identifiable using genome-wide, high resolution CMA. This expands our understanding of the causes, clarifies the risk of future loss(s) and helps counsel families on the need for potential genetic counseling.

## 228

**Genomic Augmentation of Newborn Screening.** *B. Solomon, D. Bodian, R. Iyer, K. Huddleston, R. Hastak, A. Chu, A. Black, G. Eley, J. Vockley, J. Niederhuber.* Division of Medical Genomics, Inova Translational Medicine Institute, Falls Church, VA.

Newborn Screening (NBS) aims to efficiently and cost-effectively identify neonates with treatable diseases. By design, NBS yields relatively frequent false positives, often requiring repeat NBS and further work-up. We aimed to generate preliminary data that NBS can be augmented by performing parallel DNA sequencing. These data were generated through our trio-based (parents+ newborn), IRB-approved whole-genome sequencing (WGS) studies of early childhood health, which have generated > 6,000 whole genomes on comprehensively phenotyped patients in ~3 years. We conducted pilot NBS analyses on the first 702 newborns enrolled in our preterm birth study, including both full-term (FT) and preterm (PT) neonates. We bioinformatically analyzed variants in 127 selected genes corresponding to all disorders currently/planned to be included in blood-based NBS, and compared results to standard NBS. Among these infants, a total of 966 standard NBS were performed, with 117 infants (17%) receiving ≥1 abnormal/invalid result. PT infants were 8x more likely than FT to receive an abnormal/invalid result from standard NBS (39% vs 4.7%, p<0.0001). WGS analysis revealed 2,216 distinct variants in the 127 genes, 229 annotated in HGMD as disease mutations and 128 as pathogenic in ClinVar. Other than hemoglobinopathies, no infants with abnormal standard NBS had a confirmed NBS disorder, concurrent with WGS results. WGS analysis ascertained individuals with glucose-6-phosphate dehydrogenase deficiency, a disorder not included in NBS in the state where the studies take place. As an example of WGS analysis helping rule-out false positives, 4 infants had a positive initial NBS for cystic fibrosis, though none received a confirmatory diagnosis. WGS detected 56 distinct, heterozygous CFTR variants (30 HGMD disease mutations) but no bi-allelic mutations. Since many NBS conditions are rare, we focused on sickle cell anemia to test the ability of WGS to reproduce positive standard NBS, and found complete agreement between standard NBS and WGS. We are now conducting blinded sequencing for known positives from >50 NBS conditions, as well as known negatives and carriers. We have also designed several custom platforms to perform parallel sequencing (in real-time) with standard NBS, and can easily add additional genes of interest. Among other ancillary benefits, these studies are enabling the construction of a unique genomic resource related to WGS-detected NBS variants.

## 229

***CETN1* variations cause idiopathic male infertility.** *D.V.S. SUDHAKAR[1], A. KHATTRI[1], R. PHANINDRANATH[1], A.K. SHARMA[1], J. RESHMA DEVI[1], M. DEENADAYAL[2], N.J. GUPTA[3], S. PRASAD[4], S. YOGENDRA[1], K. THANGARAJ[1].* 1) Centre for Cellular and Molecular Biology, Hyderabad, India; 2) Infertility Institute & Research Centre, Secunderabad, India; 3) Institute of Reproductive Medicine, Salt Lake, Kolkata, India; 4) Sridevi Nursing Home, Warasiguda, Hyderabad, India.

Centrins are calmodulin-like, EF-hand containing calcium-binding proteins that are found in all eukaryotic cells from yeast to mammals. Centrins are encoded by 3 homologous genes (*CETN1*, *CETN2, CETN3*) in humans. The expression of centrin-1 (*CETN1*) is testis-specific, spermatogenic cell-specific and developmental stage-related. Our previous studies on several Y-chromosome, autosomal and mitochondrial genes revealed 25% of genetic causes responsible for male infertility. Therefore, our aim is to identify additional genetic factors that are associated with male infertility. We collected testicular biopsy from 6 obstructive and 5 Non-obstructive Azoospermic (NOA) men, isolated mRNA and performed gene expression analysis using Affymetrix array (2.0). We found several genes were differentially expressed, of which *CETN1* was one among the highly down-regulated genes in NOA individuals. Moreover, recent studies have shown that *Cetn1* (-/-) male mice were infertile. Hence, we selected this gene for further studies. We have sequenced the coding region of *CETN1* in ethnically matched 656 infertile and 347 fertile Indian men and found five nucleotide substitutions in the entire *CETN1* gene; of which two were 5'UTR variants (rs_367716858, novel), one each was synonymous (rs114739741), non-synonymous (rs61734344) and 3'UTR variant (rs568365). Non-synonymous nucleotide substitution (rs61734344) was found to be strongly associated with male infertility ($P^{Corr}$<0.0005), replacing Methionine with Threonine (p.Met72Thr) in highly conserved region. Functional characterization of the above non-synonymous mutation (p.Met72Thr) was carried out using (overexpressed) wild type and mutant proteins. Biophysical studies revealed that the mutant protein (p.Met72Thr) is less structured and has considerable differences in the ion binding thermodynamics, compared to wild type protein. Since the mutant protein's spectroscopic and thermodynamic properties are hampered, we predicted that the mutation could probably lead to the compromised physiological functions of Centrin1, which is a calcium sensor. In addition to the substitutions, we found a novel seven base pair indel (g.581202_209) in the 3'UTR region, which is also associated ($P^{Corr}$=0.033) with male infertility. This is the first study on *CETN1* gene mutations and their association with human male infertility.

## 230

**Mutations in RPL17 expand the molecular basis of Diamond-Blackfan anemia and guide insights into unique biochemical signatures underscoring ribosomopathies.** *E.E. Davis[1], D.W. Reid[2], J. Liang[3], J.R. Willer[1], L. Fievet[1], Z.A. Bhuiyan[4], A.L. Wall[1], J.S. Beckmann[5], N. Katsanis[1], C.V. Nicchitta[2,6], F. Fellmann[4].* 1) Center for Human Disease Modeling, Duke University Medical Center, Durham, NC, USA; 2) Department of Biochemistry, Duke University Medical Center, Durham, NC, USA; 3) BGI-Shenzhen, Shenzhen, China; 4) Service de Génétique Médicale, CHUV, Lausanne, Switzerland; 5) Swiss Institute of Bioinformatics, Lausanne, Switzerland; 6) Department of Cell Biology, Duke University Medical Center, Durham, NC, USA.

Diamond-Blackfan anemia (DBA) is a rare, clinically heterogeneous disorder hallmarked by red blood cell aplasia and incompletely penetrant defects in facio-skeletal development. Associated typically with loss-of-function mutations in at least ten ribosomal components, the extensive genetic heterogeneity poses significant molecular challenges. However, the biochemically tractable organellar basis of DBA, a clinical entity within the ribosomopathies, offers the unique opportunity to investigate the mechanisms underlying disease pathology. Here, we report the genetic, functional, and biochemical dissection of a multigenerational Swiss family with anemia, neutropenia and variable craniofacial and limb abnormalities segregating under a dominant inheritance paradigm. We conducted whole exome sequencing in three affected individuals and identified a novel single nucleotide change in a splice acceptor region of the 60S ribosomal protein L17 encoding gene, RPL17; this variant segregated with all seven affected family members, and mRNA splicing studies showed that the mutation resulted in an in-frame deletion of exon 5. To investigate the physiological relevance and pathogenicity of this variant, we used in vivo complementation studies in zebrafish. First, rpl17 suppression results in micrognathia and anemia phenotypes in developing zebrafish embryos that are orthologous to those observed in patients. Subsequently, we used these relevant phenotypic readouts in zebrafish to determine that the deletion of exon 5 is pathogenic. To explore the mechanistic basis for the ribosomal dysfunction in affected individuals in this pedigree, and to elucidate further the precise roles of RPL17 in the large 60S ribosomal subunit, we conducted a series of biochemical assays using patient-derived cell lines and rpl17 zebrafish morphants. Whereas ribosome maturation was not significantly altered in mutants versus controls, ribosome profiling studies demonstrated reductions in the translation of mRNAs encoding proteins functioning in key developmental pathways relevant to craniofacial development and red cell generation and homeostasis. Strikingly, ribosome profiling of both mutant cells and morphant embryos revealed a distinct translation profile consistent with selective elongational pausing. Together, these studies highlight the power of a multidisciplinary approach to inform both disease architecture and pathogenesis, and also the molecular mechanisms of ribosome function.

## 231

**Digenic inheritance in Alport syndrome.** *M. Mencarelli[1,2], L. Heidet[3], H. Storey[4], M. van Geel[5], B. Knebelmann[3], C. Fallerini[1], L. Dosa[1,2], N. Miglietti[6], M.F. Antonucci[1,2], F. Cetta[7], A. van den Wijngaard[5], S. Yau[4], F. Mari[1,2], M. Bruttini[1,2], F. Ariani[1,2], K. Dahan[8], B. Smeets[5], C. Antignac[3,9,10], F. Flinter[11], A. Renieri[1,2].* 1) Medical Genetics, University of Siena, Siena, Italy; 2) Genetica Medica, Azienda Ospedaliera Universitaria Senese, Siena, Italy; 3) Centre de Référence des Maladies Rénales Héréditaires de l'Enfant et de l'Adulte (MARHEA), Service de Néphrologie Pédiatrique, Hôpital Necker-Enfants Malades, Paris, France; 4) Molecular Genetics Laboratory, Viapath, 5th Floor Tower Wing, Guy's Hospital, London SE1 9RT, England; 5) Clinical Genetics, Maastricht University Medical Centre, The Netherlands; 6) Clinica Pediatrica, Azienda Ospedaliera Spedali Civili, Brescia, Italy; 7) IRCCS MultiMedica, Milan, Italy; 8) Université Catholique de Louvain, Belgium; 9) Inserm UMR 1163, Laboratory of Inherited Kidney Diseases, Paris, France; 10) Paris Descartes-Sorbonne Paris Cité Université, Imagine Institute, Paris, France; 11) Department of Clinical Genetics, Guy's & St Thomas' NHS Foundation Trust, Guy's Hospital, London SE1 9RT, England.

Alport syndrome (AS) is a clinically and genetically heterogeneous inherited progressive nephropathy associated with deafness and characteristic ocular lesions. Monogenic inheritance models are well known, with semidominant X-linked inheritance, linked to COL4A5 gene, or autosomal dominant or recessive inheritance, linked to COL4A3 or COL4A4 gene. The increased availability of massive parallel sequencing permits identification of families with pathogenic mutations in more than one disease-gene, suggesting digenic inheritance models. We present a series of 11 families, in which individuals with 2 pathogenic mutations in different Alport associated genes are more severely affected than heterozygotes. The double heterozygotes mean age of renal function deterioration is intermediate with respect to the autosomal dominant form and the autosomal recessive one, in line with molecule stoichiometry of the disruption of the triple helix of the type IV collagen alpha chains building the basement membrane of the glomeruli. Furthermore, segregation analysis indicates three possible segregation models: i) digenic autosomal inheritance with linked mutations in trans mimicking a recessive disease (5 families); ii) autosomal inheritance with linked mutations in cis mimicking a dominant disease (2 families); iii) unlinked autosomal and X-linked inheritance having its own peculiar segregation (4 families including one with triallelic inheritance). In summary, in this work we provide evidence for the first time of the digenic inheritance of Alport syndrome. Our results are of interest scientifically and have implications for genetic counselling. Clinical geneticists should be made aware of the digenic model of inheritance in Alport syndrome in order to assess the exact recurrence risk and give the correct prognosis.

## 232

***PCBD1* and diabetes: a novel player with direct implications for therapy.** *D. Simaite[1, 2], J. Kofent[1], M. Gong[1,2], F. Rüschendorf[1], S. Jia[1], P. Arn[3], K. Bentler[4], C. Ellaway[5], P. Kühnen[6], G.F. Hoffmann[7], N. Blau[7,8], F.M. Spagnoli[1], N. Hübner[1], K. Raile[2,6].* 1) Max Delbrück Center for Molecular Medicine, Berlin, Germany; 2) Experimental and Clinical Research Center, Berlin, Germany; 3) Nemours Children's Clinic, Jacksonville, FL; 4) University of Minnesota Amplatz Children's Hospital, Minneapolis, MN; 5) Royal Alexandra Hospital for Children, Westmead, NSW, Australia; 6) Charité University Medicine, Berlin, Germany; 7) University Children's Hospital, Heidelberg, Germany; 8) University Children's Hospital, Zürich, Switzerland.

Monogenic diabetes is a rare genetically and clinically heterogenic disease with mutations in more than 20 genes known to cause it in a dominant or recessive manner. Interestingly, several of these genes overlap with the susceptibility loci of polygenic type 2 diabetes (T2DM), suggesting that mutations in genes causing a rare disease can contribute to the development of the common disease. However, the genetic defects in many families with presumed monogenic diabetes still remain unclear. Identification of them would give us an insight in the pathogenesis mechanisms of not only monogenic diabetes but also of polygenic T2DM as well as might suggest new treatment strategies for both diseases. Here, we employed the whole-genome sequencing of a family with early-onset diabetes and identified a novel deletion in PCBD1, coding for a dimerization cofactor of HNF1A and HNF1B transcription factors, which have previously been shown to cause monogenic diabetes. Moreover, we identified three additional diabetic cases among the patients with neonatal hyperphenylalaninemia caused by the homozygous loss of function mutations in PCBD1. Interestingly, all these patients could change from insulin to oral antidiabetic drug therapy resulting in an improved glycemic control. Besides that, functional analysis showed Pcbd1 expression in the pancreas of Xenopus and mouse embryos from early specification onward, with enrichment in endocrine progenitor cells, suggesting an evolutionary conserved function. Furthermore, a morpholino-mediated knockdown in Xenopus revealed that Pcbd1 activity is required for a proper early pancreatic fate specification. The expression of Pcbd1 was also maintained in the mouse insulinoma cell line, however, transient Pcbd1 knockdown neither affected cell viability nor the expression of several genes important for insulin production and secretion. Therefore, it is likely that the reduced pancreatic beta cell progenitor pool might be responsible for the development of the disease. In summary, we provide not only genetic but also functional evidence that PCBD1 mutations can cause early-onset monogenic diabetes. Moreover, it can be treated with oral antidiabetic drugs instead of insulin resulting in a better outcome for the patients.

## 233

**The Ankrd11 mutation in the Yoda mouse mirrors the human gene defect and provides new insights into KBG syndrome.** *K. Walz[1,2], D. Cohen[1], P.M. Neilsen[3], J. Foster II[1], F. Brancati[4,5], K. Demir[6], R. Fisher[7], M. Moffat[8], N.E. Verbeek[9], K. Bjorgo[10], A. Lo-Castro[11], P. Curatolo[11], G. Novelli[5], C. Abad[1], C. Lei[1], O. Diaz-Horta[1], J.I. Young[1], D.F. Callen[12], M. Tekin[1].* 1) Dr. John T. Macdonald Foundation Department of Human Genetics and John P. Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, FL, 33136, USA; 2) Department of Medicine, Miller School of Medicine, University of Miami, FL, 33136, USA; 3) Swinburne University of Technology Sarawak Campus, Kuching, Sarawak, Malaysia; 4) Department of Medical, Oral and Biotechnological Sciences, Gabriele D'Annunzio University, 66100 Chieti, Italy; 5) Medical Genetics Unit, Policlinico Tor Vergata University Hospital, Viale Oxford 81, 00133 Rome, Italy; 6) Division of Pediatric Endocrinology, Dokuz Eylül University Faculty of Medicine, Narlidere, İzmir, Turkey; 7) Northern Genetics Service Teesside Genetics Unit, The James Cook University Hospital Marton Road Middlesbrough TS4 3BW, UK; 8) Department of Paediatric Dentistry, Newcastle Dental Hospital and School, Richardson Road, Newcastle upon Tyne, Newcastle NE2 4AZ, UK; 9) Department of Medical Genetics, University Medical Center Utrecht, Lundlaan 6, 3584 EA Utrecht, Netherlands; 10) Department of Medical Genetics, Oslo University Hospital, Kirkeveien 166 0450 Oslo, Norway; 11) Department of Neuroscience, Pediatric Neurology and Psychiatry Unit, Tor Vergata University of Rome, Italy; 12) 3Centre for Personalised Cancer Medicine, University of Adelaide, Adelaide SA 5000, Australia.

Mutations in ANKRD11 have recently been reported to cause KBG syndrome, an autosomal dominant condition characterized by intellectual disability (ID), behavioral problems, and macrodontia. To understand the pathogenic mechanism that relates ANKRD11 mutations with the phenotype of KBG syndrome, we studied the cellular characteristics of wildtype ANKRD11 and the effects of mutations in humans and mice. The characterization of a mouse model of KBG syndrome carrying a missense mutation in Ankrd11 showed hypo-activity, increased anxiety, presence of repetitive behaviors, impaired learning and memory, and sociability and preference for social novelty for the mutant mice, consistent with the human phenotype. In addition, we show that the abundance of ANKRD11 is tightly regulated during the cell cycle through degradation by proteasome, which requires the proteasome-dependent destruction boxes (D-box) at the ANKRD11 C-terminus. Analysis of 11 pathogenic ANKRD11 variants in humans, including six reported in this work, and one reported in the Ankrd11Yod/+ mouse, shows that all mutations affect the D-box signals at the C-terminus and the mutant protein accumulates aberrantly in the nucleoli. We conclude that ANKRD11 C-terminus D-boxes play an important role in regulating the abundance of the protein during the cell cycle, and disturbance of this role by a mutation leads to KBG syndrome.

## 234

**Defects in TAPT1, involved in axial skeletal patterning, cause a complex lethal recessive disorder of skeletal development.** *S. Symoens[1], A. Barnes[2], C. Ghistelinck[1], F. Malfait[1], K. Vleminckx[1], B. Guillemyn[1], D. Syx[1], W. Steyaert[1], E. Parthoens[3], M. Biervliet[4], G. Gillessen-Kaesbach[5], J. De Backer[1], A. Willaert[1], H.P. Bächinger[6,7], A. De Paepe[1], J.C. Marini[2], P.J. Coucke[1].* 1) Center for Medical Genetics Ghent, Ghent University Hospital, Ghent, Belgium; 2) Bone and Extracellular Matrix Branch, NICHD, NIH, Bethesda, MD, United States; 3) Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium; 4) Center for Medical Genetics, Brussels University Hospital, Brussels, Belgium; 5) Institut für Humangenetik Lübeck, Universitätsklinikum Schleswig-Holstein, Lübeck, Germany; 6) Research Department, Shriners Hospitals for Children, Portland, OR, United States; 7) Department of Biochemistry and Molecular Biology, Oregon Health & Science University, Portland, OR, United States.

*TAPT1* encodes the evolutionary highly conserved Transmembrane Anterior Posterior Transformation-1 protein. ENU mutagenesis of *TAPT1* results in embryonic lethality of murine homozygotes, with posterior to anterior transformations of thoracic and lumbar vertebrae. The mechanism by which this ubiquitously expressed protein causes a specific patterning defect and lethality is unknown. We describe a Moroccan family with three lethal fetuses affected with fractures of ribs and long bones, undermineralized skull and axial skeleton, hydramnios with ascites and dilated ventricles. Because of the occurrence of multiple fractures and undermineralized skeleton, a clinical diagnosis of lethal autosomal recessive Osteogenesis Imperfecta was suggested. Although type I collagen folding was slightly delayed, causing mild overmodification of type I collagen, thorough molecular analysis of all known OI genes did not detect a causal mutation. We combined homozygosity mapping with exome sequencing, which identified a homozygous c.1108-1G>C mutation in *TAPT1*, causing in-frame skipping of exon 10. A second homozygous *TAPT1* missense mutation in exon 9 (c.1058A>T, p.(Asp353-Val)) was identified by direct sequencing in a complex Syrian pedigree with three lethal fetuses with fractures and multiple congenital anomalies of brain, face, heart and lungs. Immunocytochemical staining of dermal fibroblasts revealed co-localization of TAPT1 with the centrosomal protein γ-tubulin, while co-staining with acetylated tubulin detected that TAPT1 forms a pocket in which the primary cilium is inserted. Increased *TAPT1* expression was observed during cilium formation. Moreover, we showed in patients' dermal fibroblasts that primary cilium formation is severely disturbed. A zebrafish *tapt1b*-morpholino-approach revealed severe cartilage malformation and a delay in bone formation. Our results show that defects in *TAPT1* underlie a novel autosomal recessive disorder, which is characterized by multiple fractures in utero, generalized undermineralization of the skeleton, microbrachycephaly, ascites and pleural effusion. We also prove that TAPT1 is a centrosomal protein that is of crucial importance for proper cilium formation, thereby suggesting that this disorder is a novel ciliopathy.

## 235

**Mutations in *KITLG*, encoding KIT ligand, cause unilateral hearing loss.**

*C. Zazo Seco[1,2,3], L.S. Serrao de Castro[4,5], J.W. van Nierop[1,3], M. Schraders[1,2,3], E.J. Verver[1], M. Morín[4,5], N. Maiwald[6], M. Wesdrop[1], H. Venselaar[8], L. Spruijt[6], J. Oostrik[1,2,3], J. Schoots[6], L.H. Hoefsloot[6], J.H. Jansen[2,7], G. Huls[2,7], M.M. Van Rossum[9], H.P. Kunst[1,3], M.A. Moreno-Pelayo[4,5], H. Kremer[1,2,3,6], Baylor-Hopkins Center for Mendelian Genomics.* 1) Department of Otorhinolaryngology, Hearing & Genes, Radboud university medical center, Nijmegen, Netherlands; 2) The Radboud Institute for Molecular Life Sciences, Radboud university medical center, Nijmegen, Netherlands; 3) Donders Institute for Brain, Cognition and Behaviour, Radboud university medical center, Nijmegen, Netherlands; 4) Servicio de Genética, Hospital Universitario Ramon y Cajal, IRYCIS, Madrid, Spain; 5) Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Madrid, Spain; 6) Department of Human Genetics, Radboud university medical center, Nijmegen, Netherlands; 7) Department of Laboratory Medicine, Laboratory of Hematology, Radboud university medical center, Nijmegen, Netherlands; 8) Centre for Molecular and Biomolecular Informatics, Radboud university medical center, Nijmegen, Netherlands; 9) Department of Dermatology, Radboud university medical center, Nijmegen, Netherlands.

Familial nonsyndromic unilateral hearing loss (UHL) is uncommon with very few affected families described in literature. To date, no genes or loci are known to be involved in familial nonsyndromic UHL. In our study, we focused on elucidating the genetic cause underlying HL in a large Dutch family (W09-1628) with congenital, nonsyndromic, unilateral or asymmetric, stable, severe-to-profound HL. The UHL in family W09-1628 is inherited in a dominant manner and exhibits reduced penetrance. Our approach was a combined strategy of linkage analysis and whole exome sequencing which revealed a heterozygous deletion that creates a nonsense mutation in *KITLG* as the putative cause of the UHL. This mutation segregates with the HL in the family and it is not present in 306 Dutch control alleles, the Exome Variant Server, 1000 Genomes and in the Nijmegen exome database (n= 2096). In order to further address the involvement of *KITLG* mutations in UHL, sequence analysis of the coding regions and splice sites of *KITLG* was performed in 16 UHL index patients, mostly of Dutch origin, and in 25 UHL index patients from Spanish origin. This revealed a heterozygous deletion in a patient of Spanish origin. This mutation segregates with the HL in the family and it is neither present in 188 Spanish control alleles nor in any variation database described above. The mutation affects a highly conserved cysteine residue involved in a Cys-Cys intramolecular bond. *In silico* protein modeling predicts that this mutation affects the local structure of the protein and the biological activity of KITLG. *KITLG* encodes KIT ligand which, after binding to the KIT receptor, triggers a downstream cascade that has an effect on the proliferation, migration and cell survival of melanocytes, hematopoietic stem cells and primordial germ cells. There are mutant mice with semi-dominant mutations in *Kit* and *Kitl* that exhibit HL. The HL in the *Kit* mutant $W^v/W^v$ is unilateral and the mutant allele shows reduced penetrance. The HL in the *Kitl* mutant $Sl^d/Sl^d$ is congenital and severe. In both mutants, the level of hearing seems to be dependent on the number of melanocytes that migrate to the stria vascularis and survive. There, melanocytes are essential for generating the endocochlear potential which is the driving force for sensory hair cell depolarization. Besides the UHL, no further abnormalities were seen in these families with heterozygous mutations in *KITLG*.

## 236

**Molecular pathogenesis of Tuberous Sclerosis Complex (TSC) in patients with no mutation identified in *TSC1* or *TSC2*.**

*M.E. Tyburczy[1], Y. Chekaluk[1], K. Dies[2], M. Sahin[2], J. Glass[3], D. Franz[3], S. Camposano[4], E. Thiele[4], D. Kwiatkowski[1].* 1) Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 2) Children's Hospital, Boston, MA; 3) Cincinnati Children's Hospital Medical Center, Cincinnati, OH; 4) Massachusetts General Hospital, Boston, MA.

Tuberous sclerosis complex (TSC) is an autosomal dominant disorder caused by mutations in *TSC1* or *TSC2*. Two-thirds of TSC cases are sporadic, and mosaicism is known to occur at low frequency in TSC. The use of Sanger sequencing and deletion analysis of *TSC1* and *TSC2* results in 10-15% of TSC patients being diagnosed with no mutation identified (NMI). We hypothesized that NMI occurs due to: mosaic mutations, intronic mutations, technical failure to identify the usual mutations, and possibly the existence of a third, as yet undiscovered, TSC locus. We used a series of molecular genetic techniques, including next-gen sequencing (NGS), to investigate this hypothesis in 51 TSC NMI patients. We performed NGS of the genomic extent of *TSC1* and *TSC2* including promoter regions, all exons, and most of the introns on DNA prepared from blood cells or saliva from patients with a definite clinical diagnosis of TSC. We found a *TSC1* genomic deletion in 1 case and *TSC2* genomic deletions in 3 cases by MLPA. NGS was performed on the remaining 47 patients. Variants of known or potential pathogenic significance were identified in 32 (68%) patients: 26 (81%) variants in *TSC2* and 6 (19%) in *TSC1*. All of these variants were absent in DNA samples from unaffected parents, supporting pathogenicity, and all were confirmed by secondary assays. Mutations were mosaic in 17 (53%) patients with mutant allele frequencies ranging from 1% to 34%, and splice site mutations were seen in 14 (44%) cases. In one patient a 740 nt deletion was found in the *TSC2* promoter. We also identified 4 heterozygous mutations in *TSC1* and *TSC2* coding exons that appear to have been missed by previous analyses. We also searched for mutations in skin tumors (angiofibromas) from 3 NMI patients. In these patients, we identified mosaic mutations in angiofibromas from 2 of 3 samples studied, and neither was detected in respective blood samples (< 0.1% allele frequency). In conclusion, our study clearly indicates that NGS is necessary to detect low allelic mutations in *TSC1* and *TSC2* as these results detected mutations in 68% of TSC NMI patients. Mosaic and splice region mutations were common, and were missed by previous analyses. NGS has the potential for more efficient and sensitive mutation detection in TSC. Furthermore, analysis of accessible TSC skin tumors enables mutation detection, and defines a set of TSC patients with no detectable mutation in blood or saliva.

## 237

**RAB11FIP1 interacts with the BLOC-1 complex to retrieve melanogenic proteins from the recycling pathway and a dominant negative mutation in *RAB11FIP1* causes Hermanksy-Pudlak Syndrome Type 10 (HPS-10).** *A.R. Cullinane[1], M.A. Merideth[1], M.B. Datiles[2], J.A. Curry[1], N.F. Hansen[3], J.K. Teer[4], J.G. White[5], J.C. Mullikin[3,6], M. Huizing[1], W.A. Gahl[1].* 1) Medical Genetics Branch, NHGRI (NIH), Bethesda, MD; 2) NEI, NIH, Bethesda MD 20892, USA; 3) Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, NHGRI, NIH, Rockville, MD, USA; 4) Department of Biomedical Informatics, H. Lee Moffitt Cancer Center and Research Institute, Tampa FL 33612, USA; 5) Department of Laboratory Medicine, University of Minnesota, Minneapolis, MN 55455, USA; 6) NIH Intramural Sequencing Center, NHGRI, NIH, Rockville MD 20852, USA.

Hermansky-Pudlak Syndrome (HPS) is a genetically heterogeneous disorder of lysosome-related organelle (LRO) biogenesis and is characterized by oculocutaneous albinism and a bleeding diathesis. There are currently 9 known genes that cause HPS; all of whose protein products function in the biogenesis of LROs. The Biogenesis of Lysosome related Organelle Complex 1 (BLOC-1) contains 8 subunits but relatively little is known about the intracellular function of the complex, although a role in endosomal protein sorting has been suggested. Using his-tagged BLOC-1 subunits expressed in HEK293 cells and mass spectroscopy, we discovered that RAB11FIP1 is a novel interacting protein of the BLOC-1 complex. *RAB11FIP1* encodes a RAB11A interacting protein that homo-dimerizes to interact with RAB11A. A yeast-2-hybrid assay showed that the dysbindin subunit of BLOC-1 directly interacts with RAB11FIP1; this was confirmed by co-immunoprecipitation and confocal immunofluorescence microscopy in melanocytes. We now report a girl who had previously been screened for mutations in *HPS1* through *HPS6* and all the genes encoding the BLOC-1 complex. No mutations were found, although the patient had typical signs and symptoms of HPS and a cellular phenotype mimicking that of BLOC-1, i.e., increased plasma membrane cycling in melanocytes and endosomal accumulation of a melanogenic protein, TYRP1. Whole exome sequencing revealed a de novo heterozygous frameshift mutation in *RAB11FIP1*. The short protein fragment from this allele was expressed and interacted with the full-length protein, resulting in a dominant negative effect. Known cargos of the BLOC-1 complex in melanocytes are TYRP1 and ATP7A. How these cargos traffic to LROs was unknown, but we discovered that GFP-TYRP1 traffics to the plasma membrane, is endocytosed and only then is directed to LROs. We demonstrated that TYRP1 and ATP7A interact with the AP-1 complex, allowing this trafficking to occur. However, ATP7A appears to traffic directly to endocytic vesicles, where RAB11FIP1 and the BLOC-1 complex are required for retrieval to LROs. Taken together, these data suggest a function of the BLOC-1 complex in retarding protein recycling by forming a physical brake between early endosomes (through the BLOC-1 interactor, Syntaxin-13) and recycling endosomes (through the BLOC-1 interactor, RAB11FIP1). This would allow more time for proteins to be retrieved from the endosomal compartment (by the AP-1 complex) and directed to LROs.

## 238

**Genetic Basis and Functional Consequences of Chromatin State Variability across Individuals.** *F. Grubert[1], J. Zaugg[1], M. Kasowski[1], O. Ursu[1], D. Spacek[1], A. Martin[1], L. Steinmetz[1,2], A. Kundaje[1], M. Snyder[1].* 1) Dept Genetics, Stanford University, Stanford, CA; 2) European Molecular Biology Laboratory, Genome Biology, Heidelberg, Germany.

One of the continuing challenges in biomedical research is to understand the genetic contribution to hereditary traits and diseases. The number of associations between genetic variants (e.g. single nucleotide polymorphisms (SNPs)) and complex diseases is rising rapidly, however, our understanding of the underlying molecular mechanisms is lagging far behind. One reason for this lack of understanding is that most disease-associated loci lie in the non-coding part of the genome. Previously we have shown that disease-loci are enriched in regulatory elements that show great variability among individuals. Here we performed ChIP-Seq experiments for three different histone modifications (H3K4m3, H3K4me1 and H3K27ac) across 70 unrelated individuals to better understand the genetic contribution of the observed variability in chromatin-states. Using the histone marks as quantitative traits we identified a quantitative trait locus (QTL) for more than 10% of all regulatory elements, whereby enhancers are more likely to have a QTL than are promoters. More than 50% of the chromatin QTLs we identified for a single mark also significantly affect at least one other mark indicating a shared mechanism influencing the activity of different histone marks. A potential mechanism that could explain the genetic basis of variable chromatin activity is the disruption of transcription factor (TF) binding sites through single SNPs. These motif disruptions could prevent binding of sequence-specific TFs, which often recruit histone-modifying enzymes to their genomic targets. We also found that chromatin QTLs are significantly enriched in SNPs previously identified in genome-wide association studies (GWAS), providing further evidence of the functional implications of chromatin variability in humans. Overall our study will allow us to better understand the molecular mechanism that underlie known associations between genotype and complex genetic diseases.

## 239

**Integration of 111 reference human epigenomes helps interpret the molecular basis of complex traits and disease.** *M. Kellis, C. Roadmap Epigenomics.* Computer Science, Broad Institute, MIT & Harvard, Cambridge, MA.

The NIH Roadmap Epigenomics Consortium generated the largest collection to-date of human epigenomes for primary cells and tissues. Here, we undertake an integrative analysis of 111 reference human epigenomes profiled for histone modification patterns, chromatin accessibility, DNA methylation, and RNA expression. We establish global maps of regulatory elements, define enhancer modules of coordinated activity, and infer regulatory networks linking enhancers to regulators and target genes. We study DNA methylation across chromatin states, cell types, and during differentiation, and the epigenomic signatures that distinguish age, sex, and lineage specification. We find that epigenetic features at enhancer regions are highly dynamic features across tissues, individuals, genotypes, and disease state, in contrast to mostly invariant features at promoters. We use epigenomic information to predict pathways harboring regulatory variants associated with complex traits, to identify disease-relevant cell types and regulators, and to determine the tissue-of-origin of cancer samples. Lastly, we use model organisms to support the biological relevance of sequence variants with associations below genome-wide significance but lying in relevant epigenomic states in Alzheimer's and cardiac phenotypes. Our results demonstrate the central role of epigenomic information for understanding gene regulation, cellular differentiation, and human disease.

## 240

**An imprinting map of the human placenta based on the application of a novel population genetics approach to RNAseq data.** *C.T. Watson[1], O. Rodriguez[1], B. Jadhav[1], N. Azam[1], D.J. Ho[1], K. Cheung[1], D. Sachs[1], K. Hao[1], R.J. Wright[2], E.E. Schadt[1], A.J. Sharp[1].* 1) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY; 2) Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY.

We have developed a novel approach for identifying imprinted gene expression through analysis of population genotype frequencies in RNAseq data. Under the null model of bi-allelic expression, the observed genotype frequencies for a transcribed SNP are expected to follow classical Hardy-Weinberg equilibrium (HWE). However, where consistent mono-allelic expression occurs, such as at imprinted loci, this will manifest as a significant departure from HWE in mRNA, with a depletion of heterozygotes. We have applied this approach to analyze an RNAseq dataset from 180 placental samples from the National Children's Study. We identified 1,213 transcribed SNPs with significantly reduced rates of heterozygosity (HWE $p<0.001$), corresponding to 56 Refseq genes, three large microRNA clusters, and a further 272 "intergenic" SNPs located outside of annotated RefSeq transcripts. This list includes 18 previously identified imprinted genes. In order to validate imprinted expression of these loci, we harnessed maternal SNP genotypes available from 31 of the placenta in our study. Utilizing a novel phasing approach we were able to assign parental origin to the expressed allele at ~97% of heterozygous sites. Considering those SNPs for which there were multiple informative placentae (n=598), 96% showed monoallelic expression of the same parental allele across all samples, indicating imprinting as the predominant underlying mechanism. Based on whole genome bisulfite sequencing data, we also observed allele-specific methylation associated with many of these transcripts, and performed validations of parental-specific expression and methylation marks in an independent cohort of placenta/mother pairs. These findings demonstrate the robustness of our HWE approach, which provides the first comprehensive map of imprinting in the human placenta. Intriguingly, many of these placental-specific imprinted genes have functions that are consistent with the parental conflict hypothesis, in which it is theorized that imprinting evolved primarily to regulate transfer of nutrients from mother to offspring: we observed novel paternally-expressed genes that positively regulate angiogenesis, and maternally-expressed genes with inhibitory effects on capillary formation. Furthermore, some of these imprinted genes have recently been shown to have altered expression levels in growth-restricted fetuses, suggesting that they represent strong candidates as modifiers of fetal growth.

## 241

**Tissue-specific patterns of imprinting revealed by analysis of monoallelic expression in human populations.** *T. Lappalainen[1,2], Y. Baran[3], E. Tsang[4], T. Tukiainen[5], M.A. Rivas[6], M. Pirinen[7], M. Gutierrez-Arcelus[8], The. GTEx Consortium[9], D.G. MacArthur[5,9], S.B. Montgomery[4], N.A. Zaitlen[10].* 1) New York Genome Center, New York, NY; 2) Department of Systems Biology, Columbia University, New York, NY; 3) The Blavatnik School of Computer Science, Tel-Aviv University, Israel; 4) Departments of Pathology and Genetics, Stanford University, CA; 5) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 6) Wellcome Trust Center for Human Genetics, Oxford, UK; 7) Institute for Molecular Medicine Finland, University of Helsinki, Finland; 8) Department of Genetic Medicine and Development, University of Geneva, Switzerland; 9) The Broad Institute, Boston, MA; 10) Department of Medicine, University of California San Francisco, CA.

Imprinting is an epigenetic mechanism that leads to parent-of-origin effects via imbalanced expression of the maternally and paternally inherited alleles, and it has been shown to play a role in Mendelian and common disease and cancer. This study is the first systematic characterization of imprinting in multiple primary tissues from adult humans, using allele-specific expression data of the GTEx project with 1652 RNA-seq samples from 178 unrelated individuals. Our novel filtering and likelihood-based approach based on probabilistic generative models distinguishes imprinting from other sources of monoallelic expression. We identify 47 imprinted genes, of which 29 have been identified before. However, 34% of "known" imprinted genes are biallelic in our data set, demonstrating how poorly previous catalogs capture imprinting in primary tissues of adults. Methylation array data shows low correlation with imprinting status in expression data, supporting recent reports on complex epigenetic mechanisms underlying parental monoallelic expression. We show widespread tissue-specificity of imprinting, with 37/47 imprinted genes being biallelic in at least one tissue. The direction of imprinting can also change between tissues: the IGF2 gene with previously known maternal imprinting is in fact paternally imprinted in the brain. We also observe instances in which some individuals are imprinted while others exhibit biallelic expression. In muscle, females have less imprinting than males ($p=0.0053$) and imprinted genes have more sex-specific expression ($p=0.012$), pointing to gender-specific parental effects. While testis has less imprinting than other tissues ($p=3.8 \times 10^{-5}$), in general imprinted genes are highly expressed in tissues with endocrinological functions, consistent with their role in growth regulation. We also putatively characterize how monoallelic expression could modify the phenotypic impact and penetrance of functional coding variants in imprinted genes. Altogether, our results demonstrate that imprinting is not a stable property of a gene but can vary substantially between tissues and individuals. This adds to our understanding of the mechanisms of imprinting and its role in the function of the human genome.

## 242

**Population-scale and single-cell RNA sequencing provides insight into X chromosome inactivation.** *T. Tukiainen[1,2], A. Kirby[1,2], T. Lappalainen[3,4], A.-C. Villani[2,5], R. Satija[2], J. Maller[1,2], . The GTEx Project Consortium[6], A. Regev[5], N. Hacohen[2,5], D.G. MacArthur[1,2].* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 3) New York Genome Center, New York, NY; 4) Columbia University, New York, USA, NY; 5) Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Charlestown, MA; 6) The Genotype-Tissue Expression (GTEx) Project Consortium.

X chromosome inactivation (XCI) effectively balances the expression dosage of X chromosome genes between men and women by randomly silencing one of the two X chromosomes in each female cell. However, XCI is incomplete and variable; more than 15% of X chromosome genes have been reported to fully or partially escape from inactivation, but the full extent of XCI and its variation between tissues and individuals remains unclear. We have deployed several complementary approaches based on high-throughput RNA sequencing to comprehensively profile the landscape and variability of escape from XCI. Using detailed gene expression data from the Geuvadis and GTEx consortia we show that approximately half of the reported escape genes demonstrate male/female expression differences detectable at population-level. For these genes sex-biased expression is present and directionally similar across the various tissues studied, suggesting XCI is tightly and uniformly regulated across human tissues. Our analysis also highlights several novel candidate escape genes following similar sex-bias patterns. In line with earlier studies, we observe that even in genes with some degree of XCI escape the expression from the inactive copy of X is rarely as high as from the active X, potentially explaining why not all reported escape genes demonstrate large male/female expression differences. To confirm these observations and assess individual-level variability in escape from XCI we have analyzed high-throughput single-cell RNA-seq data from 192 cells from an exome-sequenced female sample, allowing the direct determination of expression from the inactive and active X chromosomes. In addition, we have assessed the allelic imbalance across the X chromosome in monoclonal tissue samples and in tissue samples showing skewed X inactivation. These analyses highlight well-known escape genes, replicate several of our novel candidates, and also confidently flag several additional candidate XCI escape genes with only modest sex-bias in the population-level analysis, hence extending the number of genes with variable escape and underscoring the large degree of inter-individual variability in X inactivation. Together these analyses provide a comprehensive view of the landscape of escape from XCI, essential for deeper understanding on how the process and escape genes contribute to sexual dimorphism and sex chromosome aneuploidies.

## 243

***Cis*-methylation quantitative trait loci mapping of chromosome 15q25.1 in human brain reveals novel genetic associations with nicotine dependence.** *D.B. Hancock[1], J.C. Wang[2], N.C. Gaddis[1], N.L. Saccone[2], J.A. Stitzel[3], A. Goate[2], L.J. Bierut[2], E.O. Johnson[1].* 1) RTI International, Research Triangle Park, NC; 2) Washington University, St. Louis, MO; 3) University of Colorado, Boulder, CO.

Genome-wide association studies have unequivocally found that single nucleotide polymorphisms (SNPs) on chromosome 15q25.1 contribute to nicotine dependence and other smoking behaviors. The associated SNPs span genes that encode iron-responsive element binding protein 2 (*IREB2*), hydroxylysine kinase (*HYKK*), proteasome subunit alpha type-4 (*PSMA4*), and three nicotinic acetylcholine receptors (*CHRNA5*, *CHRNA3*, and *CHRNB4*). Prior analyses of this region found that the associated SNPs have important biological functions in human brain: the missense SNP rs16969968 alters the receptor function of CHRNA5, and several upstream and intronic SNPs tag expression quantitative trait loci (eQTL) regulating *CHRNA5* mRNA expression. To identify other biologically important SNPs that may contribute to nicotine dependence, we conducted *cis*-methylation QTL (*cis*-meQTL) mapping using SNP genotypes and DNA methylation levels measured across the *IREB2-HYKK-PSMA4-CHRNA5-CHRNA3-CHRNB4* gene region in the BrainCloud and Brain QTL cohorts (total N= 175 European Americans and 65 African Americans). We found significant associations between 8 SNPs and *CHRNA5* methylation ($5.75\times10^{-5} < P < 6.04\times10^{-10}$) and between 2 SNPs and *CHRNA3* methylation ($1.29\times10^{-4} < P < 7.67\times10^{-6}$) in prefrontal cortex. We also observed significant associations of these SNPs with *CHRNA5* methylation in frontal cortex, temporal cortex, and pons ($3.42\times10^{-3} < P < 4.92\times10^{-12}$). We tested the newly identified *cis*-meQTL SNPs for association with nicotine dependence in a meta-analysis across four independent cohorts with total N=9,815 (5,549 European Americans, 2,309 Italians, and 1,957 African Americans). All of the *CHRNA5*-implicated SNPs were nominally to significantly associated with nicotine dependence ($7.99\times10^{-3} < P < 8.83\times10^{-4}$): rs2292117, rs6495306, rs680244 and rs621849 tagged known *cis*-eQTL SNPs associated with nicotine dependence, whereas rs12915366, rs3743077, rs950776, and rs11636753 represent new association signals. The SNP minor alleles were associated with higher *CHRNA5* DNA methylation and mRNA expression levels and decreased risk of nicotine dependence. This inverse relationship is consistent with previously reported rodent models. Our findings are the first to connect previously observed differences in *CHRNA5* mRNA expression and nicotine dependence risk to underlying differences in DNA methylation.

## 244

**Joint methylome- and genome-wide association studies in blood and brain identifies new disease mechanisms for schizophrenia.** *E.J.C.G. Van den Oord[1], A. Shabalin[1], G. Kumar[1], S. Clark[1], J.L. McClay[1], L.Y. Xie[1], R. Chan[1], S. wedish Schizophrenia Consort.[1,3,4], V. Vladimirov[1,2], C. Hultman[3], P.F. Sullivan[3,4], P.K.E. Magnusson[3], K.A. Aberg[1].* 1) Center Biomarker Research and Personalized Med, Virginia Commonwealth Univ, Richmond, VA; 2) Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond ,VA, USA; 3) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 4) Department of Genetics, University of North Carolina at Chapel Hill, NC, USA.

We performed joint analyses of data from methylome- (MWAS) and genome-wide association studies (GWAS) to identify schizophrenia (SZ) disease mechanisms. Methylation measurements for 28 million CpGs were obtained with methyl-binding domain enrichment followed by sequencing (MBD-seq) in post-mortem brain tissue (prefrontal cortex, N=75) and whole blood (N=1,459) of SZ patients and controls. Of the 67 million reads/sample, 47% were used after alignment and QC. After 1000 genomes imputation and selection on imputation quality and MAF > 0.05, 5 million SNPs were available. A variety of mechanisms were tested such as effects CpGs created or destroyed by SNPs (CpG-SNPs) or whether effects of cis/trans methylation quantitative trait loci (meQTLs) were altered in cases versus controls. Analyses were performed with a specifically designed analysis pipeline that was bundled in an ultra-fast and memory efficient software called RaMWAS. Critical findings were replicated using targeted (bisulfite) pyrosequencing in independent SZ case-control samples from post-mortem brain tissue (N=50) and blood (N=400-1,100). Permutation tests showed partial but significant overlap between MWAS and GWAS top findings with two SZ GWAS meta-analyses (N=32,143 and 21,953; $p < 0.031$ and 7×10-4). Overlapping findings frequently involved CpG-SNPs. One example is a CpG-SNP in interleukin receptor, IL1RAP, which was associated with the same direction of effects in both blood and brain and replicated in independent samples ($p < 1.6\times10\text{-}4$ N=368). The majority (~68%) of CpG-SNPs were likely methylated with 94% of these sites being methylated in both brain and blood. Because CpG-SNPs show individual variation in both sequence and methylation and may have similar methylation statuses across multiple tissues, we also test how relevant there sites are for other diseases. Using the NHGRI GWAS catalogue finding for all disease, we observed substantial enrichment (odds ratio = 3). Other overlapping SZ MWAS and GWAS findings implicated transcription factor binding sites where both genotype and methylation status could potentially inhibit the binding of transcription factor to their recognition elements (e.g. CREB1, replication $p < 1\times10\text{-}10$, N=1,086). The MWAS findings that did not overlap with GWAS tended to reflect environmental insults (e.g. hypoxia and inflammation). In summary, our joint analyses identified replicating sites that implicated specific hypotheses about SZ disease mechanisms.

## 245

**Whole genome bisulfite sequencing of acute lymphoblastic leukemia cells.** *P. Wahlberg[1], A. Lundmark[1], J. Nordlund[1], A. Raine[1], S. Busche[5], E. Forestier[2,4], T. Pastinen[5,6], G. Lönnerholm[3,4], A-C. Syvänen[1].* 1) Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Sweden; 2) Department of Medical Biosciences, University of Umeå, Sweden; 3) Department of Children's and Women's Health, Uppsala University, Sweden; 4) For the Nordic Society of Pediatric Hematology and Oncology (NOPHO); 5) Department of Human Genetics, McGill University, Montréal, Québec H3A0G1, Canada; 6) Department of Human Genetics, McGill University and Genome Quebec Innovation Center, Montréal, Québec H3T1C5, Canada.

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer in the developed countries. Aberrant patterns of methylation have been associated with cancer and epigenetic modifiers such as the ten eleven translocation (TET) family of enzymes and DNA methyltransferases (DNMTs) are recurrently mutated in leukemic cancers. We have previously documented large differences between ALL subtypes and normal B- and T-cells using the Illumina Human 450K BeadArray platform. To further investigate the distribution of CpG methylation in ALL cells, we generated Whole Genome Bisulfite Sequencing (WGBS) data at high coverage (20-30X) of four Swedish pediatric ALL patients with different genetic subtypes of ALL. In our study we also included low-pass (~5X) WGBS data from B- (n=8) and T-cells (n=14) and additional BCP-ALL t(12;21) patients (n=3). Comparison of the WGBS data from the ALL methylomes revealed that the ALL methylome generally have higher methylation levels than B- and T-cells. Specifically we observed aberrant CpG island methylation in ALL samples with the t(12;21)ETV-RUNX1 translocation. Principal component analysis (PCA) of methylation data from CpG islands showed limited variation between samples, with the exception of the ALL subtype t(12;21)ETV-RUNX1 that formed a separate cluster. On the contrary, PCA using methylation levels from CpG island shores reveal distinct cell-type specific T- and B-cell clusters separated from the ALL samples. The DMR analysis between ALL samples and B- and T-cell identified between 9,235 - 15,924 DMRs with an average size of 1 kb and in sizes from 200 bp up to 43 kb. In the ALL subtype t(12;21)ETV-RUNX1 we identified large-scale locus specific DMRs at cancer-related genes that were associated with increased RNA expression. ALL specific DMRs and cell-type specific B- and T-cell DMRs are distributed across the genome in a similar manner with the exception of hypermethylated DMRs annotated to CpG islands that are enriched in ALL samples. DMRs are under-represented in intergenic regions and enriched to regions associated to genes. The fact that the DMRs overlap with and are enriched to functional sequences indicates that they have a functional role in ALL.

## 246

**Completion of The 1000 Genomes Project: Results, Lessons Learned and Open Questions.** *G. Abecasis, The 1000 Genomes Project.* Ctr Statistical Gen, Univ Michigan, SPH I, Ann Arbor, MI.

Starting in 2008, the 1000 Genomes Project (1000GP) set out to use next generation sequencing technologies to generate a catalog of human genetic variation and haplotypes. This publically available catalog now includes haplotypes for >2,500 individuals from 26 populations and >80 million genetic variants, ranging from SNPs, indels and other small variants, to insertions of mobile elements and other material, to large structural variants spanning 100s of kilobases. We summarize challenges, opportunities and technical and methodological advances encountered during the course of the project. In addition, we summarize insights about human genetic variation and the utility of project results for genetic association studies. Throughout the project, we have combined cost effective strategies for generating sequence data, public data release, thorough quality control of the resulting data, and integrated multiple methods for analysis to improve results. We have also developed software tools, methods and formats that are now in widespread use and allow sequence analysis and interpretation in a wide variety of contexts. Through advances in DNA sequencing technology and a combination of analytical approaches ranging from read mapping, to local reassembly, to full-scale de novo assembly of human genomes, our number of sequenced genomes has increased from ~180 in a first pilot analysis to >2,500 in our final release, and the proportion of each genome assessed with high confidence has increased from ~80% to ~96%. We assessed the accuracy and sensitivity of our results through comparisons with deep sequencing using Complete Genomics for 427 individuals and deep PCR-free Illumina data for 24 individuals as well as targeted long-read sequencing using PacBio. Our haplotype resource can aid genetic studies of disease across a variety of populations, provides insights about human demography and aids functional interpretation of the genome. Perhaps more importantly, the principles of open data sharing, collaboration, and friendly competition embodied by the project can be implemented in many future collaborations.

## 247

**Inferring the functional effects of non-synonymous variants using experimental results from deep mutational scanning.** *R.J. Hause, V.E. Gray, J. Shendure, D.M. Fowler.* Genome Sciences, University of Washington, Seattle, WA.

Investigating the consequences of non-synonymous genetic variation furthers our basic understanding of protein function while also facilitating the interpretation of variants observed in a clinical setting. Many computational tools exist to predict the effects of amino acid substitutions on protein function (e.g. SNAP, SIFT, PolyPhen-2); however, nearly all of these methods rely solely on evolutionary, biochemical, and structural information without leveraging experimental data. Where experimental data is used, it tends to be outdated, limited to a few mutations per protein. Deep mutational scanning (DMS) is a method that uses next-generation sequencing to experimentally measure the functional effects of hundreds of thousands of variants of a protein. Because DMS surveys the sequence-function landscape of proteins based on much larger numbers of mutations than what has been available to date, models trained on these datasets may improve the performance of models for predicting mutational consequences. We set out to utilize DMS data: (1) to better understand the relationship between properties of mutations and specific protein properties, (2) to predict the functional effects of mutations in proteins, and (3) ultimately, to improve the interpretability of "variants of unknown significance" in clinically relevant genes. To these ends, we are constructing an ensemble classifier based on evolutionary, physicochemical, and structural features to predict estimates of protein functionality derived from DMS of over 86 proteins. In ongoing work that will be presented at ASHG, we will analyze feature importance both globally and for specific functions (e.g. binding, stability). Using cross-validation and external validation on unpublished DMS datasets, we will demonstrate the extent to which our classifier is predictive of the functional effects of non-synonymous mutations and compare its performance to other available algorithms. We will also assess the ability of our algorithm to distinguish common, non-synonymous variants from the 1000 Genomes Project and the Exome Sequencing Project from rare, pathogenic non-synonymous variants in Clinvar and COSMIC. We anticipate that our model will improve prediction of functional and pathogenic variants, shed light on the underlying parameters that correlate with functionality and pathogenicity, and highlight the power of incorporating available experimental data from DMS into variant effect prediction.

## 248

**Context-specific regulatory networks identify key regulators of complex traits.** *G. Quon[1,2], D. Marbach[1,2], S. Feizi[1,2], M. Grzadkowski[1], M. Kellis[1,2].* 1) CSAIL, MIT, Cambridge, MA; 2) Broad Institute, Cambridge, MA.

Genome-wide association studies (GWAS) have identified thousands of single nucleotide variants associated with diverse human traits, but understanding their combined action in complex systems remains an open challenge. With more than 80% of lead GWAS SNPs located in non-coding regions of the genome rich in regulatory elements, functionally characterizing these variants necessitates knowledge of (1) the locations of cell type specific enhancers; (2) the identity of the target genes of those enhancers; and (3) the interactions between these target genes to identify disrupted pathways and subnetworks. Using enhancer and promoter maps for 111 cell types constructed by the Roadmap Epigenomics Consortium, we have constructed directed context-specific and cell type specific networks, where nodes represent both genes and non-coding regulatory elements (enhancers, promoters), and edges lead from transcription factors to regulatory elements, and regulatory elements to genes. These meta-networks, where nodes consist of both genes and non-coding regulatory elements, enable context-specific network analysis that can yield insight into the role of different cell types in complex traits, and to our knowledge have not been previously explored. To leverage these networks for GWAS analysis, we developed an efficient probabilistic model to map GWAS variants to candidate disrupted regulatory elements, and use each context-specific network to (1) identify trait-associated genes whose regulation is disrupted by non-coding variants; (2) identify master TF regulators of the trait-associated genes; and (3) identify other genes (not proximal to GWAS variants) involved in the trait. We predicted modules of non-coding variants associated with brain, cardiovascular, lipid, and immune-mediated disorders, as well as their regulators. Predicted regulatory modules and transcription factors involved in HDL and LDL cholesterol levels replicated across multiple studies and are most highly expressed in liver cell types. Furthermore, gene mutations reported in the MGI database lead to abnormalities including perturbed circulating lipid levels and susceptibility to atherosclerosis. Modules and regulators predicted for multiple sclerosis and Crohn's disease are most highly expressed in CD4+,CD8+, and CD34+ cells, and their mutants lead to defects in B-cell and NK-cell morphology and circulating levels.

## 249

**Allele-specific alternative splicing in diploid human genomes.** *N. Raghupathy, K. Choi, S.C. Munger, G.A. Churchill.* The Jackson Laboratory, Bar Harbor, ME.

Current practices for RNA-seq analysis employ three separate pipelines to quantify gene expression abundance, allele-specific expression (ASE), and alternative splicing. Gene-level abundance is estimated from alignment of all reads genome-wide, whereas ASE is assessed by analyzing only reads that overlap known SNP locations and alternative splicing is estimated by analyzing reads overlapping annotated splice junctions. We have developed computational tools, Seqnature and EMASE, to build individualized diploid genomes from phased genetic variations, align RNA-seq reads to individualized diploid transcriptomes, and estimate transcript abundance. The EM algorithm implemented in EMASE simultaneously estimates expression at the level of the allele, isoform and gene. Here we extend EMASE to include a splice-aware feature that enables the simultaneous estimation of allele-specific alternative splicing in addition to allele-specific and total gene abundances. The EM algorithm probabilistically allocates both allele and gene multi-mapping reads to estimate effective read counts at exons and splice-junctions. These read counts can be used to derive allele, isoform, and gene expression estimates. We demonstrate the utility of the approach using simulated and real human RNA-seq data from the 1000 Genomes Project Yoruban population as well as single cell RNA-seq data.

## 250

**Developing a high-throughput CRISPR-based assay for saturation mutagenesis of human genes.** *M.L. Carpenter, C. Lee, N. Hammond, A. Li, A. Adams, C.D. Bustamante, M.C. Bassik.* Department of Genetics, Stanford University, Stanford, CA.

One of the biggest challenges currently facing the clinical translation of whole-genome sequencing is our lack of knowledge about the functional impact of the majority of variants. Although many variant annotation tools have been developed to take advantage of characteristics such as conservation, amino acid properties, population frequency, and genic location (e.g., splice site, promoter), these predictions are rarely experimentally verified. Most functional testing usually occurs on a case-by-case basis, and it is often difficult to directly compare results between laboratories. In this study, we aim to develop an experimental method to simultaneously and consistently assay the impact of many mutations in a single gene. We will describe our implementation of a CRISPR/Cas9-based approach to create a population of human cells in which each cell harbors a different mutation in the same gene, *TP53*. These cells can be grown in bulk under an appropriate selective pressure and then sequenced to determine the abundance of each mutation. This abundance serves as a proxy for the pathogenicity of each mutation, and even has the potential to reveal unexpected effects—for example, pathogenic synonymous mutations. As a test case, we have used CRISPR/Cas9 to install a set of 5 known pathogenic and 5 benign mutations in the human gene *TP53* in the MCF7 breast cancer cell line, as well as 10 mutations categorized as variants of unknown significance (VUS). We have characterized the effects of these mutations on growth and DNA damage sensitivity in cells; these measurements will allow us to assess the sensitivity and specificity of our assay and will help set benchmarks for the future assessment of VUS. At the same time, we are developing technologies for improving the efficiency of producing single mutations in individual cells in large populations using the CRISPR system. Our experimental pathogenicity measures will eventually be incorporated into the new ClinGen resource, which is being developed by our group and others to serve as a central, curated repository for clinically relevant variant information.

## 251

**Transcriptome-wide nuclease-mediated protein footprinting to identify RNA-protein interaction sites.** *I. Silverman[1,2], F. Li[1], Q. Zheng[1], B. Gregory[1,2].* 1) Department of Biology; 2) Cell and Molecular Biology Graduate Group, University of Pennsylvania, Philadelphia, PA.

RNA-binding proteins (RBPs) are intimately involved in all aspects of RNA processing and regulation and are linked to neurodegenerative diseases and cancer. Therefore, investigating the relationship between RBPs and their RNA targets is critical for a broader understanding of post-transcriptional regulation in normal and disease processes. The majority of approaches to study RNA-protein interactions focus on individual RBPs. However, there are hundreds of these proteins encoded in the human genome, and each cell type expresses a different repertoire, greatly limiting the ability of current methods to capture the global landscape of RNA-protein interactions. To address this gap, we and others have recently developed methods to globally identify regions of RNAs that are bound by proteins in an unbiased manner. Here, we present our ribonuclease-mediated protein footprint sequencing approach, termed protein interaction profile sequencing (PIP-seq). We describe the application and validation of this protocol in multiple mammalian cell lines. We identify numerous putative RBP-binding motifs, reveal novel insights into co-binding by RBPs, and uncover a significant enrichment for disease-associated polymorphisms within RBP interaction sites. Finally, we use structure-specific nuclease digestion patterns generated by these methods to reveal the local RNA secondary structure at binding sites for several RBPs, including the double-stranded RNA binding protein and component of the microRNA processing machinery, DGCR8. Intriguingly our results suggest that highly structured regions of human mRNAs are targeted by the microprocessor complex, resulting in endonucleolytic cleavage and production of functional small RNAs. Our results offer insights into global patterns of RNA-protein interactions, reveal the structural contexts of RBP binding sites, and uncover a novel mechanism for microRNA machinery-mediated regulation. Future applications of this method to study the dynamics of RNA-protein interactions and RNA secondary structure in developmental and disease processes will help to uncover the role of RBPs in post-transcriptional regulation.

## 252

**Epigenome imputation leads to higher-quality datasets and helps improve GWAS interpretation.** *J. Ernst[1], A.K. Sarkar[2,3], L.D. Ward[2,3], M. Kellis[2,3].* 1) UCLA, Los Angeles, CA; 2) MIT, Cambridge, MA; 3) Broad Institute, Cambridge, MA.

Genotype imputation has become commonplace to predict unobserved genetic variants by leveraging the increasing availability of large reference panels. The field of epigenomics is now undergoing a similar transition, with thousands of reference epigenomes becoming available, presenting an analogous opportunity to exploit their highly correlated nature for prediction of unobserved epigenomic datasets, and to generate more robust versions of existing datasets. Here, we introduce epigenome imputation, and apply it to predict 4,315 high-resolution genome-wide signal maps, consisting of 31 histone marks, DNaseI, DNA methylation, and RNA-Seq across 127 tissue/cell types. Imputed signal tracks show strong concordance with observed signal, and surpass observed datasets in effective sequencing coverage, consistency, and correspondence with relevant gene annotations, even for tissue-restricted genes. Global discrepancy between observed and imputed data reveals low-quality experiments, while local discrepancies in high-quality datasets in some cases reveal locations of tissue-specific regulation. We also use imputed datsets to generate the most comprehensive prediction of chromatin state information to date, consisting of 25 chromatin states based on 12 imputed marks across 127 epigenomes. Imputed epigenomic data has important implications for interpreting genome-wide association studies. Across 108 traits, we find that chromatin states learned using imputed data significantly improve the power to detect functional enrichments of trait-associated loci in characterized active enhancer regions, increasing the number of significant cell type-trait pairs by approximately 30%. They also show improve enrichments for variants that are weakly-associated (below genome-side signficance): for Type 1 Diabetes for example, we find increased enrichment in enhancer regions, better distinction of disease-relevant cell types and regions, and reduced enrichment for spurious cell types and regions. We expect that our method, software implementation, and imputed datasets will be a valuable community resource and that epigenome imputation will become a widely-adopted complement to large-scale experimental mapping of epigenomic information.

## 253

**Conservation of mammalian trans-regulatory circuitry under high cis-regulatory turnover.** *A.B. Stergachis[1], S. Neph[1], R. Sandstrom[1], E. Haugen[1], A. Reynolds[1], M. Zhang[2], R. Byron[2], T. Canfield[1], S. Stelhing-Sun[1], K. Lee[1], R. Thurman[1], S. Vong[1], D. Bates[1], F. Neri[1], M. Diegel[1], E. Giste[1], D. Dunn[1], S. Hansen[1,2], A. Johnson[1], P. Sabo[1], M. Wilken[3], T. Reh[3], P. Treuting[4], R. Kaul[1,2], M. Groudine[2,5], M. Bender[5,6], E. Borenstein[1], J. Stamatoyannopoulos[1,2].* 1) Department of Genome Sciences, University of Washington, Seattle, WA, USA; 2) Department of Medicine, University of Washington, Seattle, WA USA; 3) Department of Biological Structure, University of Washington, Seattle, WA USA; 4) Department of Comparative Medicine, University of Washington, Seattle, WA, USA; 5) Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; 6) Department of Pediatrics, University of Washington, Seattle, WA, USA.

The anatomical body plan and its fundamental physiological axes have been highly conserved during the extended interval of mammalian evolution separating mice and humans, though only a fraction of the human genome evinces evolutionary constraint. To quantify cis- vs. trans-regulatory contributions to the evolution of mammalian regulatory programs, we performed extensive genomic DNase I footprinting of the mouse genome across 25 cell and tissue types, collectively defining >8.6 million TF occupancy sites on the mouse genome at nucleotide resolution. Here we show that mouse TF footprints encode a regulatory lexicon of >600 motifs that is >95% similar with those recognized in vivo by human TFs. However, only ~20% of mouse TF footprints have occupied human sequence orthologs. Despite substantial turnover of the cis-regulatory landscape around each TF gene, nearly half of the cross-regulatory connections between individual TF genes have been maintained in orthologous human cell types through innovated TF recognition sequences. Strikingly, the higher-level organization of mouse TF-to-TF connections into cellular network architectures is substantially identical with human. Our results suggest that evolutionary selection on mammalian gene regulation is targeted chiefly at the level of trans-regulatory circuitry.

## 254

**A regulatory DNA association study between autoimmune disease risk and variation in regulatory regions that are highly unique to adaptive immune cells.** *A. Madar[1], D. Chang[1], A.J. Sams[1], F. Gao[1], Y. Waldman[1,2], C. Van Hout[3], A.G. Clark[1,3], A. Keinan[1].* 1) Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY; 2) Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv, Israel; 3) Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY.

Autoimmune diseases arise from the abnormal response of adaptive immune cells (T- and B- lymphocytes) against body tissue. Variation in non-coding regulatory DNA is thought to be a major genetic contributor to multiple autoimmune diseases. Here, we integrate DNase-seq data (a high throughput technology to detect regulatory DNA marked by DNase1 hypersensitive sites, DHSs) from immune and other cell types, with genome-wide association studies of autoimmune diseases and, as controls, unrelated complex traits. We describe an approach that combines DNase-seq data from multiple cell types to quantify the level of specificity of DHSs to each cell type. Based on this new approach, we identified ~1.2 million nucleotides (0.04% of the genome) of regulatory DNA that is uniquely accessible in adaptive immune cell types. Our new quantitative approach greatly improves the detection of such regions compared to the more usual use of DHS data to infer a binary open/closed chromatin state. We show that these adaptive-immune-specific regulatory regions (but not regions specific to other cell types) selectively contribute to the risk of multiple autoimmune diseases, but not to other complex traits. For multiple sclerosis and rheumatoid arthritis, for instance, regulatory DNA that is most accessible in T regulatory cells is most associated with disease risk. Finally, we performed the first regulatory DNA association study (RDAS) of autoimmune diseases that considers only variants in or near adaptive-immune-specific regulatory regions, thereby reducing the multiple testing burden compared to GWAS. Associations that we discovered using a GWAS of a small number of individuals, and that were not discovered in the original GWAS, are highly replicable in more recent, larger studies, and lead to interpretable results for non-coding regulatory DNA. Our quantitative approach for detecting cell type-specific DHSs can generalize to many other applications. RDAS of trait-relevant cell types can facilitate new discoveries from GWAS, suggest new focus regions for trait specific array designs, and is particularly well tailored for the coming era of whole-genome sequencing based GWAS, as it can take advantage of the base-pair resolution offered by sequencing data, while avoiding the pitfalls of increasing the number of tests as a result of this resolution boost.

## 255

**Novel kernel methods for detecting gene-environment.** *K.A. Broadaway[1], R. Duncan[1], L.M. Almli[2], K.J. Ressler[2], B. Bradley[2,3], M.P. Epstein[1].* 1) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA; 2) Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA; 3) Atlanta VA Medical Center, Atlanta, GA.

The etiology of complex traits likely involves the effects of genetic and environmental factors, along with complicated interaction effects between them. Consequently, there has been interest in applying genetic association tests of complex traits that account for potential modification of the genetic effect by the presence of an environmental exposure. One can perform such an analysis using a joint test of gene and gene-environment interaction (GxE). GxE testing is recommended when a study of interactions is expected to provide evidence of an association between genotype and phenotype that would not be found if only the main effects of exposures were examined; for example, in a situation of 'complete' interaction, where a genotype has an effect on phenotype in the presence of an environmental exposure, but no effect in absence of the exposure. However, when the genotypic effect in the absence of environmental exposure is greater than zero, a main effect test is expected to rival or outperform the joint test. When GxE is suspected, an optimal association test would be one that remains powerful under a variety of models, ranging from those of strong GxE effect ('complete' interaction) to little or no GxE effect. To fill this demand, we have extended a kernel-machine based approach for association mapping to consider joint tests of gene and GxE by incorporating a garrote parameter into the kernel framework. The kernel-based approach to GxE testing is promising for several reasons. First, since multiple typed markers are likely to be in linkage disequilibrium with the causal variant, joint consideration of these variants will capture the effect of a true causal variant more effectively than independent testing. Second, grouping variants together into sets along the genome allows epistatic interactions within the gene to be implicitly considered in the association test. Third, the kernel approach readily allows for inclusion of covariates, such as principal components to account for population stratification. We illustrate the method using simulated data of continuous phenotypes. We show that our kernel-machine approach typically outperforms the traditional joint test under strong GxE models and further outperforms the traditional main-effect association test under less strict models of weak or no GxE effects. We also illustrate our test using genome-wide association data from the Grady Trauma Project.

## 256

**Gene-environment dependence creates spurious gene-environment interaction.** *F. Dudbridge[1], O. Fletcher[2,3].* 1) London School of Hygiene and TM, London, United Kingdom; 2) Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, UK; 3) Division of Breast Cancer Research, The Institute of Cancer Research, London, UK.

Gene environment interactions have the potential to shed light on biological processes leading to disease and to improve the accuracy of epidemiological risk models. However relatively few such interactions have yet been confirmed. In part this is because genetic markers such as tag SNPs are usually studied, rather than the causal variants themselves. Previous work has shown that this leads to substantial loss of power and increased sample size when gene and environment are independent. However, dependence between gene and environment can arise in several ways including mediation, pleiotropy and confounding, and several examples of gene environment interaction under gene environment dependence have recently been published. Here we show that under gene environment dependence, a statistical interaction can be present between a marker and environment even if there is no interaction between the causal variant and the environment. We give simple conditions under which there is no marker environment interaction and note that they do not hold in general when there is gene environment dependence. Furthermore, the gene environment dependence applies to the causal variant, and cannot be assessed from marker data. For example, an interaction recently reported between rs10235235 and age at menarche on the risk of breast cancer could be explained by a causal variant with minor allele frequency 2%, moderately strong effects on both disease and environment, but no interaction with environment. In addition to existing concerns about mechanistic interpretations, we suggest further caution in reporting interactions between genetic markers and environmental exposures.

## 257

**Discovery of gene-environment and epistatic interactions affecting gene expression in the TwinsUK cohort via association mapping of variance and monozygotic twin discordance.** *A. Brown[1,2,3], A. Buil[2], A. Viñuela[4], M. Davies[4], K. Small[4], T. Spector[4], E. Dermitzakis[2], R. Durbin[1].* 1) Wellcome Trust Sanger Institute, Cambridge, United Kingdom; 2) Dep. Genetic Medicine and Development, University of Geneva, Geneva, Switzerland; 3) NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway; 4) Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom.

The identification of non-additive interactions between genetic variants (epistasis), or between genetic variants and environment (GxE), can give insight into mechanism and contribute to our understanding of the genetic architecture of traits. Here we explore epistasis and GxE affecting gene expression, considered as a set of quantitative traits measured using RNA-seq, in fat, LCLs, skin and whole blood from the TwinsUK cohort (N=~850). Because epistasis and GxE affect trait variance in a genotype dependent fashion, we used the strategy of prioritising SNPs associated with variance in expression (v-eQTL) to look for interactions. Similarly, SNPs associated with discordance of expression between monozygotic twin pairs (d-eQTL) indicate presence of GxE. We found 1198 v-eQTL in LCLs, 620 in fat, 368 in skin and 39 in blood, and 73, 211, 63 and 1 d-eQTL in those tissues. A greater proportion of v-eQTL acted as d-eQTL (21-44% depending on tissue) than vice versa (0-29%), consistent with d-eQTL as a subset of v-eQTL induced by GxE. Functional overlap with ENCODE data shows LCL v-eQTL significantly depleted in transcriptionally repressed regions (odds ratio, 0.82) and enriched in enhancers (OR 1.71); d-eQTL are enriched in promoters (OR 5.29). Skin d-eQTL are enriched in H3K36me3 regions (OR 4.02), a mark of active transcription. To find environments involved in GxE signatures, we tested all v-eQTL and d-eQTL for interactions with age, BMI and 20 expression principal components (PCs), having previously seen PCs can be highly heritable. There were three Bonferroni significant interactions between genotype and BMI affecting fat expression ($p<1.94e-5$). There were many interactions with PCs: 2, 10, 39 and 66 in blood, fat, skin and LCLs ($p<9.70e-7$). Analysis of separate dermis and epidermis data suggested some skin d-eQTL are cell specific eQTL. Finally, as v-eQTL can be induced by epistasis, we scanned the cis window for SNPs interacting with v-eQTL. Initial results found epistasis in all tissues, frequently shared across tissues. We replicated epistasis found in all 4 tissues using RNA-seq LCL data from GEUVADIS samples: for 100%, 52%, 67% and 15% of interactions in blood, fat, LCLs and skin we observed $p<0.05$. In summary, we detect widespread variance and discordance effects in gene expression. V-eQTL provide a way to discover replicating epistasis while d-eQTL consistently have more success at mapping GxE with phenotypes, PCs and tissue composition measures.

## 258

**A statistical approach to distinguish genetic pleiotropy from clinical heterogeneity: application to autoimmune diseases.** *B. Han[1-3], D. Diogo[1-4], E.A. Stahl[5], S. Eyre[6,7], S. Rantapää-Dahlqvist[8], J. Martin[9], T.W. Huizinga[10], P.K. Gregersen[11], J. Worthington[6,7], L. Klareskog[12], P.I.W. de Bakker[13,14], S. Raychaudhuri[1-4,6].* 1) Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 2) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 3) Partners HealthCare Center for Personalized Genetic Medicine, Boston, MA 02115, USA; 4) Division of Rheumatology, Immunology, and Allergy, Brigham and Women's, Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA; 5) The Department of Psychiatry, Mount Sinai School of Medicine, New York, New York, USA; 6) Arthritis Research UK Epidemiology Unit, Musculoskeletal Research Group, University of Manchester, Manchester Academic Health Sciences Centre, Manchester M13 9PT, UK; 7) NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester NHS Foundation Trust, Manchester Academic Health Sciences Centre, Manchester M13 9PT, UK; 8) Department of Public Health and Clinical Medicine / Rheumatology, Umeå University, S-901 85 Umeå, Sweden; 9) Instituto de Parasitologia y Biomedicina Lopez-Neyra, Consejo Superior de Investigaciones Cientificas (CSIC), 18100 Armilla, Granada, Spain; 10) Department of Rheumatology, Leiden University Medical Centre, 2300 RC Leiden, The Netherlands; 11) The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, NY 11030, USA; 12) Rheumatology Unit, Department of Medicine, Karolinska Institutet and Karolinska University Hospital Solna, SE-171 76 Stockholm, Sweden; 13) Department of Epidemiology, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands; 14) Department of Medical Genetics, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands.

**Motivation:** Recent studies have demonstrated that many medically relevant phenotypes have a shared genetic structure. For example, many autoimmune diseases have shared alleles and exhibit cross-heritability, but it is uncertain whether this is the consequence of a common genetic basis (pleiotropy) or the consequence of clinical heterogeneity. Clinical heterogeneity occurs when a presumably phenotypically homogeneous patient cohort consists of genetically distinct subgroups, either (1) because different phenotypes were misclassified as one or (2) the diagnosed trait in a subset of individuals was caused by another trait. **Method:** We developed a novel statistical approach to distinguish genetic pleiotropy from clinical heterogeneity. Our method examines a patient cohort to assess if there is evidence of a subgroup that is enriched for the risk alleles for a second trait compared to the rest of the cohort. **Results:** Based on simulations, we demonstrate that our approach has >90% power to detect clinical heterogeneity with 50 risk alleles in 2,000 samples. We applied this approach to seronegative (CCP-) rheumatoid arthritis (RA), which is known to share genetic structure with seropositive (CCP+) RA. We examined 71 CCP+ RA risk alleles in 3,273 CCP- RA cases, and identified statistically significant clinical heterogeneity (P=0.003). Since the two RA subtypes are not causal to each other by definition, this was evidence of misclassifications. Specifically, our method suggested that 24% of CCP- RA cases were likely to be misclassified CCP+ RA patients, which was consistent with our previous observations that the shared genetic structure between the two RA subtypes might be in part attributable to misclassifications (Han *et al.* AJHG 2014). On the other hand, examining CCP+ RA risk alleles within WTCCC cases of Type 1 diabetes, also known to share genetic structure with RA, we observed no evidence of clinical heterogeneity (P=0.8), suggesting that these two conditions have true pleiotropic genetic effects. **Conclusions:** Our statistical approach effectively distinguishes pleiotropy from clinical heterogeneity. This is a key advantage compared to previous approaches to assess shared genetic structure, such as polygenic modeling or Mendelian randomization, which are both unable to make this distinction. Our method is widely applicable to misdiagnosis detection and causal inference between traits.

## 259

**Analysis of variants obtained through whole-genome sequencing provides an alternative explanation to apparent epistasis.** *A.R. Wood[1], M.A. Tuke[1], M. Nalls[2], D. Hernandez[2,3], S. Bandinelli[4,5], A. Singleton[2], D. Melzer[6], L. Ferrucci[7], T.M. Frayling[1], M.N. Weedon[1].* 1) Genetics of Complex Traits, University of Exeter Medical School, Exeter, United Kingdom; 2) Laboratory of Neurogenetics, National Institute of Aging, Bethesda, Maryland, USA; 3) Department of Molecular Neuroscience and Reta Lila Laboratories, Institute of Neurology, UCL, London, United Kingdom; 4) Tuscany Regional Health Agency, Florence, Italy, I.O.T. and Department of Medical and Surgical Critical Care, University of Florence, Florence, Italy; 5) Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy; 6) Institute of Biomedical and Clinical Sciences, University of Exeter Medical School, Barrack Road, Exeter, United Kingdom; 7) Longitudinal Studies Section, Clinical Research Branch, Gerontology Research Center, National Institute on Aging, Baltimore, Maryland, USA.

It has proven hard to find examples of “gene-gene” interaction (epistasis) in humans. The first evidence of epistasis affecting traits was recently described by Hemani *et al.* (Nature, 2014). They detected 30 pairwise interactions influencing gene expression levels that were replicated in additional studies. We sought further replication but used genotypes derived from low-pass whole-genome sequencing to capture more completely the variation around the putatively interacting variants. Using 450 unrelated individuals from the InCHIANTI study with genome-wide expression profiling captured on the Illumina HT-12 expression microarray from whole-blood, we replicated 14 of the reported pairwise interaction effects ($P<0.05$). However, in each case, a third variant captured by whole-genome sequencing could explain all of the apparent epistasis in our data. This third variant was often moderately correlated with each of the two putatively interacting variants, despite very low levels of correlation between the original pair. For example, evidence for putative interactions between pairs of SNPs in cis influencing *FN3KRP* ($P$=3×10-12), *CSTB* ($P$=8×10-07), *MBLN1* ($P$=3×10-06) and *NAPRT1* ($P$= 6×10-06) disappeared on correction for a single confounding sequenced variant (*FN3KRP* ($P$=0.43), *CSTB* ($P$=0.99), *MBLN1* ($P$=0.16) and NAPRT1 ($P$=0.84)) Our results provide an alternative explanation for the apparent epistasis observed for gene expression traits in humans.

## 260

**A joint testing framework uncovers paradoxical SNPs, improves power, and identifies new sources of missing heritability in association studies.** *B.C. Brown[1], N.A. Patsopoulos[2], A. Price[3,4], L. Pachter[1,6,7], N. Zaitlen[5].* 1) Computer Science Department, UC Berkeley, Berkeley, CA; 2) Department of Neurology, Brigham & Women's Hospital, Harvard Medical School Cambridge, MA; 3) Department of Epidemiology, Harvard University Cambridge, MA; 4) Department of Biostatistics, Harvard University Cambridge, MA; 5) Department of Medicine, UCSF San Francisco, CA; 6) Department of Mathematics, UC Berkeley Berkeley, CA; 7) Department of Molecular and Cell Biology, UC Berkeley Berkeley, CA.

Variants identified via GWAS of complex human phenotypes account for only a small fraction of the total heritability. While there are many proposed explanations for this missing heritability (Manolio et al Nat. 2009), an overlooked issue is that of “linkage masking” (LM), in which linkage disequilibrium between SNPs masks their signal under gold standard marginal tests of association, preventing their discovery in GWAS. Previous examinations of this phenomenon have focused only on known associated loci (Wood et al HMG 2011). In this work, we show that (1) LM is an instance of a Simpson’s paradox, where an effect is visible in subgroups but not in the population as a whole, (2) mixed-model based estimates of $h^2_g$ include their signal although GWAS may never find them, (3) intelligent joint testing without an interaction term will improve power in the presence of proximal causal variants, including masked SNPs, without a substantial increase in multiple testing burden.

Joint testing of SNPs has been under-utilized due to the immense computation time required and the large multiple testing penalty. We avoid these issues by using a sliding window approach wherein we perform joint tests only on markers with squared correlation exceeding a threshold R. We also provide a method for estimating the null distribution 500x faster than a permutation test, making application computationally efficient. We detail, via extensive simulation, the power gain/loss under different disease models, window sizes, R, and LD patterns. We find significant power gains when multiple causal variants are proximal, reaching as high as 32.1% in the case of LM. The increase in multiple hypothesis testing penalty is relatively minor for reasonable window sizes and R, preventing severe power loss when causal variants are distant. We applied our method to three WTCCC data sets (RA, CD, T1D) with a window size of 100 SNPs and R=0, discovering 47% more loci from the NHGRI database over the marginal test (22 vs 15 loci). For example, in RA rs2104286 has p-value 7e-06, but this drops to 3e-09 when jointly tested with rs1570527, revealing the later-discovered association at 10p15.1. Additionally, we find all classically discovered loci, lending further evidence to recent work suggesting most loci harbor multiple variants (Gusev et al PG 2013). In all, our framework provides evidence that joint testing can improve power and uncover sources of missing heritability.

## 261

**Valid permutation testing in the presence of polygenic variation.** *M. Abney.* Dept Human Gen, Univ Chicago, Chicago, IL.

This abstract discusses the difficulties in performing valid permutations to obtain an empiric null distribution when testing for quantitative trait loci in the presence of polygenic effects. Although permutation testing is a popular approach for determining statistical significance of a test statistic with an unknown distribution -- for instance, the maximum of multiple correlated statistics or some omnibus test statistic for a gene, gene-set or pathway -- naive application of permutations may result in an invalid test. The risk of performing an invalid permutation test is particularly acute in complex trait mapping where polygenicity may combine with a structured population, for instance due to the presence of families, cryptic relatedness, admixture or population stratification. I give both analytical derivations and a conceptual understanding of why typical permutation procedures fail and suggest an alternative permutation based algorithm that succeeds. In particular, I examine the case where a linear mixed model is used to analyze a quantitative trait and show that both phenotype and genotype permutations may result in an invalid permutation test. The problems are due to a lack of exchangeability that arises from confounding that exists between the genotype being tested and the polygenic effect. Based on analytical derivations I provide a metric that predicts the amount of inflation of the type 1 error rate in the empiric permutation distribution depending on the correlation structure of the polygenic effect in the sample and the heritability of the trait. I validate this metric by doing simulations, showing that the permutation distribution matches the theoretical expectation, and that my suggested permutation based test obtains the correct null distribution. Finally, I discuss situations where naive permutations of the phenotype or genotype are valid and the applicability of the results to other test statistics.

## 262

**Sparse Bayesian latent factor decompositions for identifying trans-eQTLs.** *V. Hore[1], J. Marchini[1,2].* 1) Department of Statistics, University of Oxford, Oxford OX1 3TG, UK; 2) Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK.

Expression quantitative trait loci (eQTL) mapping aims to uncover the role of genetic variation in gene regulation. Many approaches to eQTL mapping employ mass univariate regression between nearby genes and SNPs, and although these techniques have been successful in finding cis-eQTLs, it is harder to identify trans-eQTLs due to their non-local nature. These techniques do not explicitly model the joint effect of genetic variation across networks of genes, nor do they naturally extend to allow for simultaneous analysis of multiple tissues.

We have developed a method for identifying trans-eQTLs in multiple tissues by jointly modeling correlations between genes and tissues. Our approach is a general framework for decomposing matrices and tensors into sparse latent factors, where latent factors consist of networks of covarying genes. The model also determines the subset of tissues in which each latent factor is active. We fit our model using variational Bayes, which allows for relatively fast inference on large data sets, and imputation of missing data. In addition, the method is capable of integrating other information to inform correlations in gene expression, such as genotypes (both directly or through a kinship matrix) and measured covariates.

Via simulations we are able to demonstrate that the method can reliably identify trans-eQTLs, in the presence of cis-eQTLs, multiple confounding factors and realistic levels of experimental noise. We will also report results of applying our method to the GTEx Project data, which consists of gene expression in up to 30 tissues for each individual.

## 263

**Mutations in CNTNAP1 and ADCY6 are responsible for severe arthrogryposis multiplex congenita with axoglial defects.** *J. Melki[1], J. Maluenda[1], A. Camus[1], L. Fontenas[2], K. Dieterich[3], F. Nolent[1], N. Monnier[4], P. Latour[5], J. Lunardi[4], M. Bayes[6], PS. Jouk[3], S. Sternberg[7], J. Warszawski[8], I. Gut[6], M. Gonzales[9], M. Tawk[2], A. Laquérriere[10].* 1) Unité Mixte de recherche (UMR)-986, Inserm and University Paris 11, 94276 Le Kremlin Bicêtre, France; 2) Unité Mixte de recherche (UMR)-788, Inserm and University Paris 11, 94276 Le Kremlin Bicêtre, France; 3) Département de Génétique, CHU Grenoble, Inserm U-836, Institut des Neurosciences, 38043 Grenoble, France; 4) Laboratoire de Biochimie et Génétique Moléculaire, CHU Grenoble, 38043 Grenoble, France; 5) Service de Neurobiologie, CHU de Lyon, 69677 Bron, France; 6) Centro Nacional de Análisis Genómico, Barcelona, 080028, Spain; 7) Assistance Publique Hôpitaux de Paris, Hôpitaux Universitaires Pitié-Salpêtrière, Service de Biochimie Métabolique, 75651 Paris, France; 8) UMR-1018, Inserm et Université Paris 11, Service d'Epidémiologie-Santé Publique, CHU Bicetre, 94276 Le Kremlin-Bicêtre, France; 9) Service de Génétique et d'Embryologie Médicales, Université Paris VI, Hopital Trousseau, 75571 Paris, France; 10) Pathology Laboratory and NeoVasc Region-Inserm Team ERI28, Institute of Research for Innovation in Biomedicine, University of Rouen, 76031 Rouen, France.

Non-syndromic arthrogryposis multiplex congenita (AMC) is characterized by multiple congenital contractures resulting from reduced fetal mobility. Genetic mapping and whole exome sequencing were performed in 31 multiplex and/or consanguineous undiagnosed AMC families. We report pathogenic mutations in two new genes. Homozygous frameshift mutations in CNTNAP1 were found in four unrelated families. Patients showed a marked reduction in motor nerve conduction velocity (less than 10m/sec) and transmission electron microscopy (TEM) of sciatic nerve in the index cases revealed severe abnormalities of both nodes of Ranvier width and myelinated axons. CNTNAP1 encodes CASPR, an essential component of node of Ranvier domains which underly saltatory conduction of action potentials along myelinated axons, an important process for neuronal function. A homozygous missense mutation in Adenylate Cyclase 6 gene (ADCY6) was found in another family characterized by a lack of myelin in the Peripheral Nervous System (PNS) as determined by TEM. Morpholino knockdown of the zebrafish orthologs led to severe and specific defects in peripheral myelin in spite of the presence of Schwann cells. ADCY6 encodes a protein that belongs to adenylate cyclase family responsible for the synthesis of cAMP. Our data indicate an essential and so far unknown role of ADCY6 in PNS myelination likely through the cAMP pathway. Additional AMC families have been included and using the same approach, two new genes involved in the development of the neuromuscular system have been identified and will be presented.

## 264

**A mutation in TMTC2 reveals a new mechanism causing sensorineural hearing loss.** *M. Olivier[1], A. Indap[2], Y. Zhou[3], J.W. Kent Jr.[1], E. King[4], C.B. Erbe[4], R. Cole[5], J. Littrell[5], K. Merath[5], S. Mleziva[4], J. Jensen[4], L.S. Burg[4], F. Rüschendorf[6], J.E. Kerschner[4], G. Marth[2], N. Hübner[6], H.H.H. Göring[1], D.F. Friedland[4], W.-M. Kwok[3], C.L. Runge[4,7].* 1) Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX., USA; 2) Department of Biology, Boston College, Chestnut Hill, MA, USA; 3) Department of Pharmacology and Toxicology, Medical College of Wisconsin, Milwaukee, WI, USA; 4) Department of Otolaryngology and Communication Sciences, Medical College of Wisconsin, Milwaukee, WI, USA; 5) Biotechnology and Bioengineering Center, Medical College of Wisconsin, Milwaukee, WI, USA; 6) Max-Delbrück Centre for Molecular Medicine, Berlin-Buch, Germany; 7) Department of Anesthesiology, Medical College of Wisconsin, Milwaukee, WI, USA.

Sensorineural hearing loss (SNHL) is commonly caused by pathologies affecting cochlear structures or the auditory nerve. Over 32 independent genetic loci have been demonstrated to cause nonsyndromic autosomal dominant SNHL, and 190 mutations in 31genes have been described so far. Similarly, 53 loci have been identified for recessive forms of SNHL. Despite this large number, genes causing SNHL identified to date only explain a fraction of the overall genetic risk for this debilitating disorder. Here, we report on a six-generation family of Northern European descent with 18 individuals displaying bilateral, symmetric, progressive SNHL. Genome-wide exome sequencing identified a rare, fully penetrant variant in an uncharacterized gene, TMTC2, that segregates with SNHL in this family (Val381Ile, rs35725509, $p=6\times10^{-13}$). In contrast, no previously reported hearing loss mutations were identified in this family. Analysis of a cohort of unrelated individuals with SNHL (N=182) revealed that 3% also carried this same mutation, compared to 0.8% of individuals in the general population, as assessed by the NHLBI Exome Sequencing Project ($p=5\times10^{-6}$), making rs3572559 the third most common mutation causing autosomal dominant SNHL reported to date. The electrophysiological characterization of TMTC2 reveals that the Val381Ile mutation significantly accelerates the inactivation kinetics of ion channels when an expression plasmid with the coding sequence of TMTC2 carrying the Val381Ile variant is transiently transfected into HEK293 cells and compared to cells transfected with the TMTC2 wildtype expression plasmid. This suggests that TMTC2 is a regulatory protein modulating inner ear function and hearing, and is not an ion channel or structural protein as other genes previously implicated in SNHL. Such a functional role has not been described before for genes contributing to hearing loss, and may suggest a novel regulatory mechanism affecting auditory nerve function in normal hearing.

## 265

**Understanding Pathogenesis of Lissencephaly with Patient-Derived Induced Pluripotent Stem Cells.** *M. Bershteyn[1], A. Kriegstein[1], A. Wynshaw-Boris[2].* 1) Edyth and Eli Broad Institute of Regeneration Medi, University of California, San-Francisco School of Medicine, San Francisco, CA; 2) Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland OH, USA.

Normal development of the cerebral cortex requires a complex series of cellular events, including specification, proliferation, migration and differentiation, to establish the proper structure and function. Mutations that disrupt these key developmental processes give rise to cortical malformations. Miller Dieker Syndrome (MDS) is a genetic developmental disorder characterized by severe cortical malformations including reduced brain size (microcephaly) and nearly absent cortical folding (lissencephaly), with devastating neurological consequences such as mental retardation and intractable epilepsy. MDS is caused by heterozygous deletions of human band 17p13.3, harboring several dozen genes, including PAFAH1B1. Analyses of Pafah1b1 mutant mice revealed defects in neuronal migration, which is considered to be the main cellular deficiency in lissencephaly. However, it is unknown whether induction, proliferation or differentiation of neural stem cells are also disrupted in MDS, and the roles of most of the other deleted genes from locus 17p13.3 in cortical development or MDS pathogenesis have not been examined. To study the cellular and molecular mechanisms of MDS pathogenesis, we generated induced pluripotent stem cells (iPSCs) from MDS patients. We found that MDS iPSCs can proliferate, self-renew and generate all three germ layers in vitro and in vivo. Moreover, using 2-D and 3-D in vitro differentiation methods, we observed efficient induction of neuroepithelial stem cells and radial glia, exhibiting characteristic mitotic behaviors such as interkinetic nuclear migration and mitotic somal translocation. These results suggest that haploinsufficiency for locus 17p13.3 does not grossly impair these key features of cortical development. However, upon high-density culture conditions, MDS neural progenitors exhibited increased apoptosis. Our results so far suggest that reduced viability of neural stem cells may be a primary factor in MDS pathogenesis that precedes any defects in neuronal migration. Additional studies are underway to characterize other aspects of human cortical development such as proliferation of various types of progenitors, migration and differentiation.

## 266

**Hypomorphic PCNA mutation underlies a novel human DNA repair disorder.** *E.L. Baple[1], H. Chambers[2], H.E. Cross[3], H. Fawcett[4], Y. Nakazawa[5,6], B.A. Chioza[1], G.V. Harlalka[1], S. Mansour[7], A. Sreekantan-Nair[1], M.A. Patton[1], M. Muggenthaler[1], P. Rich[8], K. Wagner[9], R. Coblentz[9], C.K. Stein[10], .J.I. Last[11], A.M.R. Taylor[11], A.P. Jackson[12], T. Ogi[5,6], A.R. Lehmann[4], C.M. Green[2,13], A.H. Crosby[1].* 1) Medical Research, RILD Wellcome Wolfson Centre, University of Exeter, Exeter, Exeter, United Kingdom; 2) University of Cambridge, Cambridge, United Kingdom; 3) Department of Ophthalmology, University of Arizona College of Medicine, Tucson, Arizona, USA; 4) Genome Damage and Stability Centre, University of Sussex, Falmer, Brighton, United Kingdom; 5) Nagasaki University Research Centre for Genomic Instability and Carcinogenesis (NRGIC), Nagasaki, Japan; 6) Department of Molecular Medicine, Atomic Bomb Disease Institute, Nagasaki University, Nagasaki, Japan; 7) SW Thames Regional Genetics Service, St. George's Healthcare NHS Trust, London, United Kingdom; 8) Department of Neuroradiology, St. George's Hospital, London, United Kingdom; 9) Windows of Hope Genetic Study, Walnut Creek, Ohio, USA; 10) SUNY Upstate Medical University, Syracuse, New York, USA; 11) School of Cancer Sciences, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom; 12) MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom; 13) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.

A number of human disorders, including Cockayne syndrome, UV-sensitive syndrome, xeroderma pigmentosum and trichothiodystrophy, result from the mutation of genes encoding molecules important for nucleotide excision repair. We describe a novel DNA repair disorder identified amongst the Ohio Amish community. The cardinal clinical features of this disorder include postnatal growth retardation, hearing loss, premature aging, telangiectasia, neurodegeneration, photophobia, photosensitivity and predisposition to sun induced malignancy. Assuming autosomal recessive inheritance and that a founder mutation was responsible for the condition, we used a combination of autozygosity mapping and linkage analysis to identify the underlying molecular cause. Our genetic investigation identified a homozygous missense (p.Ser228Ile) sequence alteration of the proliferating cell nuclear antigen (PCNA) associated with the disease phenotype. PCNA is a highly conserved sliding clamp protein essential for DNA replication and repair. Due to this fundamental role, mutations in PCNA that profoundly impair protein function would be incompatible with life. Interestingly, while the p.Ser-228Ile alteration appears to have no effect on protein levels or DNA replication, patient cells exhibit significant abnormalities in response to UV irradiation displaying substantial reductions in both UV survival and RNA synthesis recovery. Importantly the defective transcriptional responses to UV light are completely rescued by wild type PCNA molecule, demonstrating that the mutation is causative. Furthermore we show that the p.Ser228Ile change profoundly alters the capacity of PCNA to interact with a specific subset of partner proteins. These proteins include XPG and the DNA metabolism enzymes Flap endonuclease 1 and DNA Ligase 1, molecules fundamental to genomic integrity, thus providing an explanation for the separation of function effect. Taken together our findings detail the first mutation of PCNA in humans, associated with a unique neurodegenerative disease displaying clinical and molecular features common to other DNA repair disorders, which we show to be attributable to a hypomorphic amino acid alteration. Further investigation of the altered biological processes underlying this syndrome should provide valuable insight into the neurodegenerative disease mechanisms involved in DNA damage tolerance and repair disorders.

## 267

**Profound Neuropathy Target Esterase impairment results in Oliver-McFarlane syndrome.** *R.B. Hufnagel[1], G. Arno[2], N.D. Hein[3], J. Hersheson[4], L.A. Krueger[1], T.J. Jaworek[5], L.C. Gregory[4], S. Hull[2], V. Plagnol[6], C.M. Willen[7], T.M. Morgan[8], C.A. Prows[1], R.S. Hegde[9], S. Riazuddin[10], G.A. Grabowski[1], R.J. Richardson[11], J.P. Martinez-Barbera[4], T. Huang[1], M.T. Dattani[12], R.A. Sisk[5], H. Houlden[4], J.K. Fink[3], A.T. Moore[2], Z.M. Ahmed[1,5].* 1) Human Genetics, Cincinnati Children's Hospital, Cincinnati, OH; 2) Ophthalmology, Moorfields Eye Hospital, London, UK; 3) Neurology, University of Michigan, Ann Arbor, MI; 4) Neurology, University College London, London, UK; 5) Pediatric Ophthalmology, Cincinnati Children's Hospital, Cincinnati, OH; 6) Statistical Genetics, University College London, London, UK; 7) Ophthalmology, University of Kentucky, Lexington, KY; 8) Human Genetics, Vanderbilt University, Nashville, TN; 9) Developmental Biology, Cincinnati Children's Hospital, Cincinnati, OH; 10) Pediatric Otolaryngology, Cincinnati Children's Hospital, Cincinnati, OH; 11) Environmental Health Sciences, University of Michigan, Ann Arbor, MI; 12) Endocrinology, University College London, London, UK.

A half century mystery elucidated. Oliver-McFarlane syndrome [MIM 275400], first described in 1965, is a rare disorder with the triad of congenital trichomegaly, chorioretinal atrophy, and hypopituitarism. In our patients, thyroid and growth hormone replacement during childhood successfully improved the pituitary sequelae of intellectual disability and short stature. However, vision loss was progressive and devastating. Designing therapeutic approaches first requires discovery of the genetic and biologic mechanisms. In this study, through whole exome sequencing, we identified seven novel mutations in the *PNPLA6* gene in five patients with Oliver-McFarlane syndrome. *PNPLA6* (19p13.2) encodes Neuropathy Target Esterase (NTE), a phospholipase that hydrolyzes phosphatidylcholine and maintains axonal integrity. Previously, recessive alleles of *PNPLA6* have been reported for two adult onset neurologic disorders, Spastic Paraplegia type 39 (SPG39) and Boucher-Neuhäuser syndrome. Oliver-McFarlane syndrome is distinguished from these by the congenital onset, trichologic findings, childhood pituitary sequelae, and ocular disease severity. In accord with the clinical findings, we found *PNPLA6* expression in the developing human eye, pituitary, and brain. Molecular modeling suggested impaired function of the NTE patatin domain in the Oliver-McFarlane alleles. Concurrently, we confirmed this hypothesis in two model systems. In zebrafish, the *pnpla6* curly-tailed morphant phenotype was fully rescued by wild type human *PNPLA6* mRNA but not with Oliver-McFarlane or SPG39 mutation-harboring human mRNAs. Second, NTE enzymatic activities were measured in patient fibroblast cell lines to determine if the congenital onset of Oliver-McFarlane correlates with reduced hydrolase activity compared to adult onset SPG39. Indeed, NTE activity was significantly decreased in individuals with Oliver-McFarlane (≤25% of normal activity) compared to SPG39 (73% of normal activity). In conclusion, Oliver-McFarlane syndrome is caused by *PNPLA6* mutations that result in early and severe loss of NTE function. Identification and functional characterization of this new *PNPLA6*-opathy reveals a broad spectrum of neurodevelopmental and neurodegenerative disorders caused by NTE impairment, which appears to determine both age of onset and affected tissues in a dose-dependent manner. Ongoing studies will focus on the relationship between NTE activity and disease prognosis, as well as therapeutic strategies.

## 268

**A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement.** *V. Paquis[1,2], S. Bannwarth[1,2], S. Ait-El-Mkadem[1,2], A. Chaussenot[1,2], E.C. Genin[1], S. Lacas-Gervais[3], K. Fragaki[1,2], L. Berg-Alonso[1], Y. Kageyama[4], V. Serre[5], D.G. Moore[6], A. Verschueren[7], C. Rouzier[1,2], I. Le Ber[8,9], G. Augé[1,2], C. Cochaud[2], F. Lespinasse[1], K. N'Guyen[10], A. de Septenville[8], A. Brice[8], P. Yu-Wai-Man[6], H. Sesaki[4], J. Pouget[7].* 1) Department of Medical Genetics, IRCAN, UMR CNRS 7284/INSERM U1081/UNS, Nice, France; 2) Department of Medical Genetics, National Centre for Mitochondrial Diseases, Nice Teaching Hospital, France; 3) Joint Center for Applied Electron Microscopy, Nice Sophia-Antipolis University, France; 4) Department of Cell Biology, Johns Hopkins University School of Medicine, Baltimore, MD, 21205, USA; 5) UMR7592 CNRS, Jacques Monod Institute, Paris Diderot University, France; 6) Wellcome Trust Centre for Mitochondrial Research, Institute of Genetic Medicine, International Centre for Life, Newcastle University, Newcastle upon Tyne NE1 3BZ, UK; 7) Department of Neurology, Timone Hospital, Marseille Teaching Hospital, France; 8) Sorbonne Université, UPMC Univ Paris 06, UM75, Inserm U1127, Cnrs UMR7225, Institut du Cerveau et de la Moelle épinière (ICM), F-75013 Paris, France; 9) National Reference Centre on Rare Dementias, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Paris, France; 10) Department of Medical Genetics, Timone Hospital, Marseille Teaching Hospital, France.

Mitochondrial DNA (mtDNA) instability disorders are responsible for a large clinical spectrum, among which amyotrophic lateral sclerosis-like symptoms and frontotemporal dementia are extremely rare. We report a large family with a late-onset phenotype including motor neuron disease, cognitive decline looking like frontotemporal dementia, cerebellar ataxia and myopathy. In all patients, muscle biopsy showed ragged-red and COX negative fibres with combined respiratory chain deficiency and abnormal assembly of complex V. The multiple mtDNA deletions found in skeletal muscle revealed a mtDNA instability disorder. Patient fibroblasts present with respiratory chain deficiency, mitochondrial ultrastructural alterations and fragmentation of the mitochondrial network. Interestingly, expression of matrix-targeted photoactivable GFP showed that mitochondrial fusion was not inhibited in patient fibroblasts. By whole-exome sequencing (WES), we identified a missense mutation (c.176C>T; p.Ser59Leu) in the CHCHD10 gene that encodes a coiled-coil helix coiled-coil helix protein, whose function is unknown. We show that CHCHD10 is a mitochondrial protein located in the intermembrane space and enriched at cristae junctions. Overexpression of CHCHD10 mutant allele in HeLa cells led to fragmentation of the mitochondrial network and ultrastructural major abnormalities including loss, disorganization and dilatation of cristae. The observation of a frontotemporal dementia-amyotrophic lateral sclerosis (FTD-ALS) phenotype in a mitochondrial disease led us to analyse CHCHD10 in a cohort of 21 families with pathologically proven FTD-ALS. We identified the same missense p.Ser59Leu mutation in one of these FTD-ALS families. This work opens a novel field to explore the pathogenesis of FTD-ALS clinical spectrum by showing that mitochondrial disease may be at the origin of some of these phenotypes.

## 269

**Comprehensive investigation of *CASK* and other relevant genes in 41 patients with intellectual disability, microcephaly and disproportionate pontine and cerebellar hypoplasia (MICPCH) using next-generation sequencing.** *S. Hayashi[1], N. Okamoto[2], J. Takanashi[3], J. Inazawa[1].* 1) Dept Molec Cytogenetics, Tokyo Med & Dental Univ, Tokyo, Japan; 2) Dept. Planning and Research, Osaka Medical Center and Research Institute for Maternal and Child Health, Osaka, Japan; 3) Department of Pediatrics, Kameda Medical Center, Chiba, Japan.

The *CASK* gene [OMIM: *300172] at Xp11.4, encoding a member of the MAGUK (membrane-associated guanylate kinase) proteins, is highly expressed in the mammalian central nervous system of both adults and fetuses and plays several roles in neural development and synaptic function. While loss of function of *CASK* raised by mutation or genomic copy-number variant (CNV) causes intellectual disability and microcephaly with pontine and cerebellar hypoplasia (MICPCH) [OMIM: #300749] mostly in females, insufficiency of *CASK* probably leads to lethality in males. We reported a first case of MICPCH with heterozygous deletion at Xp11.4p11.3 including *CASK* in 2008, and thereafter we have recruited 41 patients presenting with typical MICPCH in order to investigate *CASK* aberrations. So far, we have detected various types of *CASK* aberrations in 27 of 41 patients (65.9%): large deletions in 6 patients, intragenic duplication or complex rearrangement in 3 patients, point mutations in 18 patients and other aberrations in 3 patients. Subsequently, we have investigated the *CASK*-negative cases using next-generation sequencing. We screened the remaining 14 cases by target-resequencing of all exons of 17 selected genes responsible for microcephaly and/or pontocerebellar hypoplasia, along with the entire region of *CASK* including all exons, introns and promoter region, and we identified novel candidate mutations in two cases. Also, we screened a trio of a patient and their parents by whole exome sequencing. Our research comprehensively clarified a correlation between these genotypes and phenotypes of MICPCH, and suggested that not only *CASK* but also other several genes are involved in the etiology of MICPCH.

## 270

**ABAT is a novel human mitochondrial DNA depletion syndrome gene linking gamma-aminobutyric acid (GABA) catabolism and mitochondrial nucleoside metabolism.** *P. Bonnen[1], A. Besse[1], P. Wu[1], F. Bruni[2], T. Donti[1], B. Graham[1], W. Craigen[1], R. McFarland[2], P. Moretti[1], S. Lalani[1], K. Scott[1], R. Taylor[2].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Wellcome Trust Centre for Mitochondrial Research, Institute of Neuroscience, The Medical School, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK.

ABAT is a key enzyme responsible for catabolism of principal inhibitory neurotransmitter gamma-aminobutyric acid (GABA) in the mitochondrial matrix. We report a novel role for ABAT in a seemingly unrelated pathway, mitochondrial nucleoside salvage, and demonstrate that mutations in this enzyme cause neurometabolic dysfunction and are a novel cause of mtDNA depletion syndrome (MDS). MDS is a group of autosomal recessive disorders that share the hallmark of decreased copy number of the mitochondrial genome and clinically manifest as encephalopathy, encephalomyopathy or hepatocerebral. Whole exome sequencing of a child with severe psychomotor retardation, intractable seizures, hypotonia and hyperreflexia revealed a homozygous missense mutation in ABAT. Compromised neurometabolic activity was confirmed with in vivo proton magnetic resonance spectroscopy showing significantly increased levels of GABA in the subject's brain. Muscle biopsy exhibited abnormal mitochondrial morphology and global decrease of electron transport chain activity motivating investigations into additional potential roles for ABAT in mitochondrial function. RNAi-mediated inhibition of ABAT in fibroblasts caused depletion of mtDNA. By employing a first-in-kind lentiviral vector to simultaneously express ABAT-UTR-specific shRNA hairpins and mutant ABAT cDNA-sans-UTR in healthy cells we demonstrated that all ABAT mutations in known subjects cause mtDNA depletion. Nucleoside rescue and co-IP experiments pinpoint that ABAT functions in the mitochondrial nucleoside salvage pathway to facilitate conversion of dNDPs to dNTPs, binding other proteins (SUCLA2, SUCLG1, SUCLG2, and NME4) previously thought to participate in this process. This work reveals ABAT as a connection between GABA metabolism and nucleoside metabolism and defines a novel genetic cause of MDS.

## 271

**GWAS meta-analysis of ten studies identifies five novel loci associated with gallstone disease in European ancestry individuals.** *A.D. Joshi[1], C. Andersson[2], S. Buch[3], M. Gala[4], R. Noordam[5], A. Teumer[6], S. Stender[7], B.G. Nordestgaard[7], L. Weng[8], A.R. Folsom[8], P.L. Lutsey[8], D. Ellinghaus[9], W. Lieb[10], C. Shafmayer[11], B. Boehm[12], A. Tybjærg-Hansen[7], U. Völker[13], H. Völzke[6], L. Rose[14], P.E. Weeke[15], D.M. Roden[15], J.C. Denny[15], W. Tang[8], B.H. Stricker[5], J. Hampe[3], D.I. Chasman[14], A.D. Johnson[2], A.T. Chan[4,16].* 1) Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, MA; 2) The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA; 3) Department of General Internal Medicine, University Hospital Schleswig-Holstein, Kiel, Germany; 4) Division of Gastroenterology and Hepatology, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA; 5) Department of Internal Medicine and Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands; 6) Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany; 7) Copenhagen University Hospitals and Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; Department of Clinical Biochemistry, Rigshospitalet, Copenhagen, Denmark; 8) Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, MN; 9) nstitute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany; 10) Institute of Epidemiology, Christian Albrechts Universität Kiel, Niemannsweg 11, Kiel, Germany; 11) Department of General, Abdominal, Thoracic and Transplantation Surgery, University of Kiel, Kiel, Germany; 12) Department of Internal Medicine I, Ulm University Hospital, Ulm, Germany; 13) Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Germany; 14) Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 15) Department of Medicine, Vanderbilt University, Nashville, TN; 16) Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.

In 2007, a modestly sized genome wide association study (GWAS) consisting of 280 cases identified the hepatic cholesterol transporter *ABCG8* locus as a susceptibility factor for human gallstone disease. To date, no other population-based GWAS has been reported for this phenotype. Therefore, we performed a large-scale GWAS meta-analysis of 9,298 gallstone cases and 62,007 controls of European ancestry pooled from 8 cohort and 2 case-control studies to discover additional genetic variants associated with gallstone disease. In the case-control sets nested within each discovery study described in Table 1, we used age and sex adjusted logistic regression models to obtain effect size estimates. Inverse variance weighted meta-analysis of study-specific estimates was performed and 91 SNPs were observed to be individually associated with gallstone disease at a genome-wide significance level ($p < 5\times10^{-8}$) from a total of 5 distinct genomic regions. Conditional analyses were performed in genomic regions (10 mega-base windows around SNPs) that showed nominal significance ($p < 5\times10^{-6}$) at the meta-analyses stage to identify SNPs that were independently associated with gallstones disease after adjusting for the top-hits in the region. After conditional analyses, two SNPs in *ABCG8*, *rs11887534* ($p=7.2\times10^{-53}$) and *rs4245791* ($p=4.0\times10^{-31}$), were independently associated with gallstone disease. Additionally, we identified novel independent associations for *rs1025447* ($p=4.7\times10^{-11}$) in *DYNC2LI1*, which plays a structural role in cilia formation; *rs9843304* ($p=3.3\times10^{-9}$) in *TM4SF4*; *rs2547231* in *SULT2A1*, which is a sulfo-conjugation enzyme that act on hydroxysteroids and cholesterol-derived sterol bile acids ($p=1.9\times10^{-9}$); and *rs1260326* ($p=1.4\times10^{-8}$) in *GCKR*, a glucokinase regulator. A borderline genome-wide significant association was observed for *rs6471717* ($p=1.09\times10^{-7}$) near the *CYP7A1* gene that codes for an enzyme to catalyze formation of bile salts from cholesterol. Replication was performed for six SNPs in five loci (except the *DYNC2LI1* locus), and all of them were observed to be associated with gallstone disease in independent replication samples ($p < 5\times10^{-6}$ ; 6,489 cases and 66,366 controls) from three studies (Table 1). In this large-scale GWAS meta-analysis of gallstone disease, we identified five biologically plausible novel loci, which have putative functions in cholesterol transport and metabolism, cilia structure/bile flow, and sulfonylation of bile acids/hyrdoxysteroids.

Table 1. Sample sizes of discovery and replication studies included in the meta-analysis

| **Discovery studies** | **Study Design** | **Cases** | **Controls** |
|---|---|---|---|
| Women's Genome Health Study (WGHS) | Nested case control study | 2,853 | 20,436 |
| Nurses' Health Study (NHS-1/2) and Health Professionals Follow-Up Study (HPFS) | Nested case control study | 2,472 | 10,155 |
| Study of Health in Pomerania (SHIP) | Nested case control study | 843 | 3,134 |
| Atherosclerosis Risk In Communities (ARIC) prevalence study | Case-control study (prevalent cases) | 832 | 8,032 |
| Rotterdam Study | Prospective cohort study | 705 | 5,269 |
| ARIC incidence study | Longitudinal cohort study | 687 | 7,311 |
| Framingham Heart Study (FHS) | Nested case control study | 515 | 3,783 |
| BioVU - (Vanderbilt University) | Hospital-based case-control study | 202 | 2,542 |
| SPC (PopGen cohort) | Nested case control study | 122 | 527 |
| SHIP-TREND (Germany) | Nested case control study | 67 | 818 |
| All discovery samples | | 9,298 | 62,007 |
| **Replication studies** | **Study Design** | **Cases** | **Controls** |
| Copenhagen City Heart Study and Copenhagen General Population Study | Nested case-control study | 3,599 | 60,958 |
| Kiel University | Hospital-based case-control study | 2,104 | 2,225 |
| NHS1/HPFS-replication set | Nested case control study | 786 | 3,183 |
| All replication samples | | 6,489 | 66,366 |
| **Combined discovery+ replication** | | 15,787 | 128,373 |

.

## 272

**Association analyses of 100,720 individuals reveal new loci associated with body fat percentage providing new insights in related cardiometabolic traits.** *Y. Lu[1], F. Day[2], S. Gustafsson[3], T. Kilpeläinen[4], R. Loos[1,2] on behalf of the Genetics of Body Fat Consortium.* 1) The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 2) MRC Epidemiology Unit, University of Cambridge, Institute of Metabolic Science, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ, UK; 3) 3, Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala 75185, Sweden; 4) Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen 2100, Copenhagen, Denmark.

Large-scale meta-analyses of genome-wide association studies (GWAS) for readily-available adiposity measures, such as BMI, waist-to-hip ratio adjusted for BMI (WHRadjBMI) and obesity risk have identified at least 75 loci that contribute to body weight and fat distribution in adults and children of diverse ancestry. While these commonly studied adiposity traits are easily assessed in large populations and thus allow statistically well-powered meta-analyses, they represent heterogeneous phenotypes and do not distinguish between lean and fat mass. To increase our understanding of the genetic basis of adiposity and its links to cardiometabolic disease risk, we conducted a genome-wide association meta-analysis of body fat percentage (BF%), which more accurately assesses adiposity. In our primary meta-analysis, we combined the results of genetic associations with BF% for up to 100,720 individuals from 43 GWAS (n up to 76,138) and 13 MetaboChip studies (n up to 24,582), which were predominantly of European ancestry (n up to 89,300). In secondary analyses, we stratified by sex and/or ancestry. For loci that reached genome-wide significance ($P<5\times10^{-8}$), we examined their association with cardiometabolic traits. SNPs in 12 loci reached genome-wide significance. Two (near *IRS1*, *SPRY2*) of the 12 loci had been identified in previous GWAS for BF%, and six (in or near *FTO*, *MC4R*, *TMEM18*, *TOMM40*, *TUFM/SH2B1*, and *SEC16B*) have previously been reported for BMI. Four of the 12 loci, near *GRB14/COBLL1*, *IGF2BP1*, *PLA2G6*, and *CRTC1*, were novel. SNPs in the 12 established loci increase body fat percentage by 0.24 to 0.51 SD/allele, explaining 0.58% of the BF% variance (between 0.03- 0.13% per locus). Some loci had association signatures with other cardiometabolic traits that were discordant with observed phenotypic correlations. E.g. the BF% increasing allele of the *COBLL1/GBR14* locus was associated with reduced WHRadjBMI, an improved lipid profile, and decreased risk of type 2 diabetes (T2D). The BF%-increasing allele of the *PLA2G6* locus was associated with reduced triglyceride levels and that of the *TOMM40/APOE* locus with reduced risk of cardiovascular disease, but increased risk of T2D. Our meta-analysis for BF% identifies novel loci, and reveals patterns of association that suggest a role of peripheral mechanisms involved in adipocyte and lipid metabolism and insulin sensitivity, complementing the central nervous pathways that are highlighted in GWAS for BMI and obesity risk.

## 273

**Genome-Wide Analysis in Africans Provides Novel Insight into the Genetic Basis of the Metabolic Syndrome.** *F. Tekola-Ayele, A.P. Doumatey, G. Chen, D. Shriner, A.R. Bentley, J. Zhou, A. Adeyemo, C.N. Rotimi.* Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD., USA.

The metabolic syndrome (MetS) is a constellation of heritable risk factors that increase the risk of developing several metabolic disorders including type 2 diabetes and cardiovascular diseases. Understanding the genetic basis of MetS is likely to provide insight into the biological pathways and mechanisms shared by the components of the MetS. We conducted a genome-wide association study (GWAS) of MetS in 1,427 West Africans (602 MetS and 825 non-MetS) recruited from Ghana and Nigeria, followed by replication testing and meta-analysis in East Africans recruited from Kenya as participants in the Africa America Diabetes Mellitus study. MetS status was assigned to each sample based on the definition of the National Cholesterol Education Program. A continuous MetS score (cMetS) was assigned to each sample based on the sum of the standardized residuals of the MetS components. Samples were genotyped on the Affymetrix Axiom® PANAFR SNP array that contains ~2.2 million SNPs. Imputed dosage data were analyzed using logistic regression model with adjustment for age, sex, and the first three principal components. We found a low-frequency (1.6%) variant near CA10 that confers a strong risk of MetS in the discovery sample (P=3.86×10-8, OR= 6.80). This variant had the same frequency in the 1000 genomes West Africa population samples, but was absent in all other population samples including East Africans. In meta-analysis of the West and East African samples, we found two variants that reduce risk of MetS: an African population-specific, low-frequency (1.7%) variant in CTNNA3 (P=1.63×10-8, OR= 0.35) and a common variant (46.4%) near RALYL (P=7.37×10-9, OR= 0.11). Analyses of the samples in the two extreme 33.3% and 25% tails of the empirical distribution of cMetS identified two variants that reduce risk of MetS: one near KSR2 (P=4.52×10-8, Pmeta=7.82×10-9, OR=0.53, allele frequency=30%), and an African population-specific, low-frequency (4%) variant near MBNL1 (Pmeta=3.51×10-8). In all, this first GWAS of MetS identified five loci that may have pleiotropic effects in several metabolic and cardiovascular diseases. The study also showed variants present only in African populations influencing risk of MetS, suggesting differences in genetic architecture of MetS among human populations, and the significance of studying ancestrally diverse populations to identify novel genetic variants.

## 274

**Contribution of low-frequency variants to variation in body mass index (BMI).** *V. Turcot[1,2], Y. Lu[3], J. Czajkowski[4], H.M. Highland[5], N.G.D. Masca[6], A. Giri[7], T.L. Edwards[7], T. Esko[8,9], M. Graff[10], A.E. Justice[10], C. Medina-Gomez[11], C. Schurmann[3], R.A. Scott[12], K. Sin Lo[1], S.S. Sivapalaratnam[13,14], L. Southam[15,16], K. Stirrups[15], T.W. Winkler[17], H. Yaghootkar[18], K.L. Young[10], A.L. Cupples[19], T.M. Frayling[18], J.N. Hirschhorn[20,21,22], G. Lettre[1,2], C.M. Lindgren[23], K.E. North[10,24], I.B. Borecki[4], R.J.F. Loos[3] For the BBMRI, the GOT2D, the CHARGE, and the GIANT Consortia.* 1) Montreal Heart Institute, Montréal, Québec, Canada; 2) Faculty of Medicine, Université de Montréal, Montréal, Québec, Canada; 3) The Genetics of Obesity and Related Metabolic Traits Program, The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 4) Department of Genetics Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA; 5) Human Genetics Center, University of Texas Health Science Center, Houston, TX, USA; 6) NIHR Leicester Cardiovascular Biomedical Research Unit, University of Leicester, Leicester, UK; 7) Center for Human Genetics Research, Division of Epidemiology, Department of Medicine, Vanderbilt University, Nashville, TN, USA; 8) Estonian Genome Center, University of Tartu, Tartu, Estonia; 9) Children's Hospital Boston & Broad Institute, MA, USA; 10) Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; 11) Netherlands Consortium for Healthy Aging (NCHA), Departments of Epidemiology and Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; 12) MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK; 13) Academic Medical Centre, Amsterdam, The Netherlands; 14) CGHR, Boston, MA, USA; 15) The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; 16) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 17) Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany; 18) University of Exeter Medical School, Exeter, UK; 19) Boston University School of Public Health, Boston, MA, USA; 20) Divisions of Endocrinology and Genetics and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA; 21) Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA; 22) Department of Genetics, Harvard Medical School, Boston, MA, USA; 23) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; 24) Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Genome-wide association studies (GWAS) have identified >90 loci for BMI. While low-frequency exonic variants are known to cause extreme and early-onset obesity, little is known about their role in obesity susceptibility in the general population. To estimate the contribution of low-frequency (MAF<5%) variants to variation in BMI, we performed a meta-analysis of exome array data in up to 249,395 individuals of predominantly European descent from 79 studies. Each study tested up to 246,127 autosomal and 5,072 X-chromosomal single nucleotide variants (SNVs) for association with inverse normally transformed residuals of BMI, adjusted for age and sex. Study-specific association results were combined using inverse variance-weighted meta-analysis and associations were considered significant if P<5×10-7. Three low-frequency autosomal SNVs were significantly associated with BMI. A rare SNV with a large effect on BMI was found in an unknown protein-coding gene (KIAA0754, MAF: 0.04%, P=4.8×10-7, effect size±SE: 0.60±0.12 SD/minor allele [≈2.5 kg.m-2, or 7.2 kg for a 1.7m-tall person]). Two SNVs were located in genes (GPR61, ZBTB7B) also harboring common SNVs associated with BMI. The SNV in GPR61 (3.7%, P=2.8×10-21, 0.08±0.01SD/MA) is intronic, but flanked by a more common coding SNV (6.6%, P=1.3×10-10, 0.04±0.01SD/MA) in SYPL2, and an intronic SNV (4%) in GNAT2 that was previously GWAS-identified. While fine-mapping will be needed to identify the causal gene in this locus, of interest is that GRP61-deficient mice exhibit obesity and hyperphagia. The second low-frequency SNV is a missense variant in ZBTB7B (3.7%, P=3.8×10-8, 0.05±0.01SD/MA), located at the downstream-end of a long-range association peak (1Mb) of common SNVs (MAF>25%) that include coding variants in EFNA1 and UBQLN4. Conditional analyses to determine whether the low-frequency and common SNVs represent independent signals are ongoing. No associated X-chromosomal low-frequency SNVs were identified. Despite our large sample size, we identified only three low-frequency SNVs. We cannot exclude the possibility of other low-frequency SNVs with smaller effect sizes. Preliminary results indicate that associated low-frequency SNVs may be located near common SNVs and conditional analyses are needed to determine which SNVs are driving the associations. Ongoing analyses also include an expansion of the sample (>400,000), gene-based analyses and functional follow-up.

## 275

**Genome-wide identification of novel genetic variants associated with erythrocyte membrane fatty acids.** *A.E. Locke[1], A.U. Jackson[1], A. Stancáková[2], Y. Wu[3], T.M. Teslovich[1], C. Fuchsberger[1], N. Narisu[4], P. Chines[4], R. Welch[1], H.M. Stringham[1], X.L. Sim[1], J. Huyghe[1], M. Civelek[5], N.K. Saleem[5], A. He[6], C. Tilford[6], P. Gargalovic[6], T. Kirchgessner[6], A.J. Lusis[5], K. Mohlke[3], M. Boehnke[1], M. Laakso[2].* 1) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI; 2) Department of Medicine, Kuopio University Hospital and University of Eastern Finland, Kuopio, Finland; 3) Department of Genetics, University of North Carolina, Chapel Hill, NC; 4) Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 5) Department of Medicine, University of California, Los Angeles, CA; 6) Bristol-Myers Squibb, Pennington, NJ.

Several genes have been associated with fatty acid levels. However, most studies have measured fatty acids from plasma, which have a short half-life (~2-4 minutes) and are heavily influenced by diet. In contrast, erythrocyte membrane bound fatty acids (EMFAs) persist for life of the erythrocyte (~ 3 months), are less influenced by dietary intake, and thus can better reflect genetic factors contributing to fatty acid synthesis and metabolism. As part of the METabolic Syndrome In Men (METSIM) study, we used gas chromatography to measure 20 fatty acids (six saturated, five mono-unsaturated, and nine poly-unsaturated), the cumulative measures of all saturated, mono-unsaturated, and poly-unsaturated fatty acids, and six ratios measuring enzymatic activity in 2,299 adult males from Kuopio, Finland. Following genotyping with Illumina Omni Express and Illumina Metabochip and imputation using the GoT2D reference panel, we performed single variant and gene-based association analysis with each of the 28 fatty acid measures. Twenty-four of the 28 traits were associated with at least one variant, and in total 45 single variant associations at 21 different loci reached genome-wide significance. In addition to established associations with fatty acid desaturases (*FADS1-2-3*), *ELOVL2*, and *PDXDC1*, we also identified associations near strong functional candidate genes: fatty acid elongases (*ELOVL5* and *ELOVL6*), acyltransferases (*LPCAT3*, *AGPAT4*, *AGPAT5*, and *CERS4*), and palmitoyltransferases (*ZDHHC21*, *CPT1A*, and *SPTLC3*). In conditional analyses, we identified three genomic regions with multiple associations for six different traits. Finally, we included putative functional variants (protein truncating and missense) in gene-based tests and identified low frequency variants in *ABHD3* significantly associated with myristic acid. We used eQTL results from subcutaneous adipose tissue expression data from 1,410 METSIM individuals to identify potential functional candidate genes at these loci. In reciprocal conditional analyses, the fatty acid GWAS SNP at four loci accounted for the expression signal of a candidate gene (*FADS1*, *SCD*, *PDXDC1*, *CERS4*) and *vice versa*. Measurement and association testing of fatty acids derived from erythrocyte membranes, along with expression data from the same individuals, has considerably enhanced understanding of the biological mechanisms of fatty acid metabolism.

## 276

**Filtering for Genomic Nonsense to Find Biological Significance: *SLC13A1* Nonsense Variants Enriched in Founder Population are Associated with Reduced Serum Sulfate and Increased Aspartate Aminotransferase Levels.** *C.G. Perry, J.A. Perry, J.R. O'Connell, L.M. Yerges-Armstrong, A.R. Shuldiner.* University of Maryland School of Medicine, Baltimore, MD.

Inorganic sulfate ($SO_4^{2-}$) is an important micronutrient vital for numerous cellular and metabolic processes in human development. Sulfate participates in the biotransformation of multiple compounds via sulfate conjugation (sulfation) mediated by sulfotransferases. Decreased sulfation capacity due to inadequate levels of circulating sulfate alters metabolism and activities of a variety of endogenous compounds and reduces detoxification, increasing toxicity to xenobiotics and some drugs. Additionally, impaired sulfation capacity is associated with autism, neurological diseases, skeletal dysplasias, and premature pubarche; yet little is known about sulfate regulation in humans. Using genomic applications to glean insights into human biology, we systematically identified nonsense, single nucleotide polymorphisms (SNPs) that are rare in the general population but enriched in the Old Order Amish due to a founder effect. Filters and bioinformatics tools were applied to Illumina Human Exome BeadChip data obtained from 1648 Amish subjects. Two non-linked, nonsense SNPs (c.34C>T, p.R12X and c.144G>A, p.W48X) in *SLC13A1*, which encodes a sulfate transporter responsible for sulfate (re)absorption in the kidneys and intestine, were discovered to be enriched in this Amish cohort compared to outbred populations (1.9-fold (0.49% vs. 0.26%) and 7.2-fold (0.94% vs. 0.13%), respectively, compared to 1000 Genomes European ancestry allele frequencies). Fasting serum sulfate concentration was measured in 175 individuals recruited by genotype. Both R12X and W48X SNPs were significantly associated with a 27-31% decrease in serum sulfate levels ($P=2.0\times10^{-5}$ and $P=6.2\times10^{-7}$, respectively; $P=4.3\times10^{-19}$ for both SNPs combined). We also queried the Amish Disease Research database in a hypothesis-free, phenome-wide association study (PheWAS) manner to identify novel associations with *SLC13A1* nonsense SNPs. We observed a PheWAS significant association between *SLC13A1* nonsense genotype carriers (R12X or W48X) and a 50% increase in aspartate aminotransferase (AST) levels ($P=2.0\times10^{-6}$). All association studies adjusted for age, age$^2$, sex and relatedness. The increase in AST levels is consistent with low-grade liver damage, possibly due to decreased sulfate levels and an inability to detoxify xenobiotics. Additional studies are warranted to better understand the importance of sulfate in human physiology and its potential role in disease and drug toxicity.

## 277

**Integrating metabolite, BMI and genetic data in phenotypic extremes, drawn from a population of 50,000 samples, to assess causality of metabolite levels in obesity.** *T. Esko[1,2,3], A. Metspalu[1], C. Clish[3], JN. Hirschhorn[1,3].* 1) Division of Endocrinology, Children's Hospital Boston, Boston, MA; 2) Estonian Genome Center, University of Tartu, Tartu, Estonia; 3) Broad Institute, Cambridge, MA.

Obesity is a disorder of energy metabolism and signaling between tissues. As such, it may be possible to obtain additional clues about particular causal pathways in obesity by examining metabolites in whole blood from lean and obese individuals. However, although many metabolites are correlated with obesity, causality is more difficult to ascertain. Mendelian randomization offers an approach to assess causality. In order to identify metabolites correlated with obesity, we carried out both targeted (340 metabolites) and untargeted metabolite profiling using LC-MS in 100 normal weight, 100 lean and 100 obese individuals drawn from the extremes of a population of 50,000 samples. As a measure of obesity, we analyzed body mass index (BMI), using z-scores adjusted for age and gender from the entire cohort. From the targeted approach we identified 155 metabolites associated with BMI. Many associated metabolites were highly correlated and included species of di- and triacylglycerols, amino acids with their metabolic derivatives and molecules related to the choline pathway. In order to assess whether these metabolic alterations could play a causal role or are simply reflections of the obese state, we applied Mendelian randomization (MR), which uses genetic variants that directly influence a potential mediator (in this case, a metabolite level) as instrumental variables to assess causal influences on an outcome (in this case, obesity). Genome-wide association analyses were conducted with each of the 155 metabolites correlated with BMI. In total we identified 15 sequence variants at array-wide significance level for 6 compounds, which on average explained 8 to 13 percent of the respective trait variance. By using the best sequence variant per metabolite as instrumental variables, we provide initial evidence that long-chain triacylglycerols (C50:4, C52:5 and C56:5, respective MR p-values 1.8e-4, 1.3e-4 and 9.6e-5) and branched-chain amino acid valine (p-value 1.1e-4) are causally linked to obesity. In conclusion, by using a well-powered extremes design, high throughput metabolite profiling and Mendelian randomization we were able to provide evidence that most metabolites correlated with BMI are likely downstream effects of obesity but that long-chain triacylglycerols, branched-chain amino acids and related pathways may be causally linked to obesity.

## 278

**Systems genetics analyses of human adipose tissue gene expression identify *cis* and *trans* regulatory networks for cardio-metabolic traits.** *M. Civelek[1], Y. Wu[3], C. Pan[2], A. He[4], C. Tilford[4], N.K. Saleem[5], C. Fuchsberger[6], A. Locke[6], H.M. Stringham[6], A.U. Jackson[6], N. Narisu[7], P.S. Chines[7], Y. Zhao[8], P.S. Gargalovic[4], J. Kuusisto[5], P. Pajukanta[2], K. Hao[9], X. Yang[8], T.G. Kirchgessner[4], F.S. Collins[7], M. Boehnke[6], M. Laakso[5], K.L. Mohlke[3], A.J. Lusis[1,2].* 1) Department of Medicine, University of California, Los Angeles, CA; 2) Department of Human Genetics, University of California, Los Angeles, CA; 3) Department of Genetics, University of North Carolina, Chapel Hill, NC; 4) Bristol-Myers Squibb, Pennington, NJ; 5) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 6) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI; 7) National Institutes of Health, Bethesda, MD; 8) Department of Integrative Biology and Physiology, University of California, Los Angeles, CA; 9) Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY.

Genome-wide association studies (GWAS) have identified numerous loci that are associated with complex traits related to Metabolic Syndrome (MetSyn). However, the molecular mechanisms by which these loci affect MetSyn are usually not known. We used transcript abundance as an intermediate trait to map expression quantitative trait loci (eQTL). We tested the association of 621,695 common single nucleotide polymorphisms (SNPs) and the expression of 18,279 genes in subcutaneous adipose tissue of 1,381 Finnish males who are part of the METabolic Syndrome In Men (METSIM) study. We identified 7,941 genes for which local SNPs (<1 Mb) showed significant association at 1% FDR (local eQTLs). We focused on SNPs that are associated with 170 cardio-metabolic traits and diseases in GWAS. Of the 1,542 SNPs, 621 had been directly genotyped, representing 121 independent loci ($r^2<0.3$). Using conditional analysis, we determined that 49 of the 121 loci were also local eQTLs for 54 genes. For example, rs8077889 has been associated with triglyceride levels in GWAS and is associated with the expression of two nearby genes, *DUSP3* ($P=1.5\times10^{-12}$) and *MPP3* ($P=3.0\times10^{-26}$). Only *DUSP3* is significantly correlated with triglyceride levels ($r=0.11$, $P=2.3\times10^{-6}$) whereas *MPP3* is not ($r=-0.06$, $P=2.7\times10^{-2}$), prioritizing *DUSP3* as the causal gene. The large number of subjects profiled in METSIM also allowed the study of distant acting loci (>1Mb). *KLF14* locus has previously been shown to regulate 10 adipose genes. We replicated these associations and identified an additional 20 genes that are trans regulated by this locus ($P=5\times10\text{-}8$). These genes are *C20orf194, CIB2, COX20, DYNLT1, EIF4E3, GALNT11, MAGED2, MSRA, NEO1, NGRN, PDCL3, PLIN5, SIRT3, SYNC, TECR, TUBB, UBE2Q2, UNC13B, ZNF219, ZNF226.* Twelve of the genes, for which an expression probe was available, replicated in another eQTL study in omental and subcutaneous adipose tissue in 556 and 741 people ($P=7.8\times10^{-3}$-$1.8\times10^{-9}$). This locus was not a trans regulator of gene expression in liver (N=566 people), peripheral blood (N=5,311), monocytes and macrophages (N=745) suggesting an adipose-specific mechanism of regulation. All of the additional 20 genes had transcript levels significantly correlated with metabolic traits ($|r|=0.33\text{-}0.11$; $P=6.6\times10^{-37}$-$9\times10^{-4}$). Our studies extend the mechanistic understanding of the *KLF14* locus and highlight the power of systems genetics approaches for dissecting complex traits to identify causal genes and pathways.

## 279

**Origin, frequency and functional impact of de novo structural changes in the human genome.** *K. Ye[1], W. Kloosterman[2], L.C. Francioli[2], F. Hormozdiari[3], T. Marschall[4], J.Y. Hehir-Kwa[5], A. Abdellaoui[6], E.W. Lameijer[7], M.H. Moed[7], V. Koval[8], I. Renkens[2], M.J. van Roosmalen[2], P. Arp[8], L. Karssen[9], B.P. Coe[3], R.E. Handsaker[10], E. Cuppen[2], D.T. Thung[5], M.C. Wendl[1,11], A. Uitterlinden[8], C.M. van Duijn[9], M. Swertz[12,13], C. Wijmenga[12,13], G. van Ommen[14], P.E. Slagboom[7], D.I. Boomsma[6], A. Schonhuth[5], E.E. Eichler[3], P.I.W. de Bakker[2,15], V. Guryev[16].* 1) Washington University in St Louis, St Louis, MO; 2) Center for Molecular Medicine, Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands; 3) Department of Genome Sciences, University of Washington, Seattle, USA; 4) Centrum voor Wiskunde en Informatica, Life Sciences Group, Amsterdam, The Netherlands; 5) Department of Human Genetics, Radboud University Medical Center, Nijmegen, Nijmegen, The Netherlands; 6) Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands; 7) Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; 8) Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; 9) Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; 10) Department of Genetics, Harvard Medical School, Boston, MA, USA; 11) Department of Mathematics, Washington University, St. Louis, MO, USA; 12) Center Groningen, Department of Genetics, Groningen, The Netherlands; 13) University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands; 14) Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; 15) Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands; 16) European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.

Small insertions and deletions (indels) and large structural variations (SVs) are major contributors to human genetic diversity and disease. However, mutation rates and characteristics of de novo indels and SVs in the general population have remained largely unexplored. We report 325 de novo structural changes identified in whole genomes of 250 families, including complex indels, retrotransposon insertions and interchromosomal events. These data indicate a mutation rate of 2.87 indels (1-20bp) and 0.16 SVs (>20bp) per genome per generation. Structural changes affect 4.1kbp of genomic sequence and 43 coding bases per generation - 65-90 times more nucleotides than de novo substitutions. A significantly larger proportion (66%) of structural changes originated from fathers. Additionally, we observed a non-uniform distribution of de novo SVs across offspring, suggesting unequal familial susceptibility to genomic rearrangements. These results reveal the mechanisms that govern changes in genome structure across generations.

## 280

**Parental somatic mosaicism contributes an under-recognized source of potentially recurrent new mutations.** *I.M. Campbell[1], B. Yuan[1], C. Robberecht[2], R. Pfundt[3,4,5], P. Szafranski[1], M.M. McEntagart[6], S.C.S. Nagamani[1,7], A. Erez[1,7], M. Bartnik[8], B. Wisniowiecka-Kowalnik[8], K.S. Plunkett[1], A.N. Pursley[1], S.H.L. Kang[1], W. Bi[1], S.R. Lalani[1,7], C.A. Bacino[1], M. Vast[6], K. Marks[6], M. Patton[6], P. Olofsson[9], A. Patel[1], J.A. Veltman[3,4,5], S.W. Cheung[1], C.A. Shaw[1], L.E.L.M. Vissers[3,4,5], J.R. Vermeesch[2], J.R. Lupski[1,7,10,11], P. Stankiewicz[1,8].* 1) Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX; 2) Centre for Human Genetics, University Hospital, K.U. Leuven, Leuven, Belgium; 3) Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Radboud university medical center, Nijmegen, The Netherlands; 4) Nijmegen Centre for Molecular Life Sciences, Radboud university medical center, Nijmegen, The Netherlands; 5) Institute for Genetic and Metabolic Disorders, Radboud university medical center, Nijmegen, The Netherlands; 6) Centre for Human Genetics, St. George's University of London, Cranmer Terrace, United Kingdom; 7) Texas Children's Hospital, Houston, TX; 8) Department of Medical Genetics, Institute of Mother and Child, Warsaw, Poland; 9) Mathematics Department, Trinity University, San Antonio, TX; 10) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 11) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.

New human mutations are thought to originate in germ cells, thus making a recurrence of the same mutation in a sibling exceedingly rare. However, increasing sensitivity of genomic technologies has anecdotally revealed mosaicism for mutations in somatic tissues of apparently healthy parents. Such somatically mosaic parents may also have germline mosaicim that can potentially cause unexpected intergenerational recurrences. Here we show that somatic mosaicism for transmitted mutations among parents of children with sporadic genetic disease is more common than currently appreciated. Using the sensitivity of patient-specific breakpoint PCR, we prospectively screened 100 families having children with genomic disorders due to rare variant deletion CNVs determined to be de novo by clinical analysis of parental DNA. Surprisingly, we identified four cases of low-level somatic mosaicism for the transmitted CNV in DNA isolated from parental blood. Integrated probabilistic modeling of gametogenesis developed in response to our observations predicts that identification of mutations in parental blood increases recurrence risk substantially compared to parents with mutations confined to the germline. Moreover, despite maternally transmitted mutations being the minority of alleles, our model suggests that sexual dimorphisms in gametogenesis result in a greater proportion of somatically mosaic transmitting mothers who are thus at increased risk of recurrence. Therefore, somatic mosaicism together with sexual differences in gametogenesis may explain a considerable fraction of unexpected recurrences of X-linked recessive disease. Overall, our results underscore an important role for somatic mosaicism and mitotic replicative mutational mechanisms in transmission genetics.

## 281

**Analysis of the Genetic Variation and Age Effects on Gene Expression Using RNA-seq Data from Multiple Tissues.** *A. Viñuela[1], A.A. Brown[2], A. Buil[3], M.N. Davies[1], P. Tsai[1], J.T. Bell[1], K.S. Small[1], E.T. Dermitzakis[3], R. Durbin[2], T.D. Spector[1].* 1) Department of Twin Research & Genetic Epidemiology, King's College London, London, London, United Kingdom; 2) Wellcome Trust Sanger Institute, Hinxton, United Kingdom; 3) Department of Genetic Medicine, University of Geneva, Switzerland.

Aging can be seen as the cumulative sum of all environments an organism is exposed to over the course of its life. Effects of aging on gene expression are diverse, manifesting on mean and variance in expression, splicing, and environmental changes to genetic regulation (GxE). We analysed RNA-seq data from adipose, skin, whole blood, and lymphoblastoid cell lines (LCLs) from ~850 female adult twins from the TwinsUK cohort (39-85 years old). We identified 905 (fat), 4307 (skin), 480 (blood), and 7 (LCLs) genes where the level of expression was associated with age. Only 430 genes were associated in two or more tissues, suggesting the aging process acts in tissue specific ways. Using fat methylome data available for 552 of the twins (Grundberg, 2013) and applying Bayesian Networks, we tested whether changes in expression with age were mediated by epigenetic markers. In most cases we found little evidence that epigenetic markers were involved in differential expression. To further understand whether these changes were due to GxE or environmental factors we investigated changes in discordance of expression within monozygous twin (MZ) pairs with age. We found 2 genes in adipose, 152 in skin, 1 in blood and 26 in LCLs where discordance was age-dependent. As MZ twins are genetically identical, age-dependent differences must be due to a changing environmental component. Decomposition of variance showed that age related genes had a larger genetic component, and that the sources of variation were highly tissue specific. A more sophisticated approach looked at changes in heritability of expression with age; observing in general that genetic factors explained decreasing proportions of variance as people aged. While this could be due to increased stochasticity, it is plausible that some of this effect is due to eQTL being modified as individuals age. To identify such SNPs whose effects on gene expression changed with age we performed a genome-wide scan looking for age-genotype interactions (GxA). One gene, CD82, was genome-wide significant in fat, showing a concrete example of the how genetic control of expression is modified over time. Interestingly this gene, a metastasis suppressor, showed increased expression with age in individuals with a particular, potentially protective, allele. In summary, we have produced a comprehensive description of how aging affects expression and its genetic control, observing that these effects are frequently tissue specific.

## 282

**Dynamics Personal Omics Profiles During Periods of Health, Disease, Weight Gain and Loss.** *M. Snyder[1], W. Zhou[1], B. Piening[1], K. Kukurba[1], K. Contrepois[1], C. Craig[2], R. Chen[1], G. Mias[1], J. Li-Pook-Than[1], S. Mitra[1], L. Jiang[1], B. Hanson[4], B. Leopold[4], S. Leopold[4], B. Cooper[4], L. Liu[2], V. Sikora-Wohfield[3], A. Butte[1,3], H. Tang[1], E. Sodergren[4], G. Weinstock[4], T. McLaughlin[2], M. Snyder[1].* 1) Department of Genetics, Stanford University, Stanford, CA; 2) Department of Medicine, Stanford University, Stanford, CA; 3) Department of Pediatrics, Stanford University, Stanford, CA; 4) Jackson Laboratory for Genomic Medicine, Farmington, CT.

To understand both how omic information can be incorporated into health care and how physiology changes during periods of healthy and stress periods, we have performed integrated Personal Omics Profiling (iPOP), combining genomic, DNA methylome, transcriptome, proteome, cytokine, metabolome, and microbiome (gut, nasal and other) information, in a cohort of 60 individuals during healthy and aberrant periods. The individuals are sampled frequently during times of respiratory illness, and twenty of the subjects have experienced a high caloric diet (and weight gain) for thirty days followed by low caloric diet (and weight loss) for sixty days. Our iPOP analysis of blood, urine and microbiome components revealed extensive, dynamic and broad changes in diverse molecular components and biological pathways across healthy and disease conditions as well as during weight gain and loss. Our study describes the biochemical and omic pathways associated with respiratory and weight gain stresses at a systems-wide level and those pathways that differ and are in common between different times of environmental stress.

## 283

**Longitudinal Study Of Whole Blood Transcriptomes In a Twin Cohort.** *J. Bryois[1,2,3], A. Buil[1,2,3], P. Ferreira[1,2,3], N. Panoussis[1,2,3], A. Planchon[1,2,3], D. Bielser[1,2,3], A. Viñuela[4], K. Small[4], T. Spector[4], E.T. Dermitzakis[1,2,3].* 1) Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; 2) Institute of Genetics and Genomics in Geneva (iGE3), Geneva, Switzerland; 3) Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland; 4) Department of Twin Research and Genetics Epidemiology, King's College, London, United Kingdom.

The majority of genes in human were recently found to be regulated by expression quantitative trait loci (eQTLs). Altough eQTLs studies vastly improved our understanding of the genetics of gene expression, they only provide a snapshot of the genetics of gene expression. In order to investigate the temporal aspect of the genetics of gene expression, we used RNA-seq on whole blood of females from the TwinsUK adult registry (21 DZ twin pairs, 19 MZ twin pairs, 25 Singleton) at two timepoints separated on average by 1.8 years. The twin structure of the data allowed to estimate the heritability of gene expression at the first and second timepoints (mean=0.3) and to discover that the difference in gene expression is also heritable (mean= 0.2). Using ~60 unrelated invididuals, we discovered 999 genes with a cis-eQTL (5% FDR) at the first timepoint and 1018 genes with a cis-eQTL at the second timepoint. The cis-eQTLs detected at both timepoint were largely shared ($\pi$1=88%) indicating that the genetic signal on gene expression is mostly stable over time. We found that 1556 genes are differentially expressed (DE) between the two timepoints and that the heritability of their difference in expression is significantly greater than for genes not DE. In addition, DE genes are on average more heritable than stable genes at the first timepoint but not at the second time point, indicating that a loss of the genetic control of gene expression is a possible cause of the differential expression. Furthermore, we found that DE genes are strongly enriched (FDR < 0.01%) in age related GO terms (ribosome, oxydative phosphorylation, parkinson's disease, hungtington's disease, alzheimer's disease and spliceosome). We observed that gene expression is on average weakly correlated (mean=0.2) between the two timepoints. However, genes with an eQTL had a significantly higher correlation between timepoints than genes without an eQTL. In addition, DE genes were significantly less correlated between timepoints than stable genes. We observed that the correlation of gene expression between timepoints is significantly correlated to the heritability of gene expression indicating that part of the correlation of gene expression between timepoints is due to genetics. Finally, using conservative approaches we discovered 39 genes (5% FDR) with a genetic effect on gene expression change between the two timepoints, providing examples of genes where the change in gene expression is genetically driven.

## 284

**High-throughput Determination of Long INterspersed Element-1 Integration Preferences in the Human Genome.** *D.A. Flasch[1], A. Macia[5], T. Widmann[5], J.L. García-Pérez[5], T.E. Wilson[1,2], J.V. Moran[1,3,4].* 1) Department of Human Genetics; 2) Department of Pathology; 3) Department of Internal Medicine, 1241 E. Catherine Street, University of Michigan Medical School, Ann Arbor, Michigan 48109-5618, USA; 4) Howard Hughes Medical Institute; 5) GENYO (Centre for Genomics and Oncological Research), Granada, Spain.

Long INterspersed Element-1 (LINE-1 or L1) retrotransposon-derived sequences comprise ~17% of the human genome reference sequence (HGR). However, since the majority of L1 retrotransposition events occurred millions of years ago, Darwinian selective pressures have skewed their initial genomic distributions. Thus, new and unbiased assessments are needed to accurately survey L1 integration preferences.

Here, we have exploited engineered L1s to generate *de novo* L1 retrotransposition events in various human cell lines. We used PCR-based strategies to specifically amplify the 3' ends of engineered human L1 retrotransposition events and their associated flanking genomic DNA sequences. The resultant amplicons then were sequenced using the Pacific Biosciences circular consensus DNA sequencing platform and passed through a bioinformatics pipeline to call integration sites at single nucleotide resolution with high accuracy and sensitivity. To date, we have characterized ~23,000 L1 insertions in HeLa cells, ~30,000 insertions in ovarian carcinoma cells, ~500 insertions in human embryonic stem cells (hESCs), and ~900 insertions in hESC-derived neural progenitor cells. This large data set should provide the statistical power to determine if L1 preferentially integrates into specific genomic regions and whether L1 integration preferences differ between cell types.

Our preliminary data revealed that, depending upon the observed cell type, approximately 22%-38% of the engineered L1 insertions resided within introns. Approximately 40-45% of these insertions occurred within the largest intron of the gene. Collectively, we discovered over 700 L1 integration events into the 5' untranslated region (UTR), coding exon, or 3'UTR of genes; such insertions are relatively infrequent in the HGR. These two observations suggest that euchromatic regions of the genome are accessible and susceptible to *de novo* L1 integration events. We currently are exploring whether other features (*e.g.,* DNaseI hypersensitive sites, acetylation and methylation sites, replication origins, transcription start sites, intergenic regions, *etc.*) render genomic DNA vulnerable to L1 integration. In sum, our strategy allows an accurate assessment of L1 integration site preferences before being blurred by selective pressures that occur over evolutionary time.

## 285

**Cryptic splicing adversely affects LINE-1 retrotransposition.** *P.A. Larson[1], C.R. Beck[1], J.V. Moran[1,2,3].* 1) Department of Human Genetics; 2) Department of Internal Medicine; 3) Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI, 48109 USA.

Long INterspersed Element 1 (LINE-1 or L1) retrotransposons are a prolific family of mobile genetic elements that comprise approximately 17% of the human genome reference sequence (HGR). Most L1s have been rendered inactive by mutational processes and can be considered to be molecular fossils. However, the average human genome contains approximately 80-100 full-length, retrotransposition-competent L1s (RC-L1s). Ongoing RC-L1 retrotransposition contributes to both intra- and inter- individual genetic diversity, and on occasion causes sporadic cases of human disease. Previous studies from the Belancio and Deininger laboratories revealed that human L1 mRNAs contain cryptic splice sites, and that their utilization results in internally deleted L1 mRNAs. Given that the proteins (ORF1p and ORF2p) encoded by RC-L1s exhibit cis-preference (i.e., they preferentially bind to their respective encoding mRNA to enable retrotransposition), we sought to determine the effects of splicing on L1 retrotransposition. Here, we report the identification, frequency, and characterization of two separate classes of Spliced Integrated Retrotransposed L1 Elements (SpIREs) in the HGR. Strikingly, over the last 20 million years, SpIREs have been responsible for approximately 2% of presumed full-length L1s. The first class of SpIRE is generated by the retrotransposition of L1 mRNAs containing an intra-5' untranslated region (UTR) splicing event, which deletes cis-acting transcription factor binding sites critical for L1 transcription. Northern blot and luciferase-based assays reveal that the 5'UTRs of these SpIREs have severely reduced promoter activity. The second class of SpIRE is generated from L1 mRNAs containing an intra-5' UTR/ORF1 splicing event. These L1s lack promoter activity and cannot produce a full-length ORF1 protein. Using a cultured cell retrotransposition assay, we determined that neither class of SpIRE could retrotranspose efficiently in cis. Our data demonstrate that SpIREs are essentially 'dead on arrival' and likely cannot undergo subsequent retrotransposition. Thus, the use of cryptic splice sites within L1 mRNAs is detrimental to L1 retrotransposition and may be a mechanism to limit and/or regulate L1 retrotransposition in a developmental or cell-type specific manner.

## 286

**Discovery of a novel retrotransposon family in the *Callithrix jacchus* genome.** *M.K. Konkel[1], B. Ullmer[2], J.A. Walker[1], R. Hubley[3], E.L. Arceneaux[1], S. Sanampudi[1], C.C. Fontenot[1], A.F.A. Smit[3], M.A. Batzer[1], The Common Marmoset Genome Sequencing and Analysis Consortium.* 1) Department of Biological Sciences, Louisiana State University, Baton Rouge, LA; 2) School of Electrical Engineering and Computer Science, Center for Computation and Technology (CCT), Louisiana State University, Baton Rouge, LA; 3) Computational Biology, Institute for Systems Biology, Seattle, WA.

The genome of the common marmoset (*Callithrix jacchus*) represents the first sequenced and analyzed New World monkey (NWM, platyrrhine) genome. *C. jacchus* is also a common animal model for studying human disease, including neuroscience and infectious diseases. In its draft genome assembly [calJac3.2], we identified a novel mobile element, dubbed "Platy-1." Based on our comparative analyses, we determined that Platy-1 elements are unique to NWMs. A full-length Platy-1 element is just over 100 bp in length and does not appear to contain coding sequence. Platy-1 elements exhibit facets such as termination in an Adenosine tail of varying length, target site duplications, and an endonuclease cleavage site characteristic of non-LTR retrotransposons. The combination of these features strongly suggests that Platy-1 elements are inserted in the genome by the enzymatic machinery of L1 through a mechanism called target-primed reverse transcription. Based on our whole genome analysis, we determine that the *C. jacchus* draft assembly contains more than 2000 Platy-1 elements. Our Platy-1 subfamily reconstruction revealed the presence of older subfamilies that most likely were active early in NWM evolution and have ceased activity, as well as subfamilies of more recent origin. Further supported by our phylogenetic analyses, we conclude that the founder Platy-1 element arose prior to the radiation of NWMs. We determine that Platy-1 has propagated throughout the evolution of NWMs in the lineage leading to *C. jacchus*. Furthermore, the identification of Platy-1 elements identical to their respective consensus sequence and polymorphic within common marmoset populations indicates ongoing retrotransposition activity until at least very recently. While the overall repeat content of the common marmoset genome is comparable to other primate genomes, the finding of a novel retrotransposon family specific to NWMs illustrates that each primate lineage evolves uniquely. Furthermore, a better understanding of the composition of the *C. jacchus* genome will support advances in biomedical science.

## 287

**Comprehensive Phenotypic Analysis of 19 Individuals with Goltz Syndrome (Focal Dermal Hypoplasia).** *V. Sutton[1], H. Herce[2], T.R. Hunt[3], A.L. Smith[3], K.J. Motil[4], A.F. Bree[5], M. Fete[6], R.W. Goltz[7].* 1) Mol. & Human Genetics, Baylor College of Medicine, Houston, TX; 2) Department of Ophthalmology, Baylor College of Medicine, Houston, TX; 3) Department of Orthopedic Surgery, Baylor College of Medicine, Houston, TX; 4) Department of Pediatrics, Baylor College of Medicine, Houston, TX; 5) Dermatology Private Practice, Houston, TX; 6) National Foundation for Ectodermal Dysplasia, Fairview Heights, IL; 7) Department of Dermatology (Emeritus), University of California, San Diego, CA.

Goltz syndrome is an X-linked dominant disorder caused by mutations in the gene *PORCN*. This pleiotropic disorder manifests in skin, skeleton and eyes. Information regarding the incidence of various phenotypic features in Goltz syndrome is limited and has come from small case reports and synthesis and summary of these individual reports. We present the largest clinical analysis of individuals with Goltz syndrome to date. Nineteen individuals from 18 families (one father-daughter pair) were studied. The average age was 12 years with a median age of 9 years and range of 1-55 years. Facial characteristics included: asymmetry (52%); hypoplastic alae nasi (87%); simple or crumpled ears, ear tags, ptosis, long columella, facial cleft and diastasis recti. All had skin atrophy that followed lines of Blashko and tended to be erythematous at birth and evolved over time into hypopigmented areas. There was a high frequency of sun sensitivity and fragility/bleeding with 47% having persistent crusted erosions. Telangectasias were seen in 84%. Papillomas were reported in 63%. The characteristic yellowbrown or pink lipomatous changes were seen in 68%. Patchy alopecia was seen in 79% and 42% had diffusely sparse scalp hair. Scanning electron microscopy revealed hair shaft abnormalities in 89% including narrow diameter, flattened shaft, pili torti and trichorrehexis nodosa; 16% had the uncommon finding of pili trianguli et canaliculi. Ridging and atrophy of the nails was seen in 84%. Major limb abnormalities were seen in 90% of individuals and included: Ectrodactyly (74%); syndactyly (84%); transverse limb deficiency (11%). Severe eye malformations were seen at rates higher than previously reported or summarized from the literature and included (percentages in eyes): right anophthalmia (5%); microphthalmia (42%); iris coloboma (47%); horioretinal coloboma (58%). Gynecological examination revealed bilateral labial hypoplasia in 88%, short perineum in all and vulvar papillomas in 18%. DNA mutation analysis of the *PORCN* gene did not identify any clear genotype-phenotype correlation. This is the largest analysis of phenotypic features in individuals with Goltz syndrome and provides for improved diagnosis, prognosis and management. In addition, the tissue repository for transformed lymphoblasts and fibroblasts will provide a resource to better understand the role of PORCN in WNT signaling and to screen and test potential new therapies for Goltz syndrome.

## 288

**Multiple Symmetric Lipomatosis - New Aspects of a Forgotten Syndrome.** *J. Schreml[1], A. Lindner[1], O. Felthaus[2,3], S. Klein[2], C. Pallouras[4], S. Schreml[5], I. Harsch[6], T. Meitinger[7], T.M. Strom[7], J. Altmüller[8,1], S. Staubach[1], F.G. Hanisch[10,11], H. Thiele[8], P. Nürnberg[8,9,10], L. Prantl[2,3].* 1) Institute of Human Genetics, University of Cologne, Cologne, Germany; 2) Center of Plastic Surgery, Department of Trauma Surgery, University Medical Center Regensburg, Regensburg, Germany; 3) Applied Stem Cell Research Center Regensburg, University of Regensburg, Regensburg, Germany; 4) Dermatology, University of Cologne, Cologne, Germany; 5) Department of Dermatology, University Medical Center Regensburg, Regensburg, Germany; 6) Internal Medicine II, Endocrinology/Diabetology, Thüringen-Kliniken Georgius Agricola, Saalfeld/Saale, Germany; 7) Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; 8) Cologne Center for Genomics (CCG), University of Cologne, Cologne, Germany; 9) Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases, University of Cologne, Cologne, Germany; 10) Center for Molecular Medicine Cologne, University of Cologne, Cologne, Germany; 11) Medical Faculty, Institute of Biochemistry II, University of Cologne, Cologne, Germany.

Multiple Symmetric Lipomatosis (MSL, OMIM# 151800) is a rare disorder of the adipose organ with regional, tumorous growth of non-encapsulated adipose tissue reaching often disfiguring proportions. Knowledge on the phenotype, etiology and molecular mechanisms involved is scarce. Patients are suffering from extremely reduced quality of life and the only effective therapy to date is surgery. MSL is relatively unknown to many health care professionals, often mistaken for obesity and therefore likely underdiagnosed. Most cases are considered sporadic and only few familial cases have been reported with both, mitochondrial and autosomal dominant, mode of inheritance suggested. We have collected a total of 24 patients, among them 12 patients from 5 families and 12 sporadic cases. All pedigrees are compatible with autosomal dominant inheritance with one family showing paternal transmission. Previous studies have suggested that MSL tissue differs on a histological and molecular biological level from regular white fat. To further characterize these differences we are using human adipose derived stem cells (hASCs) from affected and unaffected adipose tissue from MSL patients and control patients. We have found that patient hASCs exhibit abnormal response towards differentiation stimuli. In absence of differentiation media the stem cells show increased proliferation and spontaneously exhibit pronounced lipid droplet formation after prolonged cultivation. Conversely, under differentiation conditions, patient hASCs show reduced propensity for the final phase of adipocyte differentiation. MicroRNA 183, a driver of adipocyte differentiation, was differentially expressed in the patient cells. Moreover, we observed a shift in the expression of the three isoforms of SHC1, a protein known to be involved in the regulation of fat accumulation, ageing and oxidative stress response, in affected patient cells. We are currently performing proteomics analysis (iTRAQ = isobaric Tags for Relative and Absolute Quantitation) to further investigate the involved signaling pathway(s) and are at the same time performing genetic analysis by exome sequencing of familial cases. Elucidation of the molecular mechanisms involved promises new insight into general adipose tissue biology and regulation. Finally, establishing histological/biochemical hallmarks that can be used for diagnosis in a clinical setting can lead to a dramatic improvement of patient care in the future.

## 289

**Obstetric and Gynecologic Health in Patients with Xeroderma Pigmentosum.** *M. Merideth[1], D. Tamura[2], J. DiGiovanna[2], K. Kraemer[2].* 1) Medical Genetics Branch, NHGRI, NIH, Bethesda, MD; 2) Dermatology Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD.

Objective: To evaluate the obstetric/gynecologic health in women with xeroderma pigmentosum (XP) Background: XP is a rare autosomal recessive DNA repair disorder with marked sun sensitivity, freckle like skin pigmentation and 1000-fold increase in UV-induced skin cancer. About 25% of XP patients also have progressive neurological degeneration. There are 7 XP nucleotide excision repair complementation groups (XP-A to XP-G) and an XP variant with defective DNA polymerase eta. There is little information available about the obstetric and gynecologic issues in this patient population. Methods: Twenty XP females (2 XP-A, 5 XP-C, 4 XP-D, 4 XP/TTD, 2 XP-E, 1 XP-variant, 2 unknown) were examined at the NIH from 2004 to 2014. Evaluation included ob/gyn history, medical record review, physical exam, blood testing and pelvic imaging, if medically indicated. Results: The patients ranged in age from 9-52 years; median age at menarche was 13 years (range 9-16.5, n=14). Thirteen of the patients reported monthly menses and 2 reported less frequent menses. Menstrual flow ranged from 3-20 days (median=5 days); 5 patients reported heavy flow, 7 reported dysmenorrhea, and 1 had premenstrual syndrome. One patient had precocious puberty and 1 XP-V patient had infertility. FSH (n=19) and estradiol (n=16) levels were in the expected range except for 3 XP-C patients with premature ovarian insufficiency. Nine XP patients were pregnant with a total of 22 pregnancies; of these, 3 resulted in early miscarriage; 1 XP/TTD patient had an emergent Cesarean section (C/S) for uterine rupture at 20 weeks, a C/S for placenta previa at 31 weeks and a C/S at term; the remaining 16 pregnancies resulted in normal vaginal deliveries at term. Four patients out of 13 had mild cervical abnormalities on Pap smear, 3 of which were treated with cryotherapy. One patient had a dysplastic nevus on the vulva and 1 patient had an invasive SCC of the jaw diagnosed during pregnancy. Conclusions: Our pilot study suggests an increased risk of premature ovarian insufficiency, precocious puberty, possible increase in growth of neoplasms during pregnancy and, as seen in mothers of TTD patients, increased pregnancy complications in XP/TTD patients. We did not observe an increased risk of menstrual irregularities or dysplasia of the cervix. The finding of a vulvar dysplastic nevus emphasizes the importance of including the external genitalia as part of the routine skin exam of XP patients.

## 290

**Adams-Oliver syndrome: Refining the diagnostic phenotype.** *S. Hassed, S. Li, J. Mulvihill, S. Palmer.* Dept Pediatrics/Gen, Univ Oklahoma HSC, Oklahoma City, OK.

Adams-Oliver syndrome (AOS) is a rare genetic disorder defined as aplasia cutis congenita (ACC) and limb reduction defects; fetal vasculopathy has been proposed as the pathogenesis. Associated anomalies are not well characterized. Four genes are known: DOCK6 [MIM 614219, 614194], EGOT [MIM 615297, 614789], ARHGAP31 [MIM 100300, 610911], and RBPJ [MIM 614814, 147183]; the last previously reported by one author (SH). We describe 15 unreported individuals with AOS; none have identifiable mutations, so additional genes remain to be found. A total literature review of 270 cases from 155 families was analyzed for anomalies; 17 have known genotype. Great variability is observed regardless of genotype. As expected, 94% of probands had ACC with limb defects; 34% of affected relatives had both cardinal manifestations. Other limb defects included syndactyly, ectrodactyly, and polydactyly. All but 1% of ACC were on the scalp, and only 52% of scalp lesions had underlying bone aplasia. The third most common finding was brain abnormalities, 22%, higher than previously reported. These were quite varied and many could not be attributed to vascular events. Various brain malformations were seen (common hypoplasias, microcephaly, hydrocephalus, with few vascular malformations), but some sequelae of vascular dysfunction (ischemia, infarct, periventricular leukomalacia and calcifications) also occurred. Most unexpected was a high rate, 22%, of CNS migration defects (pachygyria, polymicrogyria, heterotopia, colpocephaly). A deformation mechanism from cranial defects is not likely a major cause of brain abnormalities; half of patients with brain anomalies had intact skulls, though focal hypoplasia or abnormal ventricles were twice as common in cases where calvarial defects were present. There were no consistent associations between vascular anomalies or events and brain malformations. A wide variety of congenital heart defects was seen in 21%. Cutis marmorata telangiectasia congenital was present in 19%. Other vascular anomalies were seen in 13%; most involved scalp vessels (7%), but internal vascular anomalies were seen in 5%, and 2% involving hepatoportal disease. The remaining anomalies had no predictable pattern, with intraocular anomalies in 3%, cleft lip/palate in 2%, and renal anomalies in 1%. Fourteen childhood deaths were reported from complications, including 3 from hemorrhage. Defining clinical diagnostic criteria seems feasible.

## 291

**Alpha-fetoprotein assay on dried blood spot for hepatoblastoma screening in children with Beckwith-Wiedemann syndrome and Isolated Hemihyperplasia.** *A. Mussa[1], V. Pagliardini[1], C. Molinatto[1], G. Baldassarre[1], A. Corrias[1], F. Fagioli[2], M. Cirillo Silengo[1], G.B. Ferrero[1].* 1) Department of Pediatrics, University of Torino, Torino, Italy; 2) Paediatric Onco-Haematology Unit, Regina Margherita Children's Hospital, Turin, Italy.

Beckwith-Wiedemann syndrome (BWS) and Hemihyperplasia (HH) are overgrowth disorders with embryonal tumor predisposition. As hepatoblastoma complicates 6% of cases, for its early diagnosis patients undergo a cancer screening based on the repeated dosage of the sensitive tumor marker alpha-fetoprotein (αFP). However, the burden connected with the frequent blood drawns causes compliance issues and poor adherence to the surveillance protocol. Many centres opt not to perform this screening test given its unfavourable cost-effectiveness, the relatively low incidence of the tumor, and the poor adherence of the patients. We sought to analyse feasibility and reliability of αFP dosage by a micromethod based on blood spot dried on filter paper (DBS), aiming at developing a reliable laboratory test more tolerable for patients. Two-hundred and fifty coupled αFP determinations (plasma+ DBS) collected simultaneously were compared. A hundred-two determinations were performed in 45 patients affected by BWS/HH, 147 in healthy controls, and 1 in a patients with non-syndromic hepatoblastoma. The plasma αFP dosage method (AutoDELFIA hAFP, PerkinElmer) was adapted to DBS adsorbed on paper matrix for newborn screening. For the assay was used a paper disk containing 1.3 µl of blood. Reaction reading was carried out by immunometric fluorimeter for delayed fluorescence of europium. There was strong correlation between plasmatic and DBS αFP ($r^2 = 0.999$, $p < 0.001$). Cohen's k coefficient for correlation was 0.96 for diagnostic cut-off of 10 U/ml ($p < 0.001$), the threshold employed in clinical practice for tumor screening. αFP determinations on serum and DBS in cases evaluated longitudinally were overlapping and highly consistent across the entire wide range of physiological fluctuations, starting from neonatal concentrations of the magnitude of 100,000 U/ml to those infantile, ranging from 0 to 10 U/ml. Longitudinal evaluations were reliable as revealed in both serum and DBS a similar physiological decreases of αFP concentrations during the first years of life. DBS and plasma FP in the patient with non-syndromic hepatoblastoma were of comparable magnitude, 540 and 700 U/ml respectively, both consistent with the diagnosis. The DBS method allowed to dose αFP reliably and consistently for the concentrations commonly employed in clinical practice, representing a reliable strategy for conducting cancer screening in overgrowth syndromes.

## 292

**Clinical and radiographic study of 93 patients with a molecularly proven non-lethal type 2 collagen disorder.** *G.R. Mortier[1,23], R.J.A.J. Nievelstein[2], E.J.J. Verver[3], V. Topsakal[3], P. Van Dommelen[4], K. Hoornaert[5], M. Le Merrer[6], A. Zankl[7], M.E.H. Simon[8], S.F. Smithson[9], C. Marcelis[10], B. Kerr[11], J. Clayton-Smith[11], E. Kinning[12], S. Mansour[13], F. Elmslie[13], L. Goodwin[14], A.H. van der Hout[15], H.E. Veenstra-Knol[15], J.C. Herkert[15], A.M. Lund[16], R.C.M. Hennekam[17], A. Mégarbané[18], M.M. Lees[19], L.C. Wilson[19], A. Male[19], J. Hurst[19,20], N.V. Knoers[21], P. Coucke[22,23], P.A. Terhal[21].* 1) Dept Medical Genetics, Antwerp Univ Hosp, Antwerp (Edegem), Belgium; 2) Department of Radiology, University Medical Centre Utrecht, The Netherlands; 3) Department of Otorhinolaryngology and Head & Neck Surgery, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Utrecht, The Netherlands; 4) Department of Life Style, TNO, Leiden, The Netherlands; 5) Department of Ophthalmology, University Hospital Ghent, Ghent, Belgium; 6) Department of Genetics, INSERM UMR_1163, Paris Descartes-Sorbonne Paris Cité University, Imagine Institute, Hôpital Necker-Enfants Malades, Paris, France; 7) Discipline of Genetic Medicine,The University of Sydney, Sydney, Australia and Academic Department of Medical Genetics, Sydney Children's Hospital Network (Westmead), Sydney, Australia; 8) Department of Clinical Genetics, Erasmus Medical Centre, University Medical Centre, Rotterdam, The Netherlands; 9) Department of Clinical Genetics, St. Michael's Hospital, Bristol, United Kingdom; 10) Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Institute for Genetic and Metabolic Disease, Radboud University Medical Centre, Nijmegen, The Netherlands; 11) Manchester Centre For Genomic Medicine, University of Manchester, St Mary's Hospital, Manchester M139WL, United Kingdom; 12) Department of Clinical Genetics, Southern General Hospital, Glasgow G51 4TF, United Kingdom; 13) SW Thames Regional Genetics Service, St George's NHS Trust, London, United Kingdom; 14) Department of Genetics, Nepean Hospital, Penrith, Australia; 15) Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands; 16) Centre for Inherited Metabolic Diseases, Department of Clinical Genetics, Copenhagen University Hospital, Copenhagen, Denmark; 17) Department of Pediatrics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; 18) Unité de Génétique Médicale et Laboratoire Associé Institut National de la Santé et de la Recherche Médicale UMR-S910, Université Saint-Joseph, Beirut, Lebanon; 19) Department of Clinical Genetics, Great Ormond Street Hospital, London WC1N 3JH, United Kingdom; 20) Department of Clinical Genetics, Churchill Hospital, Oxford, United Kingdom; 21) Department of Medical Genetics, Division of Biomedical Genetics, University Medical Centre Utrecht, Utrecht, The Netherlands; 22) Department of Medical Genetics, Ghent University Hospital, Ghent, Belgium; 23) Ghent University, Ghent, Belgium.

The type 2 collagenopathies are a clinically heterogeneous group of chondrodysplasias caused by a heterozygous mutation in the COL2A1 gene. With the goal to improve counseling and develop more evidence-based management guidelines, we performed a thorough analysis of the phenotypic features and clinical course in a group of 93 patients with a non-lethal type of spondyloepiphyseal dysplasia due to a heterozygous COL2A1 mutation. The study group included 51 females and 42 males. The identified mutations were 68 missense mutations substituting a glycine residue in the triple helical domain, 9 splice site mutations, 5 arginine-to-cysteine substitutions, 5 in-frame deletions or duplications and 6 mutations in the C-terminal propeptide. The majority of the patients (80/93) had disproportionate short stature with on radiographs features of SEDC (n=66), SEMD (n=5), Kniest dysplasia (n=7) or spondyloperipheral (Torrance-like) dysplasia (n=2). The remaining 13 patients had a mild SED phenotype with premature degenerative joint disease but normal stature, in some cases resembling Stickler syndrome or multiple epiphyseal dysplasia. Cleft palate was present in 22% of the patients. At birth, 9% of the children presented with a clubfoot deformity and 26% of the neonates experienced respiratory problems. Myopia was found in 45% of the patients. In at least two patients, myopia was detected in childhood despite a normal ophthalmological examination in infancy. A spontaneous retinal detachment occurred in 12% of the patients (median age: 14 years; youngest age: 3.5 years). Ophthalmological anomalies tended to be more common and severe in the splice site mutation group, rather mild and infrequent for glycine-to-serine substitutions, and absent in patients with an arginine-to-cysteine mutation. Complaints of hearing loss were reported in 32 cases (37%) of whom 17 needed hearing aids (6 of those patients had a splice site mutation). More than 50% of the patients underwent orthopedic surgery, usually for scoliosis (present in 48% of all patients), hip replacement or femoral osteotomy. Patients with a glycine to a non-serine substitution had generally a more severe skeletal phenotype. Odontoid hypoplasia was present in 56% of the patients. A correlation between odontoid hypoplasia and short stature was observed in our study group. Atlanto-axial instability, found in 28% of the patients who underwent flexion-extension films, rarely resulted in neurological problems.

## 293

**Diagnostic criteria for Stickler syndrome based on comprehensive clinical and molecular analysis.** *F. Acke[1,2], P. Coucke[2], O. Vanakker[2], K. Hoornaert[3], I. Dhooge[1], A. De Paepe[2], E. De Leenheer[1], F. Malfait[2].* 1) Department of Otorhinolaryngology, Ghent University Hospital, Ghent, Belgium; 2) Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium; 3) Department of Ophthalmology, Ghent University Hospital, Ghent, Belgium.

Stickler syndrome is a heterogeneous disorder variably affecting the ocular, orofacial, auditory and skeletal system. Mutations in *COL2A1*, *COL11A1* and *COL11A2* have been found to cause Stickler syndrome and result in slightly distinct phenotypes, referred to as *type 1*, *type 2* and *type 3* respectively. Due to the large phenotypic variability, no consensus about minimal clinical diagnostic criteria exists. Currently, diagnosis is mainly based on expert opinion and positive mutation analysis. The aim of this study is to better define the syndrome and its different types by creating clinically-based guidelines. Medical records of more than 250 probands with a clinical suspicion of Stickler syndrome were reviewed for relevant symptoms and molecular results. *COL2A1* analysis was performed in all patients, and *COL11A1* and *COL11A2* were subsequently analyzed in the *COL2A1*-negative patients. In 90% of the probands, the disease-causing mutation was detected, of which 82% was located in the *COL2A1* gene, 14% in *COL11A1* and 4% in *COL11A2*. Most *COL2A1* mutations lead to haploinsufficiency (nonsense mutations, out-of-frame deletions), whereas the majority of mutations in *COL11A1*/*COL11A2* exhibit a dominant-negative effect (in-frame exon deletions, glycine substitutions). Symptoms that are more present in the mutation-positive patients and thus stronger direct towards Stickler syndrome, are high myopia, retinal detachment, cleft palate and a positive familial history suggesting autosomal dominant inheritance. Hearing loss, joint hypermobility and premature arthropathy are frequently found in Stickler syndrome, but are less specific. The main characteristics differentiating the three types of Stickler syndrome are appearance of the vitreous (membranous in *type 1*, beaded in *type 2* and normal in *type 3*) and severity of hearing loss (mild high-frequency hearing loss in *type 1*, moderate pan-frequency hearing loss in *type 2* and *type 3*), although this distinction is not absolute. Based on these clinical as well as molecular results, diagnostic criteria for the different types of Stickler syndrome using a point-scale of different symptoms are proposed. These novel criteria may guide clinicians to better diagnose Stickler syndrome and might help to select the correct molecular analysis.

## 294

**Updated Cardiac Description in Loeys Dietz Syndrome.** *G.L. Oswald[1], E.M. Reynolds[1], H.C. Dietz[1,2], J.P. Habashi[3], genTAC consortium investigators.* 1) Genetics, Johns Hopkins, Baltimore, MD; 2) Howard Hughes Medical Institute, Bethesda, MD; 3) Pediatric Cardiology, Johns Hopkins University, Baltimore, MD.

Loeys-Dietz syndrome (LDS) is an autosomal dominant connective tissue disorder characterized by a triad of aortic root enlargement with arterial tortuosity, hypertelorism and bifid uvula. In the original description in 2005, an increased prevalence of cardiac features were noted including patent ductus arterioles (PDA), patent foramen ovale (PFO)/atrial septal defect (ASD), bicuspid aortic valve (BAV), and ventricular septal defect (VSD). The purpose of this study was to analyze the 73 patients with LDS with confirmed TGFBR1 or 2 mutations in the GenTAC (National Registry of Genetically Triggered Thoracic Aortic Aneurysms and Cardiovascular Conditions) registry, and compare the findings to the 695 patients with a diagnosis of Marfan syndrome (MFS) confirmed by the Ghent criteria. The average age of diagnosis for LDS was 17.1 and 68.5% of patients were diagnosed at an age<21 years. In LDS, the prevalence of PDA (11.0%) and PFO/ASD (26.0%) was significantly greater than reported in MFS (0.1%, 2.0%, respectively; p<0.0001). The prevalence of BAV was two fold higher in LDS (8.2%) versus MFS (3.9%) and the prevalence of any aortic regurgitation (AR) was also significantly greater (46.6% v. 27.1%; p<0.001). Mitral valve prolapse (MVP) was not significantly different between LDS 32.9%) and MFS (26.6%), however the prevalence of greater than trace mitral regurgitation (MR) was greater in MFS (54.0%) as compared to LDS (30.1%; p<0.0001. Aortic dimensions prior to any aortic surgery were available in 33 LDS patients and revealed that 91% of LDS patients had an aortic root z score >2 and 79% had a z score>3. Furthermore, only 39% had an ascending aortic z score >2 and in only 18%, the z score was >3. No patients had a dilated ascending aorta in the absence of aortic root dilation. Analysis of the GenTAC database confirms the increased prevalence of certain congenital heart lesions in patients with LDS particularly in comparison to MFS. The area of aortic enlargement in LDS, as in MFS, shows a strong predilection for the root. The presence of these abnormalities should raise the consideration of LDS in patients with aortic aneurysm.

## 295

**Mosaic loss of chromosome Y (LOY) in blood cells is associated with shorter survival and higher risk of cancer in men.** *L.A. Forsberg[1,2], C. Rasi[1,2], N. Malmqvist[1,2,3], H. Davies[1,2], S. Pasupulati[1,2], G. Pakalapati[1,2], J. Sandgren[4], T. Diaz de Ståhl[4], A. Zaghlool[1,2], V. Giedraitis[5], L. Lannfelt[5], J. Score[6], N.C.P. Cross[6], D. Absher[7], E. Tiensuu Janson[3], C. Lindgren[8,9], A.P. Morris[8], E. Ingelsson[2,3], L. Lind[3], J.P. Dumanski[1,2].* 1) Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden; 2) Science for Life Laboratory, Uppsala University, Uppsala, Sweden; 3) Department of Medical Sciences, Uppsala University, Uppsala, Sweden; 4) Department of Oncology-Pathology, Cancer Center Karolinska, Karolinska Institutet, Stockholm, Sweden; 5) Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden; 6) Faculty of Medicine, University of Southampton, Southampton, UK; 7) HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA; 8) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 9) Broad Institute of MIT and Harvard University, Cambridge, Massachusetts, USA.

It is well known that men have an overall shorter life expectancy compared with women. However, it is less well recognized that incidence and mortality for sex-unspecific cancers are higher in men, a fact that is largely unexplained. Age-related loss of chromosome Y (LOY) is frequent in normal hematopoietic cells and it was first described more than 50 years ago, but the phenotypic consequences of LOY have been elusive. Our latest results suggest that LOY could be a key factor to explain the higher mortality of men. Survival analyzes performed in the Swedish ULSAM-cohort (Uppsala Longitudinal Study of Adult Men) with >1100 participants analyzed on 2.5M Illumina SNP-array indicated that LOY in peripheral blood could be associated with risks of all-cause mortality (hazards ratio (HR) = 1.91, 95% confidence interval (CI) = 1.17-3.13; 637 events) as well as non-hematological cancer mortality (HR = 3.62, 95% CI = 1.56-8.41; 132 events). Among the elderly men in this cohort, followed clinically for up to 20 years, at least 8.2% of the subjects were affected by LOY in a significant fraction of blood cells. The median survival time in men affected with LOY was half, i.e. 5.5 years shorter, compared to the men without mosaic LOY in blood cells. The association of LOY with risk of all-cause mortality was validated (HR = 3.66, 95% CI = 1.27-10.54; 59 events) in the independent PIVUS-cohort (Prospective Investigation of the Vasculature in Uppsala Seniors) in which 20.5% of men showed LOY. Our discovery of a correlation between LOY and all-cause mortality as well as non-hematological cancer mortality will be published in Nature Genetics. These results illustrates the impact of post-zygotic mosaicism such as loss of chromosome Y (LOY) on disease risk, could explain why males have a higher mortality compared to females, are more frequently affected by cancer and suggests that chromosome Y is important in processes beyond sex determination and sperm production. LOY in blood could become a predictive biomarker of male carcinogenesis.

## 296

**Genetic Heritability of Common Non-Hodgkin Lymphoma Subtypes.** *S.I. Berndt[1], L.M. Morton[1], S.S. Wang[2], L.R. Teras[3], S.L. Slager[4], J. Vijai[5], K. Smedby[6], G.M. Ferri[7], L. Miligi[8], C. Magnani[7], D. Albanes[1], A.R. Brooks-Wilson[9], E. Roman[10], A. Monnereau[11], P. Vineis[12], A. Nieters[13], B.M. Birmann[14], G.G. Giles[15], M.P. Purdue[1], B.K. Link[16], C.M. Vajdic[17], A. Zeleniuch-Jacquotte[18], C.F. Skibola[19], Y. Zhang[20], J.R. Cerhan[4], Z. Wang[1], N. Rothman[1], S.J. Chanock[1], J. Sampson[1] on behalf of the NHL GWAS Project.* 1) Division of Cancer Epidemiology & Genetics, National Cancer Institute, Rockville, MD, USA; 2) Department of Cancer Etiology, City of Hope Beckman Research Institute, Duarte, CA, USA; 3) Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA; 4) Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA; 5) Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York, USA; 6) Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden; 7) Interdisciplinary Department of Medicine, University of Bari, Policlinico - Piazza G. Cesare 11, Bari, Italy; 8) Environmental and Occupational Epidemiology Unit, Cancer Prevention and Research Institute (ISPO), Via delle Oblate 2, Florence, Italy; 9) Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada; 10) Department of Health Sciences, University of York, York, United Kingdom; 11) Environmental epidemiology of Cancer Group, Inserm, Centre for research in Epidemiology and Population Health (CESP), U1018, Villejuif, France; 12) Human Genetics Foundation, Turin, Italy; 13) Center of Chronic Immunodeficiency, University Medical Center Freiburg, Freiburg, Germany; 14) Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA; 15) Cancer Epidemiology Centre, Cancer Council Victoria, Carlton, Victoria, Australia; 16) Department of Internal Medicine, Carver College of Medicine, The University of Iowa, Iowa City, Iowa, USA; 17) Prince of Wales Clinical School, University of New South Wales, Sydney, New South Wales, Australia; 18) Department of Population Health, New York University School of Medicine, New York, New York, USA; 19) Department of Epidemiology, School of Public Health and Comprehensive Cancer Center, Birmingham, Alabama, USA; 20) Department of Environmental Health Sciences, Yale School of Public Health, New Haven, Connecticut, USA.

Non-Hodgkin lymphoma (NHL) is comprised of multiple subtypes with distinct morphologic and clinical features. Although these subtypes are hypothesized to be etiologically distinct and common variants for specific subtypes have been identified, no studies have comprehensively evaluated the contribution of common genetic variants to the heritability of individual subtypes. We, therefore, used genome-wide association data from a large, multicenter study of the four most common subtypes of NHL to estimate the contribution of common SNPs to the heritability of NHL. Participants were genotyped using the Illumina OmniExpress, and data were available from 2,179 cases of chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL), 2,142 cases of follicular lymphoma (FL), 2,661 cases of diffuse large B-cell lymphoma (DLBCL), 825 cases of marginal zone lymphoma (MZL) and 6,221 controls of European ancestry. We utilized genome-wide complex trait analysis (GCTA) to estimate the proportion of phenotypic variance explained on a liability scale and to estimate the shared heritability between subtypes, adjusting for age, sex, study, and principal components. The proportion of variance explained for the four NHL subtypes combined was estimated to be 11.9% (95% CI: 8-17%) with substantial heterogeneity among NHL subtypes: 35.5% (95% CI: 24-50%) for CLL/SLL, 20.8% (95% CI: 12-32%) for FL, 10.4% (95% CI: 4-18%) for DLBCL, and 8.3% (95% CI: 0-23%) for MZL. Small but non-significant differences were observed by sex for the individual subtypes. Although the precision in the estimates was limited, an examination of the shared heritability between subtypes showed modest genetic correlations ranging from 0.20-0.86. FL and DLBCL, for example, showed a genetic correlation of 0.57 (±0.28). Exclusion of the HLA region, did not substantially affect the estimates for the individual subtypes or the shared heritability observed between subtypes, suggesting that the HLA region captures only a small portion of the variance explained. Overall, this study indicates that although there is some shared heritability across NHL subtypes, the genetic architecture of individual NHL subtypes is distinct and consideration of NHL subtypes is essential for discovering new variants associated with susceptibility.

## 297

**A genome-wide scan identifies *NFIB* as important for metastasis in osteosarcoma patients.** *L. Mirabello[1], R. Koster[1], L. Spector[2], O.A. Panagiotou[1], P.S. Meltzer[3], B. Moriarty[2], D. Largaespada[2], N. Pankratz[2], J.M. Gastier-Foster[4], R. Gorlick[5], C. Khanna[3], A.M. Flanagan[6,7], R. Tirabosco[7], I.L. Andrulis[8], N. Gokgoz[8], J.S. Wunder[8], A. Patiño-Garcia[9], F. Lecanda[9], L. Sierrasesúmaga[9], S.R.C. de Toledo[10], A.S. Petrilli[10], M. Serra[11], C. Hattinger[11], P. Picci[11], S. Wacholder[1], L. Helman[3], M. Yeager[12], R.N. Hoover[1], S.J. Chanock[1], S.A. Savage[1].* 1) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA; 2) University of Minnesota, 420 Delaware Street, Minneapolis, MN, USA; 3) Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA; 4) Nationwide Children's Hospital, and The Ohio State University Department of Pathology and Pediatrics, 700 Children's Dr., C0988A, Columbus, OH, USA; 5) Albert Einstein College of Medicine, The Children's Hospital at Montefiore, 3415 Bainbridge Avenue, Rosenthal Room 300, Bronx, NY, USA; 6) UCL Cancer Institute, Huntley Street, London WC1E 6BT, UK; 7) Royal National Orthopaedic Hospital NHS Trust, Stanmore, Middlesex HA7 4LP, UK; 8) University of Toronto, Lunenfeld-Tanenbaum Research Institute, Mt Sinai Hospital, 600 University Ave., Toronto, Ontario, Canada, M5G 1X5; 9) Department Of Pediatrics, University Clinic of Navarra, Universidad de Navarra, Pío XII 36, 31080 Pamplona, Spain; 10) Pediatric Oncology Institute, GRAACC/UNIFESP, Rua Botucatu, 743 8º andar Laboratorio de Genética, CEP 04023-061 São Paulo SP Brazil; 11) Laboratory of Experimental Oncology, Orthopaedic Rizzoli Institute, Bologna, Italy; 12) Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, MD, USA.

Osteosarcoma (OS), the most common primary bone malignancy, typically occurs in adolescents and young adults. We recently completed the first international, multi-institutional genome-wide association study (GWAS) of OS and identified two novel loci associated with susceptibility to OS. A defining feature of OS is the high rate of metastasis to distant secondary sites, and patients presenting with metastases have a very poor prognosis. We have assembled clinical outcome data on OS cases accrued for our previous GWAS, and independent replication sets, to conduct the first GWAS of metastasis in 860 OS cases. Genotyping of germline genomic DNA was conducted using the Illumina OmniExpress SNP chip and quality control filters applied. Associations between SNPs and the risk of metastasis at diagnosis were estimated using an adjusted logistic regression log-additive genetic model. We identified one locus associated with metastasis that approached genome-wide significance in our discovery analysis of 541 European cases. Specifically, six linked intronic SNPs in the nuclear factor I/B gene (*NFIB*) on chromosomes 9p24.1 were significantly associated with an increased risk of metastasis (rs2890982, OR 2.65, $P=5.8\times10^{-7}$). To further explore this locus, we imputed SNPs across a 1 Mb region centered on the index SNP. A subsequent adjusted association analysis identified new signals that were substantially stronger than the genotyped SNPs. The top SNPs and the imputed region were replicated in three independent OS populations (319 cases). In the fixed effects meta-analysis of 860 cases, the 9p24.1 locus was strongly associated with susceptibility to metastasis at diagnosis (OR 2.49, 95% CI 1.78-3.52, $P=4.6\times10^{-8}$). We genotyped two of the linked *NFIB* SNPs and examined NFIB expression in 15 human OS cell lines. The variant allele was associated with a significant change in expression (P=0.006). We further performed proliferation and migration assays using human OS cell lines and siRNA knockdown of NFIB, and observed a significant change in migration and proliferation after knockdown of NFIB. We have identified a novel locus associated with risk of metastasis at diagnosis in OS patients, and our data suggests that variation in the *NFIB* gene is important for metastasis. We are further evaluating the function of this region and other top regions to advance our understanding of the role of germline genetics in OS metastasis.

## 298

**Enrichment of colorectal cancer associations in functional regions: insight for combining ENCODE and Roadmap Epigenomics data in the analysis of whole genome sequencing-imputed GWAS.** *S. Rosse[1], P. Auer[2,1], T. Harriason[1], C. Carlson[1], C. Qu[1], G.R. Abecasis[3], S.I. Berndt[4], S. Bézieau[5], H. Brenner[6], G. Casey[7], A.T. Chan[8], J. Chang-Claude[9], S. Chen[3], S. Jiao[1], C.M. Hutter[10], L. Le Marchand[11], S.M. Leal[12], P.A. Newcomb[1], M.L. Slattery[13], J. Smith[14], E. White[1], B.W. Zanke[15], U. Peters[1], D.A. Nickerson[14], A. Kundaje[16, 17], L. Hsu[1].* 1) Public Health Genetics, Fred Hutchinson Cancer Research Center, Seattle, WA; 2) School of Public Health, University of Wisconsin, Milwaukee, WI; 3) Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan; 4) Division of Cancer Epidemiology and Genetics, NCI, Bethesda, MD; 5) Service de Génétique Médicale, CHU Nantes, Nantes, France; 6) Division of Clinical Epidemiology and Aging Research, German Cancer Consortium (DKTK), Heidelberg, Germany; 7) Department of Preventive Medicine,University of Southern California, Keck School of Medicine, Los Angeles, CA; 8) Division of Gastroenterology,Massachusetts General Hospital and Harvard Medical School, Boston, MA; 9) Division of Cancer Epidemiology,German Cancer Research Center, Heidelberg, Germany; 10) Division of Cancer Control and Population Sciences, NCI, Bethesda, MD; 11) Epidemiology, University of Hawaii Cancer Center, Honolulu, HI; 12) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 13) Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, UT; 14) Department of Genome Sciences and Howard Hughes Medical Institute, University of Washington, Seattle, WA; 15) Division of Hematology, Faculty of Medicine, The University of Ottawa, Ottawa, ON; 16) Department of Genetics, Stanford University, Stanford, CA; 17) Department of Computer Science, Stanford University, Stanford, CA.

To investigate the role of low frequency genetic variation in colorectal cancer (CRC) susceptibility, the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) and the Colorectal Cancer Family Registry (CCFR) conducted genome-wide association studies (GWAS) of 12,661 CRC cases and 14,361 controls with imputation using a whole-genome sequence (6x coverage) reference panel of 610 CRC cases and 309 controls. These data provide a unique opportunity to investigate less-frequent and rare variants with minor allele frequencies (MAF) 0.1-5%, which contribute to the majority of the variation in the genome. However, as statistical power is limited for detection of rare variant associations we aim to address this limitation by incorporating functional data. To evaluate whether these data would be useful in discovering novel rare variant association we used ENCODE and Roadmap Epigenomics Project data to define hypothesized enhancer, promoter, and other regulatory regions in colon and rectal tissue and assessed whether these regions were enriched in GECCO-CCFR GWAS results for both aggregate-rare-variant and common-single-variant association tests. To define functional elements across the genome, we used chromatin structure across cell types and tissues to improve the resolution of CRC epigenetic signatures (uniform ChIP-seq signals). We then identified sets of rare variants and single common variants that overlapped with predicted regulatory regions to investigate enrichment of association p-values relative to those in non-CRC-regulatory regions across the genome using the Kolmogorov-Smirnov test. We found significant enrichment in regions that overlapped with predicted CRC regulatory regions for both rare variant (p=0.0001) and common variants (p=2.69e-27). In addition, we found that regulatory prediction using high resolution ChIP-seq data corresponded with in vitro identification of three previously identified CRC GWAS loci (MYC-rs6983267, CDH1-rs16260, and COLCA1-rs7130173). These results suggest that functional insight of cell-type specific regulatory mechanisms can inform the discovery of genetic associations and may be useful for incorporation into future association testing.

## 299

**Frequency and phenotypic spectrum of germline mutations in POLE and seven other polymerase genes in patients with colorectal adenomas and carcinomas.** *I. Spier[1], S. Holzapfel[1], J. Altmüller[2,3], B. Zhao[4], S. Horpaopan[1], S. Vogt[1,5], S. Chen[4], M. Morak[6,7], S. Raeder[1], K. Kayser[1], D. Stienen[1], R. Adam[1], P. Nürnberg[2], G. Plotz[8], E. Holinski-Feder[6,7], R.P. Lifton[4], H. Thiele[2], P. Hoffmann[1,9,10], V. Steinke[1], S. Aretz[1].* 1) Institute of Human Genetics, University of Bonn, Bonn, Germany; 2) Cologne Center for Genomics, University of Cologne, Germany; 3) Institute of Human Genetics, University of Cologne, Germany; 4) Departments of Genetics, Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, USA; 5) MVZ Dr. Eberhard & Partner, Dortmund, Germany; 6) Medizinische Klinik - Campus Innenstadt, Klinikum der LMU, Munich, Germany; 7) MGZ - Center of Medical Genetics, Munich, Germany; 8) Medizinische Klinik 1, Biomedical Research Laboratory, University of Frankfurt, Germany; 9) Department of Genomics, Life & Brain Center, University of Bonn, Germany; 10) Division of Medical Genetics, University Hospital Basel and Department of Biomedicine, University of Basel, Switzerland.

Purpose: In a number of families with colorectal adenomatous polyposis or suspected Lynch syndrome (HNPCC), no germline alteration in the APC, MUTYH, or mismatch repair (MMR) genes are found. Missense mutations in the polymerase genes POLE and POLD1 have recently been identified as rare cause of multiple colorectal adenomas and carcinomas, a condition termed Polymerase proofreading-associated polyposis (PPAP). The aim of the present study was to evaluate the clinical relevance and phenotypic spectrum of polymerase germline mutations. Methods: Targeted next-generation sequencing of the polymerase genes POLD1, POLD2, POLD3, POLD4, POLE, POLE2, POLE3, and POLE4 was performed (Illumina platform) using a sample of 241 unrelated patients (194 mutation negative polyposis patients and 47 familial colorectal carcinoma (CRC) cases with microsatellite stable tumours meeting the Amsterdam criteria). Data analysis was done by standard protocols using the VARBANK pipeline (CCG, Cologne). Results: The previously described pathogenic POLE mutation c.1270C>G;p.Leu424Val was detected in three families. The mutation was present in 1% (3/241) of all unrelated patients, and 7% (2/28) of familial polyposis cases. The colorectal phenotype in 13 affected mutation carriers (age at diagnosis 16-63 years) ranged from typical adenomatous polyposis to a Lynch syndrome-like manifestation, with high intrafamilial variability. The occurrence of multiple CRCs was common. Most patients (63%) had duodenal adenomas, and one case of duodenal carcinoma was reported. Additionally, various extraintestinal lesions including ovarian cancer and glioblastomas were evident. Nine further putative pathogenic variants were identified in four polymerase genes. The most promising was a de novo missense mutation in POLE (c.1306C>T;p.Pro436Ser). Conclusion: A PPAP was identified in a substantial number of our well characterized sample of polyposis and familial colorectal cancer patients. Screening for these mutations should therefore be considered, particularly in unexplained familial cases. The present study broadens the phenotypic spectrum of PPAP to duodenal adenomas and carcinomas, and demonstrated a considerable clinical overlap between tumor syndromes based on mutations in DNA repair genes. In addition, we identified novel, potentially pathogenic variants in four polymerase genes. (Supported by German Cancer Aid, BONFOR programme of the University of Bonn and NIH Centers for Mendelian Genomics).

## 300

**Impact of genetic testing on reducing colorectal cancer.** *D.W. Neklason[1, 2, 7], H.A. Hanson[2,3,7], C. Schaefer[2,7], G. Mineau[2, 6, 7], M.F. Leppert[5, 7], R.W. Burt[2,1,7], K.R. Smith[2,4,7].* 1) Internal Medicine, University of Utah, Salt Lake City, UT., USA; 2) Huntsman Cancer Institute at University of Utah, Salt Lake City, UT, USA; 3) Family and Preventive Medicine, University of Utah, Salt Lake City, UT, USA; 4) Family and Consumer Studies, University of Utah, Salt Lake City, UT, USA; 5) Human Genetics, University of Utah, Salt Lake City, UT, USA; 6) Oncological Sciences, University of Utah, Salt Lake City, UT, USA; 7) Population Sciences, Huntsman Cancer Institute, Salt Lake City, UT, USA.

In an attempt to measure the impact of genetic testing in disease prevention, we had a unique opportunity to follow a large family with >7000 individuals over 30 years. The family has an attenuated form of familial adenomatous polyposis leading to a 69% lifetime risk of colorectal cancer (CRC) due to mutation in the APC gene. Utah statewide cancer registry is linked to genealogies through the Utah Population Database which allowed measurement of hazard ratios for CRC over time. Subjects were enrolled in research starting in 1980's, after discovery of the involved gene in 1990's, and again in 2000's. The specific APC mutation was tested in 1005 participants (186 positive; 818 negative) and 669 went on to have clinical confirmation with genetic counseling around 1993 (144 positive; 525 negative). Two branches of the family, B and E had the mutation, whereas there was no evidence of the mutation in over 500 individuals tested from branch D. Analysis was restricted to individuals residing in Utah after 1963 (accurate cancer records) and excluded those < 30 years, deceased prior to age 30 or absence of a birth date. The final sample included 2,550 individuals (branch D=640, B= 250, E=1660). Although not everyone in the family chose to participate in research, we assume that genetic counseling, education and communication within the family had potential to modify behaviors within each branch. We hypothesize that the risk will decrease over time with interventions. CRC risk was evaluated by family branch using Cox proportional hazard models. We show that when controlling for sex and birth year, individuals in branches B and E combined have a 5 fold increase in the risk of CRC (p=0.02) over branch D. We then evaluated separate hazards for 1963 to 1982 (pre-hereditary knowledge), 1983 to 1993 (pre-genetic diagnosis), and >1993 (post-genetic diagnosis) in branches B/E versus branch D. This study design is controlled with the comparison branch D undergoing the same intervention, but not harboring the mutation. The resulting hazard ratios decrease over these time windows from 4.41 to 2.39 to 2.23 with a p-value of 0.001. In conclusion, we are able to demonstrate the impact of the presence of a positive genetic test combined with education and counseling to reduce cancer risk in a family setting. Although the risk of CRC is dramatically reduced, it is not reduced to the level of the branch D, suggesting that there is opportunity for additional interventions.

## 301

**Performance of Multi-Gene Panels for Familial Cancer Screening in Clinical Cases: The ColoSeq and BROCA Experience.** *B.H. Shirts[1], S. Casadei[2], A. Jacobson[1], E. Turner[1], J.F. Tait[1], M.C. King[2,3], T. Walsh[2], C.C. Pritchard[1].* 1) Laboratory Medicine, University of Washington, Seattle, WA; 2) Medicine, Division of Medical Genetics, University of Washington, Seattle, WA; 3) Genome Sciences, University of Washington, Seattle, WA.

We report the clinical performance of next-generation sequencing for hereditary cancer risk using ColoSeq and BROCA. This report includes all 356 sequential cases tested with the 19 gene ColoSeq panel and all 928 sequential cases tested with the 51 gene BROCA panel between November 2011 and February 2014. We classify variants based on consensus of four experts who consider all published and database information. We use a Bayesian framework integrating all available patient and family cancer history. Of providers ordering ColoSeq or BROCA, >95% were genetics specialists and 98% provided cancer history and/or pedigree data. Rates of positive results differed substantially by patient and family history. Actionable mutations in BROCA genes other than BRCA1 and BRCA2 were detected in 10% of patients (with or without cancer) with previous testing for BRCA1 and BRCA2 and in 15% of breast cancer patients with no previous sequencing. Overall, of 928 BROCA tests, 118 (13%) revealed pathogenic or very likely pathogenic variants, and 94 (10%) yielded either variants of uncertain significance in established cancer predisposition genes or clearly deleterious variants in emerging genes. Of 356 ColoSeq tests, 58 (16%) revealed pathogenic or likely pathogenic variants, and 48 (13%) carried either variants of uncertain significance in established cancer risk genes or clearly deleterious variants in emerging genes. Selective providers had positive results at a rate substantially higher than average for both ColoSeq and BROCA tests. We evaluated the benefit of the multi-specialist review and signout process in a subset of cases. Multi-specialist review of all variants increased sensitivity of testing, led to identification of additional actionable variants in 2-5% of cases, and decreased the number of variants of uncertain significance reported. Comprehensive multi-gene testing for patients at risk for inherited cancer syndromes is a complex process that benefits from accurate clinical histories and multiple skilled geneticists evaluating each variant. In our experience, the proportion of patients with actionable mutations depends on the panel of genes tested, patient population, provider selectivity, and combined experience of geneticists evaluating variants. We find that the number of variants assigned "uncertain significance" is less dependent on the number of genes tested than on patient population and combined experience of the geneticists evaluating the variants.

## 302

**Unanticipated germline cancer susceptibility mutations identified by clinical exome sequencing of sequentially diagnosed pediatric solid tumor patients: the BASIC3 study.** *S.E. Plon[1,2,3,4], S. Scollon[1], K. Bergstrom[1], T. Wang[4], R.A. Kerstein[1], U. Ramamurthy[1], D.M. Muzny[3], S.G. Hilsenbeck[4], Y. Yang[2], C.M. Eng[2], R.A. Gibbs[2,3], D.W. Parsons[1,3,4].* 1) Dept Pediatrics, Baylor College of Medicine, Houston, TX; 2) Dept Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 3) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 4) Dan L Duncan Cancer Center, Baylor College of Medicine, Houston, TX.

**Background**: Germline data from clinical whole exome sequencing (WES) of cancer patients can provide clinically relevant findings with regard to cancer susceptibility. We prospectively determined whether germline cancer susceptibility findings in children can be predicted based on pathologic cancer diagnosis and family history information. **Methods**: The BASIC3 study investigates the clinical implementation of CLIA-certified germline and tumor WES in an ethnically diverse cohort of sequentially diagnosed children with CNS and non-CNS solid tumors. At subject entry, the study geneticist and genetic counselors record whether specific cancer genetic tests would be considered using current clinical practice based on the cancer diagnosis, age and family history. These data were then compared with the pathogenic mutations and variants of uncertain significance (VUS) in cancer genes identified by clinical WES. **Results**: The first 115 patients enrolled comprised 41 CNS and 74 non-CNS solid tumor diagnoses. Based on the information provided genetic testing was considered for 40/115 (35%) patients, including TP53 in 23 patients. Germline WES revealed pathogenic mutations in dominant cancer susceptibility genes in 10 patients, including 4 that were suggested at study entry (*TP53, MSH2, DICER1, VHL*), 6 that were not (*BRCA1*x2, *BRCA2, CHEK2, APC*, mosaic *WT1*) and one patient with liver disease and liver cancer homozygous for a novel *TJP2* mutation. Eight patients were reported to carry single truncating mutations in recessive cancer disorder genes such as *FANCL* and *WRAP53* all without clinical evidence of the disorder. Nearly all patients harbored VUS in cancer genes (98%; median = 3), including one likely pathogenic *TP53* mutation and one Wilms tumor patient compound heterozygous for rare missense variants in *DIS3L2*. **Conclusions**: Diagnostic germline mutations were obtained in 9.6% (11/115) of sequentially diagnosed children with CNS and non-CNS solid tumors. Five of these patients carry mutations generally associated with adult cancer diagnoses and thus not prospectively recommended for testing. Comparison with 2000 WES results of non-cancer patients from the same lab suggests that *BRCA1/2* mutations maybe enriched in this childhood cancer cohort (p=0.05 exact Fischer's test). Further study is required to determine whether these cancer susceptibility mutations and variants are significantly enriched in childhood cancer patients. Supported by NHGRI/NCI-1U01HG006485.

## 303

**An Evidence-Based Dosage Sensitivity Map Towards Defining the Clinical Genome.** *E. Riggs[1], E. Andersen[2], B. Hong[2], H. Kearney[3], G. Hislop[4], S. Kantarci[5], D. Pineda-Alvarez[6], U. Maye[7], D. McMullan[8], M. Serrano[2], I. Simonic[9], S. South[2], M. Speevak[10], K. Smith[11], J. Stavropoulos[12], K. Wain[3], S. Aradhya[13], E. Thorland[3], C. Martin[1] on behalf of the Clinical Genome (ClinGen) Resource.* 1) Autism and Developmental Medicine Institute, Geisinger Health System, Lewisburg, PA; 2) Dept of Pediatrics and Pathology, ARUP Laboratories, Univ of Utah, Salt Lake City, UT; 3) Dept of Lab Medicine & Pathology, Mayo Clinic, Rochester, MN; 4) Ninewells Hospital, Dundee, United Kingdom; 5) Department of Pathology and Laboratory Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA; 6) GeneDx, Gaithersburg, MD; 7) Liverpool Women's Hospital, Liverpool, United Kingdom; 8) West Midlands Regional Genetics Laboratories, Birmingham, United Kingdom; 9) Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom; 10) CVH Site, Trillium Health Partners; Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada; 11) Sheffield Children's Hospital, Sheffield, United Kingdom; 12) Department of Pediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Canada; 13) InVitae, San Francisco, CA.

The Clinical Genome (ClinGen) Resource is an NIH-funded program dedicated to developing community resources for sharing, analyzing, and curating human genomic variant data to advance genomic medicine. Continuing work initiated as part of the International Standards for Cytogenomic Arrays (ISCA) Consortium and the International Collaboration for Clinical Genomics (ICCG), the ClinGen Structural Variation Working Group is creating a dynamic, genome-wide dosage sensitivity map towards defining the clinical genome. We are using an evidence-based process to evaluate whether haploinsufficiency (loss) or triplosensitivity (gain) of particular genes and genomic regions result in clinically demonstrable phenotypes. We defined five categories for level of evidence including: sufficient evidence, emerging evidence, little evidence, no evidence and dosage sensitivity unlikely. We have currently evaluated a total of 541 genes/genomic regions. Of these, a total of 257 met our criteria for sufficient evidence supporting a role for dosage sensitivity in human disease (47%). These genes/genomic regions are now included in our curated catalog of "known" pathogenic copy number losses and gains, which is continually updated as new genes/regions are reviewed and is publically available at www.ncbi.nlm.nih.gov/projects/dbvar/ISCA/ or through dbVar (study nstd45). This valuable resource can be used as a guide for clinical interpretation of copy number variants (CNV) throughout the genome. For example, CNVs encompassing one or more genes with sufficient evidence for dosage sensitivity should be interpreted as "pathogenic", whereas those containing genes with only limited or no evidence might be interpreted as "uncertain". The use of this resource extends beyond the individual laboratory setting, and will be used as part of the ClinGen project to mitigate inter-laboratory conflicts in CNV data submitted from clinical and research laboratories to the ClinVar and dbGaP databases at NCBI. These CNVs will be compared against our evolving dosage sensitivity map and evaluated by expert review. This process will allow us to resolve some of the current discrepancies in available data, providing additional curated resources for interpretation of CNVs throughout the genome. The ongoing evaluation of dosage sensitivity using an evidence-based process will lead to standardized clinical interpretations and play a critical role in incorporating genomic medicine into patient care.

## 304

**Next-generation sequencing of duplication CNVs reveals that most are tandem and some disrupt genes at breakpoints.** *K. Rudd, K.E. Hermetz, B. Weckselblatt, S. Newman.* Dept Human Gen, Emory Univ Sch Med, Atlanta, GA.

Copy number variation (CNV) in the form of large deletions and duplications is a major cause of neurodevelopmental disorders. Though duplications are a common finding in cytogenetic testing, the clinical consequences of duplication CNVs are particularly difficult to interpret because the genomic structure and precise breakpoints are unknown. Duplications are usually less deleterious than deletions, and could lead to disease through gene triplosensitivity, gene fusion, and/or gene disruption. We fine-mapped clinically relevant duplications in 189 subjects referred for cytogenetic testing at Emory Genetics Laboratory. Using sequence capture, massively parallel paired-end sequencing, and confirmatory Sanger sequencing, we sequenced breakpoints of 90 non-recurrent duplications to the basepair. These large duplications (27 kb- 25 Mb; median 600 kb) exist as tandem duplications (87%), insertional translocations (3%), or complex rearrangements (10%). We identified two major classes of complex duplications, those that connect two adjacent duplications (dup-dup) and those with triplications embedded in duplications (dup-trip). Sequencing breakpoint junctions revealed that both dup-dup and dup-trip structures can be tandem or inverted. Though duplicated genes may be disrupted by breakpoints, most duplications retain intact copies of genes at duplication boundaries. Few duplication junctions are predicted to result in novel isoforms of single genes (2%) or in-frame fusions of two different genes (11%). One of the three insertional translocations disrupted a gene at the insertion site and is predicted to result in an in-frame fusion of exons 1-3 of USP20 and exons 3-45 of COL4A6. Intragenic duplications of CNTN4 and TCOF1 disrupted reading frames, while duplications within OPCML, DMD, and PAFAH1B1 were completely intronic. We detected a 324-kb intergenic duplication that fuses exons 1-6 of SOS1 and exons 2-33 of MAP4K3 in-frame in a child with some features of Noonan syndrome. Our large-scale analysis of genomic gains showed that most duplications are local and tandem. Sequence analysis of duplications revealed genomic structures (triplications, inversions) and gene alterations (disruptions, fusions) that are not detectable by clinical array testing alone. These data are essential to interpret the phenotypic impact of duplications in the human genome.

## 305

**Balanced Chromosome Rearrangements Rapidly Annotate the Morbid Human Genome.** *T. Kammin[1], K.E. Wong[1], B.B. Currall[1,2], Z. Ordulu[1,2], H. Brand[2,3], V. Pillalamarri[3], C. Hanscom[3], I. Blumenthal[3], J.F. Gusella[2,3,4,5,6], E.C. Liao[7,8,9], M.E. Talkowski[2,3,4,6], C.C. Morton[1,2,6,10].* 1) Department of Obstetrics, Gynecology, and Reproductive Biology, Brigham and Women's Hospital, Boston, MA; 2) Harvard Medical School, Boston, MA; 3) Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; 4) Department of Neurology, Massachusetts General Hospital, Boston, MA; 5) Department of Genetics, Harvard Medical School, Boston, MA; 6) Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA; 7) Center for Regenerative Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA; 8) Division of Plastic and Reconstructive Surgery, MA General Hospital, Harvard Medical School, Boston, MA; 9) Harvard Stem Cell Institute, Boston, MA; 10) Department of Pathology, Brigham and Women's Hospital, Boston, MA.

The Developmental Genome Anatomy Project (DGAP) is a collaborative effort to identify novel genes critical for human development and disease pathogenesis through study of genetic loci disrupted or dysregulated by balanced chromosomal rearrangements. We perform Whole Genome Sequencing (WGS) of rearrangement breakpoints of every subject and use cellular, mouse and zebrafish models to validate the role of candidate genes. To date, we have enrolled over 295 families and sequenced 131 cases, identifying over 175 genes at 76% of breakpoints. We summarize some of our recent findings and scientific surprises, focusing on three DGAP cases to highlight the importance of analyzing apparently balanced chromosomal rearrangements to discover the etiology of rare diseases. DGAP162 presents with severe developmental delay; sequencing showed disruption of *LINC00299* resulting in its upregulated expression and revealing that a noncoding RNA can be associated with a severe developmental phenotype. DGAP056 disrupted *C2orf43*, associated with craniofacial abnormalities, sensorineural hearing loss and early onset prostate cancer (age 38 years). Our mouse KO model supports the role of this gene in hearing loss and prostate cancer. Identification of C2orf43 as a cholesterol esterase illuminates a previously unrecognized metabolic pathway common to deafness and cancer. Implementation of our large insert "jumping library" WGS approach to explore *de novo* balanced rearrangements detected prenatally permits reporting of nucleotide level precision results within two weeks, of particular value in these time-sensitive situations. Five cases analyzed to date have revealed that, compared to conventional karyotyping methods, WGS approaches are an accurate and timely method of refining chromosome rearrangements. We discuss DGAP247 and report the impact of a prenatal next-generation sequencing result from the Mother's perspective. DGAP allows us to annotate poorly defined areas of the genome and discover previously unknown developmental networks. The impact of DGAP is also therapeutic for families, providing a definitive diagnosis in an often unique clinical disorder.

## 306

**THE PERPLEXING PREVALENCE OF FAMILIAL NESTED 22q11.2 DELETIONS.** *D.M. McDonald-McGinn, L. DiCairano, J.T. Goulet, A. Capezzuto, A. Krajewski, C. Franconi, M. McNamara, D.E. McGinn, B.S. Emanuel, E.H. Zackai.* Div Human Gen, Children's Hosp Philadelphia, Philadelphia, PA.

Introduction: The 22q11.2 deletion is most often de novo and classically extends from segmental duplications A-D including loss of TBX1 an important developmental gene within the A-B region causally associated with typical features such as congenital heart disease. In contrast, TBX1 is present in patients with B-D and C-D nested deletions where overlapping features still include conotruncal anomalies and familial inheritance is common. Here we report the perplexing prevalence of familial nested 22q11.2 deletions, as well as associated phenotypes. Methods: 603/1188 individuals with 22q11.2DS (50%) in our cohort had deletion sizing by enhanced FISH, CGH, GWAS or MLPA. Phenotypic features were catalogued prospectively under IRB approval. Results: 526/603 patients (87%) had A-D deletions whereas 13% were nested. Of the latter, whose deletions did not include the A-B/TBX1 region, 20 probands had B-D and 3 C-D deletions. Parental studies in 15/20 B-D and 2/3 C-D families revealed 9/15 (60%) B-D and 1/2 C-D deletions were familial. Of these, 5/9 B-D deletions were paternally inherited, as was the familial C-D deletion. Anomalies typically associated with the A-D deletion were also identified with B-D deletions, albeit less frequently, including: CHD (31%); VPI/SMCP (38%); chronic infection (46%); hypocalcemia/growth hormone deficiency (24%); GERD (36%); and learning differences (61%). The C-D deletion also resulted in typical features including: CHD; chronic infection; hypernasal speech; and developmental delay. 3/9 parents with B-D deletions had significant findings including developmental delay, ADHD and short stature. The father with the C-D deletion has a Master's Degree in Education but reported lifelong learning deficits in math. Conclusions: In the largest series to date, we confirm that the majority of 22q11.2 deletions extend from A-D and are de novo. However, we report the noteworthy finding that nested B-D and C-D deletions are frequently familial. This may reflect a milder overall phenotype or the effect of modifier genes but this observation is critical for providing accurate genetic counseling. Moreover, this may well inform our understanding of developmental genes beyond TBX1 such as CRKL1 and SNAP29 or position effects to explain the phenotypic overlap with the standard deletion but importantly, based on this data, we urge practitioners to be vigilant in screening parents of all affected children regardless of their clinical history.

## 307

**9q33.3q34.11 microdeletion: delineation of a new contiguous gene syndrome including the STXBP1, LMX1B and ENG genes using reverse phenotyping.** *S. Nambot[1], A. Mosca Boidron[2], A. Masurel[1], M. Lefebvre[1], N. Marle[2], J. Thevenon[1,2], J. De Montléon[3], S. Perez Martin[3], M. Chouchane[3], E. Sapin[4], J. Metaizeau[4], V. Dulieu[5], F. Huet[2], C. Thauvin Robinet[1], L. Chatel[6], V. Abadie[7], G. Plessis[8], J. Andrieux[9], P. Jouk[10], G. Billy Lopez[10], C. Coutton[11], F. Morice Picard[12], M. Delrue[12], C. Rooryck Thambo[13], L. Faivre[1,2].* 1) Centre de Génétique et Centre de Référence Maladies Rares, Anomalies du Développement et Syndromes Malformatifs de l'Inter région Est, Hôpital d'Enfants, CHU Dijon, France; 2) Laboratoire de Cytogénétique, Plateau Technique de Biologie, CHU Dijon, France; 3) Service de Pédiatrie 1, Hôpital d'Enfants, CHU Dijon, France; 4) Service de Chirurgie pédiatrique, Hôpital d'Enfants, CHU Dijon, France; 5) 5Service de Soins de suite et Rééducation pédiatrique, Pôle Rééducation Réadaptation, CHU Dijon, France; 6) Service de Psychiatrie de l'enfant, CHU Dijon, France; 7) Service de Pédiatrie générale, Hôpital Necker, Paris, France; 8) Centre de Compétence des Anomalies du Développement, CHU Caen, France; 9) Laboratoire de Génétique Médicale, Hôpital Jeanne de Flandre CHRU Lille, France; 10) Service de Génétique Clinique, Département de Génétique et Procréation, Hôpital Couple Enfant, CHU Grenoble, France; 11) Laboratoire de Génétique Chromosomique, Département de Génétique et Procréation, Hôpital Couple Enfant, CHU Grenoble, France; 12) Centre de Référence des Anomalies du Développement et Syndromes malformatifs , CHU Bordeaux, France; 13) Laboratoire de Génétique Moléculaire, Plateau technique de Biologie Moléculaire, CHU Bordeaux, France.

Four patients, 3 females and 1 male aged 5 to 18 years, carrying a de novo overlapping 1.5, 3.1, 3.5 and 4.1 Mb deletion of chromosome 9q33.3q34.11 permitted to define a new contiguous gene syndrome. Patients display common clinical features including facial dysmorphism, epilepsy, intellectual deficiency of various degree and multiple congenital abnormalities. The analysis of the genes comprised in the deletions prompted us to use reverse phenotyping. The STXBP1 gene, in which de novo heterozygous mutations or deletions have been reported in patients with Ohtahara syndrome, was the best candidate to explain the cognitive and epileptic phenotype. The LMX1B gene, in which heterozygous mutations or deletions have been reported with Nail-patella syndrome, explained the presence of nail dysplasia and bone malformations, in particular patellar abnormalities. The ENG gene, in which autosomal dominant mutations or deletions have been reported in patients with Hereditary hemorrhagic telangiectasia type 1, was likely responsible for epistaxis and cutaneous-mucous telangiectases described in the oldest patients. The NR5A1 gene, deleted in one patient only, was likely responsible of his genital malformations. A high genotype-phenotype correlation was found in these 4 patients, except for the remarkable facial dysmorphism, including prominent metopic ridge, large forehead, high arched eyebrows, strabismus, bulbous nose and small mouth. This systematic analysis of genes comprised in the deletion permitted to identify genes whose haploinsufficiency is expected to lead to disease manifestations and complications that will lead to a personalized follow-up, in particular for renal, eye, ears, vascular and neurologic manifestations.

## 308

***Alu* enriched genomic structure facilitates chromosome 17p13.3 region susceptibility to diverse and complex pathogenic copy number variations.** *S. Gu[1], B. Yuan[1], I.M. Campbell[1], A. Patel[1], C. Bacino[1], P. Stankiewicz[1], S.W. Cheung[1], W. Bi[1], J.R. Lupski[1,2].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Texas Children's Hospital, Houston, TX.

Copy number variants (CNVs) in human chromosome 17p13.3, a gene-rich region, are associated with various syndromes. Deletions of *PAFAH1B1* (encoding LIS1) cause classical lissencephaly, while deletions of contiguous genes including both *YWHAE* and *PAFAH1B1* cause Miller-Dieker syndrome with more severe lissencephaly and distinctive dysmorphic facial features. Duplications in 17p13.3 are also associated with diverse phenotypes. CNVs identified within 17p13.3 region are nonrecurrent, and mechanisms underlying these CNVs warrant further investigation. We identified 40 unrelated patients with copy number changes involving known disease genes or disease candidate genes within 17p13.3 using clinical chromosome microarrays. We performed customized high-density array comparative genomic hybridization (HD-aCGH) specifically interrogating the 17p13.3 region on all the patients with CNVs in this region, including thirteen patients with duplication, three with triplication, fourteen with deletion and ten with complex rearrangements. Breakpoint junctions were mapped at nucleotide resolution by PCR and DNA sequencing. In a total number of 28 breakpoint junctions determined, 17 (60.7%) had *Alu* repeats on both sides of the breakpoint junctions, while 4 breakpoint junctions (14.3%) had *Alu* repeats on one side of the breakpoint; thus, the majority of the CNVs appear to be *Alu*-facilitated events. The human 17p13.3 region has almost no segmental duplications, but this region is highly enriched for *Alu* repeats, with a fraction of around 30% comparing to the ~10% in the whole genome and ~18% on chromosome 17. Our studies suggest that *Alu* repeats may play an important role in the formation of nonrecurrent copy number changes and structural aberrations in 17p13.3.

## 309

**Decoding NF1 intragenic copy number changes.** *M. Hsiao[1], A. Piotrowski[1,2], T. Callens[1], C. Fu[1], L. Messiaen[1].* 1) Department of Genetics, University of Alabama at Birmingham, Birmingham, AL; 2) Medical University of Gdansk, Gdansk, Poland.

Genomic rearrangements are comprised of deletions, duplications, insertions, inversions and translocations causing both Mendelian and complex disorders. Currently, several major mechanisms causing genomic rearrangements have been proposed such as non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR). However, to what extent these mechanisms contribute to rare, locus-specific pathogenic copy number changes (CNCs) remains unexplored. Furthermore, only a few studies resolved these pathogenic alterations at nucleotide-level resolution. Through array Comparative Genomic Hybridization (aCGH) as well as breakpoint-spanning PCR, we have identified the breakpoints and characterized the likely rearrangement mechanism of the *NF1* intragenic CNCs in 76 unrelated subjects. Unlike the most typical recurrent rearrangement mediated by flanking low copy repeats (LCRs), *NF1* intragenic CNCs vary in size and location. Microhomology from 1 to 41 bp was found in 57 patients (75% of 76 characterized breakpoint junctions), suggesting that the predominant mechanisms are DNA replication-based and microhomology-mediated. NAHR between repetitive elements was found in 16 individuals (21%). Additionally, *Alu* elements led to 15 *Alu*-*Alu* recombinations, and were also involved in 12 non-recurrent rearrangements including *Alu* insertion, NHEJ and FoSTeS/MMBIR. Hence, *Alu* elements were involved in as many as 27 unrelated patients (36%), suggesting their crucial role in generating *NF1* intragenic CNCs. Furthermore, several *Alu* elements located in intron 2, 3, and 50, have demonstrated much stronger ability to mediate genomic rearrangements than other *Alu* elements. In addition to NAHR, NHEJ and FoSTeS/MMBIR, complicated rearrangements such as multiple NHEJ, multiple FoSTeS/MMBIR, serial replication slippage and *Alu* insertion were found in the *NF1* intragenic CNCs. Furthermore, repetitive elements and non-B DNA structures, according to *in silico* analysis, were significantly associated with genomic rearrangements. We propose that these features might increase susceptibility to DNA double strand break or replication stalling. This locus-centered study based on a large set of breakpoint identifications provides important clues for NF1 molecular etiology, and also serves as a paradigm for other disorders involving genomic rearrangement.

## 310

***De Novo DYRK1A* Point Mutations Cause Similar Phenotypes to Those Observed in Microdeletions including 21q22.13: Further Evidence for *DYRK1A*'s Critical Role in Brain Development.** *J. Ji[1], N. Dorrani[1], J. Mann[2], J.A. Martinez-Agosto[1], N. Gallant[1], J.A. Bernstein[3], N. Gomez-Ospina[3], L. Hudgins[3], L. Slattery[3], B. Isidor[4], E. Obersztyn[5], B. Wiśniowiecka-Kowalnik[5], M. Fox[1], H. Lee[1], J. Deignan[1], E. Vilain[1], S.F. Nelson[1], W. Grody[1], F. Quintero-Rivera[1].* 1) University of California, Los Angeles (UCLA), Los Angeles, CA; 2) Kaiser Permanente, Fresno, CA; 3) Stanford University School of Medicine, Stanford, CA; 4) Centre Hospitalier Universitaire de Nantes, Paris, France; 5) Institute of Mother and Child, Warsaw, Poland.

*DYRK1A* (dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A) is a highly conserved gene located in the Down syndrome critical region. It plays an important role in controlling brain growth through the regulation of neuronal proliferation and neurogenesis. Microdeletions of chromosome 21q22.12q22.3 that include *DYRK1A* (21q22.13) contribute to a spectrum of neurodevelopmental phenotypes; however, the impact of *DYRK1A* disruption has not been fully explored. To characterize the landscape of *DYRK1A* disruptions and their associated phenotypes, we identified seven individuals ( from 17 months to 7 years old ) with de novo disruptions of *DYRK1A*; three with heterozygous microdeletions (1.7 to 4.2 Mb), and four with point mutations (2 missense and 2 nonsense). The genotype-phenotype analysis of all published cases and our cohort of patients (N=20), revealed that phenotypes were largely indistinguishable between patients with the 21q22.12q22.3 microdeletion and those with translocation or mutation limited to *DYRK1A*. All patients shared primary microcephaly (OFC <3%, -2SD), severe intellectual disability (ID), developmental delay (DD), severe speech impairment, and distinct facial features. Seizures, brain abnormalities, ataxia/abnormal gait, and feeding difficulties during infancy were present in two thirds of all patients. A less common feature was autism spectrum disorder (30%). The severity of the microcephaly varied from -2 SD to -6 SD. While *Dyrk1a* (-/-) is embryonic lethal, microcephaly and DD have been observed in *Dyrk1a* (±) deficient mouse models, and it has been demonstrated that Dyrk1a plays a vital role in shaping the brain, controlling cell density and cell morphology, and regulating developmental pathways. Although we cannot fully exclude contribution of other genes to the phenotypes in patients with a microdeletion, point mutations in *DYRK1A* are sufficient to recapitulate the neurodevelopmental abnormalities observed in the microdeletion cases. Our study further demonstrates that haploinsufficiency of *DYRK1A* results in primary microcephaly and ID in humans, and that genome-wide testing (e.g. exome sequencing, microarray) should be considered in patients with these phenotypes. Our report represents the largest cohort of patients with *DYRK1A* disruptions, and is the first attempt to achieve consistent genotype-phenotype correlations in the distinct group of subjects with 21q22.13 microdeletions and *DYRK1A* mutations.

## 311

**Prenatal whole genome SNP array diagnosis as a first-line test: nature and prevalence of abnormal results in phenotypically normal and abnormal fetuses.** *F.A.T. de Vries, M.I. Srebniak, L.C.P. Govaerts, K.E.M. Diderich, R.J.H. Galjaard, A.M.S. Joosten, A.R.M. Van Opstal.* Clinical Genetics, ErasmusMC, Rotterdam, Netherlands.

**Background**: After initially applying SNP array in selected fetuses with ultrasound (US) abnormalities, since June 2011, we routinely perform array analysis in all cases of US abnormalities and since July 2012 in all samples that we receive in our laboratory for prenatal cytogenetic studies, irrespective of the indication. Here we show the nature and prevalence of abnormal array results in cases with and without US anomalies. **Methods**: After excluding the most common trisomies and triploidy by rapid aneuploidy detection (RAD), we performed HumanCytoSNP-12 (Illumina) array on uncultured cells of 1047 cases with fetal ultrasound anomalies and of 1408 cases of uneventful pregnancies. Clinically relevant array results were classified as proposed before in causative findings (CAU) (i.e. fitting the indication/phenotype), unexpected diagnoses (UD) (i.e. NOT fitting the indication/phenotype) and susceptibility loci (SL) for neurodevelopmental disorders (Srebniak et al., 2013, Eur J Hum Genet.; doi: 10.1038/ejhg.2013.254). **Results**: In 7.4% (77/1047) of fetuses with US anomalies pathogenic array findings were detected, a microscopically detectable abnormality in 1.8% and a submicroscopic aberration in 5.6%. In 2.9% (40/1408) of cases without US abnormalities, pathogenic chromosome aberrations were found, a microscopically detectable abnormality in 0.9% and a submicroscopic aberration in 2%. The prevalence of submicroscopic UD in both groups of fetuses was ~0.5% and mainly involved early-onset untreatable diseases. The prevalence of SL for neurodevelopmental disorders was twice as high in fetuses with US (2.6%) as compared to phenotypically normal fetuses (1.3%). **Conclusion**: SNP array testing is an asset to prenatal cytogenetic diagnosis, as is demonstrated by the detection of submicroscopic pathogenic chromosome abnormalities in cases with and without US anomalies. Apart from SL, these submicroscopic aberrations account for an extra 3% (abnormal fetuses) and 0.7% (normal fetuses) of clinically relevant aberrations. Therefore, array should be a first-tier prenatal cytogenetic test for all indications.

## 312

**The clinical utility of molecular genetic testing strategies for the diagnosis of mitochondrial disorders.** *R. Bai, D. Arjona, J. Higgs, J. Scuffins, J. Juusola, P. Vitazka, J. Neidich, K. Retterer, K. Parsons, N. Smaoui, E. Haverfield, S. Suchy, G. Richard.* GeneDx Inc, Gaithersburg, MD.

Mitochondrial disorders (MtD) are genetically and clinically heterogeneous, making the choice of genetic tests for patients with suspected MtD a challenge. This study compared clinical utility of different genetic tests for MtD: whole mitochondrial genome analysis for mtDNA mutations (WMG), a nuclear NGS panel of 140 genes common in MtD (CompNuc), and whole exome sequencing (WES) to screen most nuclear genes. Patients undergoing these tests were classified into clinical groups according to MtD diagnostic criteria (NEUROLOGY 2006;67:1823). DNA samples from 635 unrelated patients were tested by WMG (121 definite, 173 probable, 285 possible, 56 unclassified); 669 by CompNuc (102 definite, 106 probable, 403 possible, 58 unclassified), and 200 by WES (26 definite, 37 probable, and 137 possible; 152 cases included both parental samples or trios). Of 635 WMG cases, 42 (6.6%) were diagnostic for definitive mutations and 29 (4.7%) for variants segregating with disease in the family, with a yield of 11.2% (71/635). The positive rate by group for definite/probable, possible, and unclassified MtD patients were 20.3% (60/294), 3.2% (9/285) and 3.6% (2/56), respectively. Of 669 CompNuc cases, diagnostic results were identified in 119 (17.8%), with a positive rate of 39.9% (83/208) for definite/probable, 7.7% (31/403) for possible, and 8.6% (5/58) for unclassified MtD patients. Of 200 patients tested by WES, 61 had a mutation identified, with a positive rate of ~30% for definite/probable MtD (19/63) as well as 30% for possible MtD (42/137). Trios yielded 49 positive results (32.2%), more successful than singleton testing (12; 25%). In 20 (32.8%) positive WES cases, mutations were found in a CompNuc gene, the remainder in genes causing a disorder with clinical overlap with a MtD. 52 patients negative for CompNuc had additional reflex testing to WES. Diagnostic mutations were identified in 15 (29%) of these, all in non-MtD genes. The technical reliability of CompNuc was higher than WES. WES missed 7% of the reportable variants identified by CompNuc, all in pseudogene/low coverage regions. For patients highly suspected of a mitochondrial disorder, combined mitochondrial genome sequencing and a comprehensive MtD nuclear gene panel may detect mutations in ~ 60% of patients, the most successful strategy. For patients without strong evidence of MtD, concurrent WES with mitochondrial genome sequencing is preferable, identifying an underlying cause in ~1/3 of cases.

## 313

**Pathogenic Variant Spectrum in Newly Described Cancer Genes on Next Generation Cancer Panels.** *L. Susswein, L. Vincent, R. Klein, J. Booker, M.L. Cremona, P. Murphy, K. Hruska.* GeneDx, Gaithersburg, MD.

Background: The past few years have witnessed a rapid expansion of the number of genes available on clinical cancer genetic tests. Twelve of the genes on Next Generation Sequencing (NGS) panels were identified more recently as associated with cancer risk: *AXIN2, ATM, BARD1, BRIP1, CHEK2, CDK4, FANCC, NBN, PALB2, RAD51C, RAD51D*, and *XRCC2*.

Methods: Since the launch of the NGS Cancer panels at GeneDx in August 2013, 3363 individuals have been tested for panels that included some or all of these newly described genes. We retrospectively queried all results to assess the spectrum of pathogenic variants identified within this group of newly described cancer genes.

Results: Overall, 515 different germline variants were identified among these 12 genes. The gene with the highest number of variants identified was *ATM*, also the largest gene of the group, with 155 different variants. *PALB2* had the next highest degree of variation with 68 variants, followed by *CHEK2* with 59 and *BRIP1* with 58. Of the variants classified as pathogenic/likely pathogenic (P/LP), 79% were truncating (frameshift, nonsense, splice mutations, and large deletions). Four genes (*ATM, BRIP1, CHEK2*, and *PALB2*) were found to have >10 different variants classified as P/LP. While all P/LP variants in *PALB2* were truncating (n=16), the other genes had some missense variants that were classified as P/LP (*ATM*: 3 missense of 29 total P/LP variants, *BRIP1*: 1 of 10, *CHEK2*: 8 of 15). *AXIN2* and *CDK4* were the only two genes in which no P/LP variants were identified; all were missense and classified as variants of unknown significance (n= 13 and 3, respectively).

Conclusion: Continued utilization of multi-gene NGS panels will enable better characterization of these newer genes. Such knowledge will allow for better understanding of the clinical consequence of mutations in these genes.

## 314

**Exome Sequencing for the Diagnosis of 46,XY Disorders of Sex Development.** *E.C. Delot[1], R.M. Baxter[1], V.A. Arboleda[1], H. Lee[2], H. Barseghyan[1], M.P. Adam[3], P.Y. Fechner[4], R. Bargman[5], K. Keegan[6], S. Travers[7], S. Schelley[8], L. Hudgins[8], R.P. Mathew[9], H.J. Stalker[10], R. Zori[10], O.K. Gordon[11], L. Ramos-Platt[12], A. Eskin[1], S.F. Nelson[1,2], E. Vilain[1].* 1) Dept Human Genetics, 5301A Gonda Bldg, David Geffen Sch Med at UCLA, Los Angeles, CA 90095; 2) Pathology and Laboratory Medicine, David Geffen Sch Med at UCLA, Los Angeles, CA; 3) Department of Pediatrics, U. Washington, Seattle WA 98195; 4) Department of Endocrinology, Seattle Children's Hospital, Seattle WA 98105; 5) Nassau University Medical Center, East Meadow, NY 11554; 6) Depts of Pediatrics and Human Genetics, Ann Arbor, MI 48109; 7) The Children's Hospital Colorado, Aurora, CO 80045; 8) Division of Medical Genetics, Stanford University, Lucile Packard Children's Hospital, Stanford, CA 94305; 9) TriStar Children's Specialists, Nashville, TN 37203; 10) Division of Pediatric Genetics & Metabolism, U. Florida Gainesville, FL 32610; 11) Cedars-Sinai Medical Center, Los Angeles, CA 90048; 12) Children's Hospital of Los Angeles, Los Angeles, CA 90027.

Disorders of sex development (DSD) are clinical conditions where there is a discrepancy between the chromosomal sex and the phenotypic (gonadal or genital) sex of an individual. Such conditions can be stressful for patients and their families and have historically been difficult to diagnose, especially at the genetic level. In particular, for cases of 46,XY gonadal dysgenesis, once variants in *SRY* and *NR5A1* have been ruled out, there are few other single gene tests available. We used exome sequencing followed by analysis with a list of all known human DSD-associated genes to investigate the underlying genetic etiology of 46,XY DSD patients who had not previously received a genetic diagnosis. We were able to identify a likely genetic diagnosis in more than a third of cases, including 22.5% with a pathogenic finding and an additional 12.5% with likely pathogenic findings. In addition, 15% had variants of uncertain clinical significance (VUS) that may be reclassified as literature evolves. Exome sequencing allowed a remarkable level of genetic diagnostic success in this cohort, especially considering that, for most patients, all other endocrine and genetic testing had been exhausted. Early identification of the genetic cause of a DSD will in many cases streamline and direct the clinical management of the patient, with more focused endocrine and imaging studies and better informed surgical decisions. When unaffected parents are also genotyped there is the additional possibility of identifying novel genes that will further enhance our understanding of these complex conditions and allow for better care and prognostic information for the patients and their families.

## 315

**Clinical exomes for hearing loss: surprising diagnoses and better yields.** *L.H. Hoefsloot[1], I. Feenstra[2], I.J. de Wijs[2], M.H. Siers[2], H.P.M. Kunst[3], R.J. Admiraal[3], R.J.E. Pennings[3], H. Scheffer[2], H. Kremer[2,3], H.G. Yntema[2].* 1) Department of Clinical Genetics, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA, Rotterdam, The Netherlands; 2) Department of Human Genetics, Radboud university medical center, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands; 3) Department of Otorhinolaryngology, Head and Neck Surgery, Radboud university medical center, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands.

Clinical exome sequencing is a test that can be used for causative mutation detection, but also for the discovery of novel gene - disease associations. Because of its high heterogeneity, non-syndromic hearing loss is an excellent disorder for exome sequencing in a diagnostic setting. Hearing loss affects one in every 1000 newborns, with approximately half of cases with a genetic background. In 30% of these, additional features lead to the diagnosis of syndromic hearing loss, but in 70% hearing loss is the only finding. We performed clinical exome sequencing in a group of 200 probands with hearing loss. In a two-tier analysis, variants in 120 genes known to be associated with hearing loss were analysed first, followed by analysis of the 'full' exome data set in case no causative mutations were identified. Analysis of variants in the panel of deafness-associated disease genes, allowed us to establish a diagnosis in 15% of cases. In another 30% of cases further studies were needed. These studies include segregation analysis for variants of unknown pathogenicity, analysis of the entire coding region and/or MLPA analysis to detect a second mutation in recessive genes, and reverse phenotyping for cases with likely-causative mutations in genes associated with syndromic hearing loss. The second tier, analyzing the exome in patients without a diagnosis in the first tier with the use of a high stringent filter (truncating, and missense variants with a PhyloP score >3.5) is ongoing. Preliminary results are for instance a homozygous mutation in the SGSH gene that was detected in a child with hearing loss and mild intellectual disability. Reverse phenotyping confirmed the diagnosis of mucopolysaccharidosis type IIIA. Furthermore, we found truncating mutations in genes that are likely candidates for deafness (based on expression pattern or homology to known deafness genes) in 3 patients. Although further studies are needed, our first results indicate that another 25% might be solved by analysis of the entire exome. In conclusion, exome sequencing with bioinformatic filtering for genes associated with hearing loss led to a molecular diagnosis in 15-45% of the patients. Exome wide analysis has revealed new candidate genes in selected cases and might add another 25% to the final diagnostic yield. More importantly, it has provided patients with a molecular diagnosis that was not anticipated, and much earlier than classic diagnostics would have revealed.

## 316

**De novo Mutations Identified in Clinical Whole Exome Sequencing.** *S. Pan[1], F. Xia[1], D. Muzny[1], S. Plon[1, 2], J. Lupski[1], A. Beaudet[1], R. Gibbs[1], C. Eng[1], Y. Yang[1].* 1) Molecular and Human Genetics, Baylor College of Medicine, HOUSTON, TX, USA; 2) Pediatrics-Oncology, Baylor College of Medicine, HOUSTON, TX, USA.

It has long been recognized that de novo pathogenic mutations contribute to genetic diseases especially early onset neurodevelopmental diseases. The occurrence of those de novo changes results in the relatively constant prevalence of related genetic disorders in human populations irrespective of severely reduced fitness. However, the exact prevalence and spectrum of de novo mutations in these diseases remain unknown. We have accumulated a large patient cohort of 2000 unrelated patients referred for clinical WES, of which 87% had early onset neurologic disorders. Through whole exome sequencing (WES) followed by family-based Sanger confirmation, molecular diagnoses were made in 504 patients, including 23 with two diagnoses. The WES diagnoses included 280 autosomal dominant (AD), 181 autosomal recessive (AR), 65 X-linked (XL) and 1 mitochondrial disorders. The total known de novo events (249) accounts for ~13% (249/2000) of all clinical WES cases in our cohort and 56% (249/504) of all WES cases with molecular diagnoses. About 74% (208/280) of the AD disorders (87% if excluding cases without parental studies), 62% (40/65) XL disorders as well as the one case with pathogenic change in the mitochondrial genome resulted from de novo mutations in the causal genes. The de novo changes included 127 missense, 42 nonsense, 11 splicing, 64 frameshift and 4 inframe changes. Of the 177 point mutations, 106 (59.9%) were C>T or G>A changes occurred at the CpG sites. Notably, mosaicism of mutant alleles was identified in 5 probands (3 with AD and 2 with XL disorders). The 249 total de novo changes were distributed in 122 genes, among which *ARID1B* (14X), *ANKRD11* (7X) and *KCNT1* (7X) were the most frequently affected. Recurrent point mutations in mutation hotspots were observed in *ACTA2* (2X), *ANKRD11* (2X), *KCNT1* (2X) and *PACS1* (3X) genes. In summary, we have identified a spectrum of de novo mutations in our large clinical WES cohort. Although individually rare, this group of de novo mutations contributed to the molecular diagnoses for a significant portion of clinical WES cases. Together with associated clinical phenotypes, this collection of de novo mutations can provide more insight in our understanding of the etiology of sporadic genetic diseases.

## 317

**Frequency of "ACMG-56" Variants in Whole Genomes of Healthy Elderly.** *L. Ariniello[1], C.S. Bloss[1], G. Erickson[1], P. Pham[2], D. Boeldt[1], O. Libiger[1], N. Schork[1], E. Topol[1], A. Van Zeeland[2], A. Torkamani[1].* 1) Scripps Translational Science Institute, La Jolla, CA; 2) Cypher Genomics, San Diego, CA.

Last year the American College of Medical Genetics (ACMG) recommended that laboratories performing clinical sequencing seek and report mutations of known or expected pathogenic mutations across a list of 56 genes implicated predominately in cancer and cardiovascular disease risk. It was recommended that this be performed for all clinical germline exome and genome sequencing, irrespective of patient age. Unfortunately, there are few studies on the collective frequency of mutations across this list of 56 genes in any given population. Such information could enable laboratories to anticipate their potential reporting burden. In the current study, we aimed to determine the frequency of mutations across the "ACMG-56" in the Scripps Wellderly Cohort. The Wellderly cohort is comprised of individuals age 80 or older without a history of any major chronic diseases or medication use. Whole genome sequencing data generated by Complete Genomics for a subset of N=454 (mean age=86.9, range 80-104 years) unrelated European individuals were analyzed. Variants were extracted per ACMG-56 guidelines using the Cypher Genomics annotation pipeline, which returns both reported pathogenic variants as well as predicted pathogenic variants based on allele frequency in various reference populations and algorithmic predictions of functional impact. Across the sample, 609 non-unique variants were identified. A total of 314 individuals (69.2%) were heterozygous for at least one variant, and a range of 0 to 6 variants per individual were identified (mean= 1.34, SD=1.27). Six Wellderly individuals (1.3%) were homozygous for one variant; however, further analysis of these instances of homozygosity revealed questionable evidence of pathogenicity for the variants identified. There are few published studies of the frequency of ACMG-56 variants across a population. Dulik et al. (2013) reported a prior analysis of a different cohort, unselected for a healthy aging phenotype, in which a range of 1 to 6 (mean=3) variants per individual were identified. Frequencies in the Scripps Wellderly cohort were lower, which would be expected given the selection criteria for this cohort. The ACMG-56 recommendations have been highly controversial for a number of reasons, and although many factors may influence laboratories' adoption of the guidelines, anticipating the frequency of variants likely to be identified, and thus the reporting burden, may help inform such decisions.

## 318

**Exploring the diagnostic yield of whole exome sequencing in a broad range of genetic conditions: the first 200 cases in the NCGENES study.** *N.T. Strande[1], C. Bizon[1,2], J.K. Booker[1], K.R. Crooks[1], A.K.M. Foreman[1], G.T. Haskell[1], M.A. Hayden[1], K. Lee[1], M. Lu[1], L. Milko[1], J.M. O'Daniel[1], P. Owen[1,2], B.C. Powell[1], C. Skrzynia[1], C.R. Tilley[1], A. Treece[1], D. Young[1,2], K.C. Wilhelmsen[1,2], K.E. Weck[1], J.S. Berg[1], J.P. Evans[1].* 1) University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Renaissance Computing Institute.

Massively parallel sequencing is an attractive modality for the diagnosis of monogenic disorders, but a major challenge with implementation in the clinical setting is determining which patients may be most amenable to diagnosis by this method. The NCGENES (North Carolina Clinical Genomic Evaluation by Next-generation Exome Sequencing) project is evaluating the use of whole exome sequencing as a diagnostic tool in patients with a variety of suspected genetic conditions where a clinical diagnosis was not made by traditional methods. We established "diagnostic gene lists" focused on certain clinical presentations in order to streamline our variant search. Variants in these genes were identified and annotated using an in-house variant analysis infrastructure and were manually reviewed to determine the pathogenicity of each variant. Variants with possible diagnostic significance (pathogenic, likely pathogenic, or variants of uncertain significance (VUS) in genes for which the patient's phenotype was consistent with the disorder) were confirmed by Sanger sequencing and returned to the patients. Defining pathogenicity of individual variants is important, but it is also necessary to consider the phenotype to assess the overall diagnostic yield of such testing. We define "positive" results as a variant or variants that provide a definitive or probable explanation for the patient's phenotype. We define several scenarios comprising the category of "uncertain" results: VUS in which there exists only uncertainty regarding whether the variant is deleterious, variants with possible contribution to disease but not entirely consistent with the phenotype, and heterozygous variants for recessive disorders consistent with the phenotype in which a second mutation was not identified. Overall, case-level results were positive (36/200) or uncertain (45/200) in 42.5% of patients. Diagnostic yield was strongly influenced by the diagnostic category. For example, among 49 patients with suspected hereditary cancer susceptibility, there were 4 (8.2%) positive, 10 (20.4%) uncertain, and 35 (71.4%) negative. In contrast, among 54 patients with primary neurological disorders, there were 13 (24.1%) positive, 14 (25.9%) uncertain, and 27 (50%) negative. Our results suggest that diagnostic yield of WES is not equal for all genetic conditions, and we discuss some reasons why this might be the case. This study aids in defining which patients ultimately may benefit most from WES.

## 319

**Clinician CME Tailored to Individual Patient's Whole Exome Results.** *M.A. Giovanni, M.F. Murray.* Genomic Medicine Institute, Geisinger Health System, Danville, PA.

Background: The Geisinger Health System "Genome Sequencing Study" (GSS) was launched in January 2014 and will generate whole exome sequence (WES) data on 100,000 unselected patient volunteers over a five-year period. The WES data is being generated in a research setting, and pathogenic or likely pathogenic variants are being identified in 75 "clinically actionable" genes. Variants will undergo CLIA validation by Sanger sequencing for clinical return of results to both patient and healthcare provider as well as documentation in the patient's electronic health record. The clinician support infrastructure includes optional continuing medical education (CME) modules targeted specifically to their individual patient's test result. Study Population: 100,000 Geisinger Health System patients will be consented and provide a sample in conjunction with a normal clinical blood draw. Geisinger clinicians throughout the healthcare system will be engaged in genomic results delivery due to their patient's participation in the GSS. Methodology: A list of 75 genes for which a pathogenic or likely pathogenic variant would be clinically actionable was developed and vetted as "the Geisinger 75". The Geisinger 75 represents 26 monogenic diseases with autosomal dominant, autosomal recessive, and X-linked inheritance patterns. Twenty-six 30-minute free-standing CME modules were developed to guide the non-geneticist clinician through the clinical evaluation of their patient in light of the genomic findings. The modules are electronically hosted on the health system's website and include pre- and post-module evaluation. Results: These 0.5 hour educational modules respond to the workflow constraints of busy clinicians by offering focused, electronic CME opportunities at the point of care. Further details on the participation and performance of clinicians following the return of results are being tracked. Conclusions: As genomic data becomes increasingly available to inform individual patient care, non-expert clinicians will most likely seek opportunities to augment their knowledge and understanding of applied genomics prior to delivering results to their patients. Individual CME modules targeted to patient-specific test results will improve clinician preparedness for delivering genomic results to their patients.

## 320

**Changing the Landscape of Genomics Education Through a Massive Open Online Course (MOOC): Genomic Medicine Gets Personal.** *B.R. Haddad, J. Russel, S. Pennestri, D. Demaree, M. Tan, B.N. Peshkin.* Georgetown University, Washington, DC.

Advances in genomics have led to a major paradigm shift in medical practice. While medicine has always been "personal," the availability of genomic data has made it possible to individualize care for many patients. It is estimated that the cost of whole genome sequencing may fall to $1,000 and such testing may become routine. Whether physicians are asked to interpret genomic information they have requested as part of patient care, or their patients have "self-prescribed" from a direct-to-consumer personal genomics company, they need to be trained to accomplish this task. However, many studies over the last decade have demonstrated that health care providers have suboptimal knowledge about genomics. Consumers also need to be educated about this field. Toward this end, we developed a cross-disciplinary MOOC that targets medical professionals as well as the general public. This 8-week course was launched on June 4, 2014 and covered 5 main themes: (1) Clinical genetics and genomics; (2) Laboratory techniques; (3) Consumer genomics; (4) Ethical, legal, and social issues; and (5) Present and future opportunities and challenges. Faculty from different disciplines including clinical and laboratory genetics, oncology, computational sciences, genetic counseling, bioethics, law, and business participated in the course. Content consisted of short video lectures, interviews, and roundtable discussions. Students were provided with online readings and resources, and were expected to complete pre- and post-course surveys, formative assessments, weekly quizzes, and a final exam. Participation in discussion boards and other exercises was also encouraged. Over 22,600 students from over 150 countries have enrolled. We will share our experience in developing the course and discuss the pros and cons of this approach to online education in genomics. Quantitative data will be presented about students' demographics, motivation, and learning achievements. Themes from the discussion board and other exercises will be discussed. We plan to maintain this course as a "living resource," updating it regularly, making it part of our medical school curriculum and continuing medical education program, and keeping it accessible to the public at large. We believe that this will allow us to provide an educational opportunity to a large audience worldwide, particularly to individuals with limited access to traditional educational resources in this cutting-edge field.

## 321

**Genome: Unlocking Life's Code - A Museum Exhibition as a Model for Informal Genomics Education.** *V. Bonham, C. Easter, E. Schonman, C. Daulton, R. Wise, B. Hurle, J. Witherly.* Education and Community Involvement Branch, NHGRI/NIH, Bethesda, MD.

Interactive, hands-on, and high-tech science museum exhibits and associated programs have the power to bring the sciences to life for large numbers of the public, to foster science literacy, and to help garner public support for scientific and health research. In 2011, the National Human Genome Research Institute and Smithsonian Institution's National Museum of Natural History partnered to create an innovative exhibit—"Genome: Unlocking Life's Code"—to give the public a window into genomic science. The exhibition, which opened in 2013, introduces the public to the role that genomics plays in their lives, in understanding health risks, disease, and treatment, our human origins, and how humans fit in as members of a family and a species in the natural world. Throughout the exhibition, the public is also asked to explore the ethical, legal, and social concerns raised by genomic science. As of June 2014, 2.5 million people have visited the exhibition while on view at the National Museum of Natural History in Washington, DC. To measure the impact of the exhibit on museumgoers, evaluations were conducted by the Smithsonian Institution's Office of Policy and Analysis in Summer 2013. These included a survey of 377 visitors entering the exhibit, a survey of 462 visitors exiting the exhibit, 33 in-depth interviews, and observations of 100 visitors as they walked through the exhibit. Quantitative and qualitative visitor responses were overwhelmingly positive, and qualitative results indicate a deepened appreciation and interest in genomics. Results also indicate that interactive aspects of the exhibit - with both volunteers and digital and mechanical interactives, such as computer touch screens, visual media, and quiz-like activities - led visitors to strongly reflect on the implications of genomic science. These evaluative results shed light on the role that informal science education can play in increasing genomic literacy, knowledge, and interest within the public.

## 322

**Human Genetics: Medical and Societal Implications. A high school course taught on a medical school campus.** *M. Godfrey[1], J.E. Bird[2].* 1) Munroe-Meyer Institute, Univ Nebraska Med Ctr, Omaha, NE; 2) UNMC High School Alliance, Univ Nebraska Med Ctr, Omaha, NE.

Patterson and Carline (2006) articulated a vision for sustained and effective partnerships between health professionals and public schools. Their blueprint was modeled by the development of a novel University of Nebraska Medical Center (UNMC) and Omaha Metropolitan Area Public High Schools partnership: the UNMC High School Alliance. The Alliance includes 21 high schools from eleven school districts in the Omaha, Nebraska area; including Council Bluffs, Iowa. Over the first four years of the Alliance program some 29% of students have come from underrepresented minority groups; compared to an 18% minority population in the Omaha area. Alliance students attend their home high schools in the morning and come to UNMC each afternoon for an entire academic year. While at UNMC they are taught by Medical Center faculty. Courses have included: Biomedical Research, Infectious Disease, The Art & Science of Decision Making, Fundamentals of Disease, Community Health, Human Anatomy, Human Genetics, and Health Science Fundamentals. Students also take part in numerous health science shadowing opportunities both on and off the UNMC campus. To ensure that high school credit is received, a certified high school teacher assists at all phases of the program. The Alliance program was not designed to attract only the best and brightest. Requirements are modest and student (as opposed to parent) interest is paramount. Here we report on the early evaluation and success of the Human Genetics: Medical and Societal Implications, a course taught by medical school faculty to high school students. Student knowledge of genetics increased each year the course has been taught. Interest in a career in genetics also increased. Students rarely thought about ELSI related issues prior to the course. Hands-on modules that related molecular biology to genetics are engaging and among the most liked lessons. Mathematical aspects of genetics are among the most difficult concepts for students. Nevertheless, many students expressed that they now had the confidence and new knowledge that would enable them to pursue a career in the health sciences.

## 323

**Whole Genome Sequencing in a Healthy Population: Processes, Challenges, and Insights.** *CS. Richards[1], P. Jain[1], MO. Dorschner[2,3], DA. Nickerson[3], GP. Jarvik[3,4], LM. Amendola[4], DK. Simpson[7], A. Rope[7], J. Reiss[7,8], K. Kennedy[8], DI. Quigley[9], J. Berg[10], C. Harding[1], M. Gilmore[7], P. Himes[7], B. Wilfond[5,6], KAB. Goddard[11] on behalf of the NextGen Project Team.* 1) Department of Molecular and Medical Genetics, Knight Diagnostic Laboratories, Oregon Health & Science University, Portland, OR; 2) Pathology, Division of Bioethics, University of Washington, Seattle, WA; 3) Genome Sciences, Division of Bioethics, University of Washington, Seattle, WA; 4) Department of Medicine, Division of Medical Genetics, Division of Bioethics, University of Washington, Seattle, WA; 5) Department of Pediatrics, Division of Bioethics, University of Washington, Seattle, WA; 6) Truman Katz Center for Pediatric Bioethics, Seattle Children's Hospital, Seattle, WA; 7) Department of Medical Genetics, Kaiser Permanente Northwest, Portland, OR; 8) Obstetrics and Gynecology, Kaiser Permanente Northwest, Portland, OR; 9) Laboratory, Kaiser Permanente Northwest, Portland, OR; 10) Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, NC; 11) Center for Health Research, Kaiser Permanente Northwest, Portland, OR.

We are performing carrier screening of a healthy population (reproductive age preconception) in a large integrated healthcare delivery system using whole genome sequencing (WGS). Our strategy is to test the female partner first, and if positive, offer testing to the partner. Our carrier screen includes about 500 genes. Our analytic pipeline includes: (1) WGS testing in the Illumina CLIA laboratory; (2) data processing including alignment and variant annotation using the SeattleSeq pipeline (Nickerson laboratory); (3) data analysis, including filtering variants and classification for the carrier test, Sanger confirmation of findings, and reporting in the Knight Diagnostic Laboratory (KDL); and (4) data analysis for incidental findings (Nickerson laboratory using the expanded incidental findings list of 114 genes described by Dorschner, 2013), and transfer of results to KDL for Sanger confirmation of findings and reporting of results. Presently, 11 patients are enrolled with 4 cases reported. To date 4 of the 5 variants reported have been novel, including two duplications leading to frameshifts (SACS gene associated with spastic ataxia, Charlevoix-Saguenay type; SLC3A1 gene associated with cystinuria), one splicing variant (CRTAP gene associated with osteogenesis imperfect type VII), and one nonsense variant (NAGA gene associated with Schindler disease types I and II). One known pathogenic missense variant was identified (AGXT gene associated with hyperoxaluria type 1). Most of these disorders are rare, and thus not present on most carrier screening panels. To date two patients carry two variants each, one patient carries only one variant, and one patient had no detectable reportable variants. Only pathogenic or likely pathogenic variants are reported, as our criteria for calling and classifying variants is very stringent. We will present metrics for the testing which include data filtering strategies for identifying disease-causing variants and time for analysis and classification of each variant. We have found that each patient's data presents a new learning experience. A Return of Results Committee composed of clinical and laboratory geneticists and genetic counselors, is consulted for expertise and guidance. We anticipate that at least 25 patient results, including incidental findings, will be returned prior to this presentation.

## 324

**Patient Perceptions about the Utility of Family History Review during Whole Genome Sequencing: Initial Findings from the MedSeq Study.** *K.D. Christensen[1], P.J. Lupo[2], J.O. Robinson[2], J. Blumenthal-Barby[2], J.L. Vassy[1,3], L.S. Lehmann[1], P.A. Ubel[4], J.S. Roberts[5], R.C. Green[6], A.L. McGuire[2], The MedSeq Study Team.* 1) Brigham & Women's Hospital and Harvard Medical School, Boston, MA; 2) Baylor College of Medicine, Houston, TX; 3) VA Boston Healthcare System, Boston, MA; 4) Fuqua School of Business and Sanford School of Public Policy, Duke University, Durham, NC; 5) University of Michigan School of Public Health, Ann Arbor, MI; 6) Partners HealthCare Center for Personalized Genetic Medicine, Boston, MA.

Background: Family history is critical to interpreting variants identified during genomic sequencing, particularly when disease and phenotypic indications are not evident. While physicians and clinicians are aware of its utility, patients may not recognize the importance of family history in the context of sequencing or as a standalone risk assessment tool.

Methods: The MedSeq Project is a randomized clinical trial exploring the use of whole genome sequencing in the care of healthy primary care patients and cardiology patients with dilated or hypertrophic cardiomyopathy. After completing a baseline questionnaire and blood draw, patient participants were randomized to a standard of care (SOC) arm where they reviewed their family history with their participating physicians, or to an experimental arm where they additionally reviewed findings from whole genome sequencing (WGS). Six weeks after this disclosure session, participants rated eight items assessing the utility of the family history review on 5-point scales. The WGS arm also rated eight parallel items assessing the utility of the sequencing information.

Results: Of 200 projected, 19 patients have completed the 6 week follow-up to date (42% WGS; 26% cardiology; mean age 55; 32% female; 100% white non-Hispanic). Within the WGS arm, utility ratings for family history and WGS did not differ (all $p>0.12$). However, participants in the WGS arm assigned higher ratings to family history than participants in the SOC arm about satisfying curiosity (4.3 vs 2.4, $p<0.001$), explaining a condition (3.9 vs 2.4, $p=0.003$), explaining a family history of disease (3.9 vs 2.8, $p=0.033$), tailoring treatments (3.7 vs 2.7, $p=0.014$), and preventing disease (3.8 vs 2.5, $p=0.016$).

Conclusions: Very early data showed that patients perceived greater utility to family health history when it was accompanied by WGS rather than as standalone information. Patients also rated family history as equally useful as personalized genetic information. Findings suggest that WGS may serve as a 'teachable moment' for addressing the benefits of understanding family history, perhaps because patients consider it an integral part of an overall genetic test experience. Ongoing work will examine disclosure sessions to understand how their content and focus be differ when family history review is coupled with WGS.

## 325

**Hereditary Cancer Communication with Underserved Patients.** *G. Joseph, C. Guerra.* Anthropology, University of California, San Francisco, San Francisco, CA.

While genetic counseling and testing for individuals and families at risk of hereditary breast and ovarian cancer (HBOC) is the standard of care, studies show that medically underserved patients (including people of color, low income and/or low literacy) have less knowledge of and access to cancer genetic counseling and testing, and even where these services are available, uptake is low compared with white, insured populations. Fewer than 13% of all women who receive BRCA testing in the US are of non-European ancestry even though people of color, who are disproportionately low-income, make up 35% of the US population. BRCA testing is now a required insurance benefit as a preventive service under the ACA; however such financial access is essential but not itself sufficient to ensure high quality genetic counseling and testing. Gaps in effective communication (when a message reaches the intended audience and where the meaning is mutually understood) are widely recognized as a major contributor to health disparities. We conducted an in-depth ethnographic study of HBOC counseling and testing at two public hospitals providing these services to their diverse underserved patients. Research included observations and audio recording of over 150 genetic counseling sessions conducted in English, Spanish and Cantonese, follow up interviews with patients and counselors, and focus groups with genetic counselors at NSGC. We identified structural barriers to effective communication, as well as strengths and limitations of counselors' communication across language, literacy level and culture. Structural barriers include substantial problems with medical interpretation for limited English speakers. The effectiveness of counselors' communication was limited by analogies and technical language that did not translate effectively across literacy level and culture. Counselors' strengths included their ability to communicate the essential purpose of the tests and to develop rapport with patients despite some miscommunication. As genetics and genomics become mainstream medicine, these advances can actually exacerbate breast cancer disparities if low-income women are unable to access and benefit from genetic risk services in the same ways as those who are affluent and insured. The unique characteristics of public health settings and the patients they serve must be understood and taken into account in order to reap the benefits of new genetic technologies.

## 326

**Cost effectiveness of adding genes to next generation sequencing panels for evaluation of colorectal cancer and polyposis syndromes.** *C.J. Gallego[1], B.S. Shirts[3], C.C. Pritchard[3], G.P. Jarvik[1], D.L. Veenstra[2].* 1) Division of Medical Genetics, University of Washington, Seattle, WA; 2) Pharmaceutical Outcomes and Research Policy Program, University of Washington, Seattle, WA; 3) Department of Laboratory Medicine, University of Washington, Seattle, WA.

Background: Next generation sequencing (NGS) panels are increasingly used in the evaluation of inherited colorectal cancer and polyposis (CRCP) syndromes. However, the added value incorporating additional genes into these panels has not been studied. This study evaluates the cost-benefit of adding groups of genes, based on their inheritance mode and penetrance. Methods: We developed a decision model comparing evaluation of CRCP syndromes with NCCN guidelines with four increasingly comprehensive NGS panels with the following groups of genes: (1) genes associated with Lynch syndrome alone, (2) adding genes associated with other autosomal dominant conditions with high colorectal cancer (CRC) penetrance, (3) adding genes associated with autosomal recessive conditions with high CRC penetrance, and (4) further adding autosomal dominant, low CRC penetrance genes. Empiric variant frequencies were obtained from a large clinical NGS laboratory. Outcomes were quality adjusted life years (QALY) and costs associated with intensive surveillance of colorectal cancer in relatives of probands identified with a CRCP syndrome. We calculated Incremental Cost-Effectiveness Ratios (ICER) of each panel compared to guidelines and compared to the next panel. Sensitivity analyses were conducted. Results: The use of a panel testing only for Lynch associated genes resulted in ICER between $125,000 and $260,000 per QALY, while the addition of high penetrance autosomal dominant genes associated with high CRC resulted in ICER of $38,000 to $69,000 per QALY. Adding high penetrance autosomal recessive genes associated with CRC had an ICER of $14,000 to $15,000 per QALY and the addition of low penetrance autosomal dominant genes yielded an ICER between $84,000 and $85,000 per QALY. Overall, the use of NGS panels for the evaluation of CRCP compared to current guidelines added costs as low as $40,000 per QALY. Sensitivity analysis did not markedly change these results. Conclusion: Incorporating autosomal dominant high penetrance genes associated with CRC to NGS panels is more cost effective than sequencing only Lynch syndrome genes. Addition of autosomal recessive high penetrance genes is also cost effective. Finally, adding low penetrance genes associated with CRC yields less benefit, but may still be cost-effective. These findings support the use of NGS panels for evaluation of CRCP syndromes and the addition of new genes to these panels, even when penetrance is modest.

## 327

**Whole-genome single-cell haplotyping, a generic method for preimplantation genetic diagnosis.** *M. Zamani Esteki[1], E. Dimitriadou[1], L. Mateiu[1], C. Melotte[1], N. Van der Aa[1], P. Kumar[1], R. Das[1], K. Theunis[1], J. Cheng[1,2], E. Legius[1], Y. Moreau[2], S. Debrock[3], T. D'Hooghe[3], P. Verdyck[4], M. De Rycke[4,5], K. Sermon[5], J.R. Vermeesch[1], T. Voet[1,6].* 1) Centre for Human Genetics, University Hospital Leuven, Department of Human Genetics, KU Leuven, Belgium; 2) Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, Belgium; 3) Leuven University Fertility Center, University Hospital Gasthuisberg, Herestraat 49, 3000 Leuven, Belgium; 4) Centre for Medical Genetics, Universitair Ziekenhuis Brussel, Laarbeeklaan 101, 1090 Brussels, Belgium; 5) Research group Reproduction and Genetics (REGE), Vrije Universiteit Brussel (VUB) Laarbeeklaan 101, 1090 Brussels, Belgium; 6) Single-cell Genomics Centre, Welcome Trust Sanger Institute, Hinxton, CB10 1SA, UK.

Preimplantation genetic diagnosis (PGD) is the genetic testing of embryos prior to implantation to avoid the transmission of germline genetic disorders or of unbalanced chromosomal rearrangements when a parent is a balanced carrier. Current single-cell PCR or FISH PGD-assays require family-specific designs and labor-intensive workup. Array comparative genomic hybridization (aCGH)-based methods, which are mainly applied for preimplantation genetic screening (PGS) to discern diploid from aneuploid embryos, enable genome-wide aneuploidy detection but do not allow diagnosing single gene disorders. Here, we present a generic method that detects in single blastomeres not only the presence of Mendelian disorders genome wide, but also chromosomal rearrangements and aneuploidies, including their parental origin as well as the meiotic or mitotic nature of chromosomal trisomies. The method interrogates single nucleotide polymorphisms (SNPs) and uses a novel computational pipeline for single-cell genome-wide haplotyping and imputation of linked disease variants (siCHILD). Following stringent single-cell QC-metrics, a bimodal approach, based on discrete SNP-calls and continuous SNP B-allele fractions, respectively, reconstructs the parental haplotypes of the biopsied single cell. The approach proved accurate on 55 embryos from 12 couples carrying either autosomal dominant, recessive or X-linked Mendelian disorders, or simple or complex translocations. The method allowed diagnosing an embryo for multiple monogenic disorders at once, and, in contrast to current PGD for translocation cases, it enabled distinguishing embryos that inherited normal chromosomes from embryos that inherited a balanced configuration of the rearranged derivative chromosomes. The method facilitates genetic selection of embryos, and broadens the range of classic PGD.

## 328

**TARGETED LOCUS AMPLIFICATION FOR HYPOTHESIS NEUTRAL NEXT GENERATION SEQUENCING AND HAPLOTYPING OF SELECTED GENOMIC LOCI.** *M.J. van Min[1], P.J.P. de Vree[2], W. de Laat[1,2], E. de Wit[1,2], M. Yilmaz[1], M. van de Heijning[1], P. Klous[1], P. ter Brugge[11], J. Jonkers[11], J. Foekens[12], J. Martens[12], H.K. Ploos van Amstel[13], P.P. Eijk[3], D. Sie[3], B. Ylstra[3], M. Ligtenberg[10], M.F. van Dooren[14], L.J.C.M. van Zutven[14], E. Splinter[1], S. Verbeek[4], K. Willems van Dijk[4], M. Cornelissen[5], A.T. Das[5], B. Berkhout[5], B. Sikkema Radatz[6], E. van den Berg[6], P. van der Vlies[6], Y. Wan[2], J.T. den Dunnen[7], M. Lamkanfi[8,9], Hubrecht Institute.* 1) Cergentis B.V., Padualaan 8, 3584 CH Utrecht, The Netherlands; 2) Hubrecht Institute-KNAW & University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands; 3) Department of Pathology, VU University Medical Center, PO Box 7057, NL-1007 MB Amsterdam, The Netherlands; 4) Department of Human Genetics and Department of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands; 5) Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; 6) Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen; 7) Leiden Genome Technology Center, Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands; 8) Department of Medical Protein Research, VIB, Albert Baertsoenkaai 3, 9000 Gent, Belgium; 9) Department of Biochemistry, Ghent University, Albert Baertsoenkaai 3, 9000 Gent, Belgium; 10) Department of Human Genetics, Radboud university medical center, Nijmegen, The Netherlands; 11) Division of Molecular Pathology and Cancer Genomics Center, The Netherlands Cancer Institute Amsterdam, The Netherlands; 12) Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands; 13) Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands; 14) Department of Clinical Genetics, Erasmus Medical Center, PO box 2040, 3000 CA Rotterdam, The Netherlands.

Current methodologies in genetic diagnostics and research are limited in their ability to uncover all genetic variation in genes of interest. Clinical genetic tests often only focus on exons and therefore miss variants in the non-coding regulatory sequences of genes. In addition, structural variants, i.e. deletions/duplications (CNVs), translocations, insertions and inversions, are difficult to uncover. Their robust detection is hampered by the hypothesis-driven nature of current targeted sequencing methodologies: the collection of probes (in hybridization-based capture methods) or primers (in polymerase or ligase-based re-sequencing approaches) determines the sequences that will be analyzed. None of these methods provide haplotyping information, ultimately needed to get complete sequence information. Here we present targeted locus amplification (TLA), a strategy to selectively amplify and sequence entire genes. TLA is based on crosslinking, fragmenting and religation steps such as performed in chromatin capture technologies. We developed a strategy that aims to ensure that (1) no sequences are lost, (2) multiple captures (i.e. maximum sequence information) are contained in the anchor-containing DNA molecules and therefore enables haplotyping , and (3) that after re-ligation, circles are formed of a size (on average 2 kb) that can still easily be PCR amplified. We show that, unlike other targeted re-sequencing methods, TLA works without detailed prior locus information as one or a few TLA primer pairs are sufficient to amplify and sequence tens to hundreds of kilobases of surrounding sequences. This, we demonstrate, enables robust detection of single nucleotide variants, structural variants and gene fusions in clinically relevant genes. TLA can characterize integrated viruses, transgenes and their integration sites. Finally, TLA can be used to haplotype across large chromosomal intervals. Data will be presented showing the ability of the TLA Technology to: -Sequence a.o. the complete BRCA1 & BRCA2 genes for the detection of germline and somatic variation in (xenograft) tumor samples. -Detect all gene-fusions in genes of interest and characterize fusions at breakpoint resolution in leukemia and other cancer types. -Sequence the HIV virus genome and its integration sites. -Haplotype the Human Leukocyte Antigen (HLA) region and other genes of interest.

## 329

**Saturation Genome Editing by Multiplex Homologous Donor Repair.** *G.M. Findlay, E.A. Boyle, R.J. Hause, J. Klein, J. Shendure.* Department of Genome Sciences, University of Washington, Seattle, WA.

Saturation mutagenesis, coupled to an appropriate biological assay, represents a fundamental means of achieving a high-resolution understanding of regulatory and protein-coding nucleic acid sequences of interest. However, mutagenized sequences interrogated by episomal expression or random genomic integration fail to account for the native context of the element's chromosomal locus with respect to the surrounding sequence, epigenetic milieu, and endogenous transcriptional activity. This shortcoming markedly limits the interpretability of the resulting measurements of mutational impact. Here, we couple CRISPR/Cas9 RNA-guided cleavage with multiplex homology-directed repair to introduce hundreds to thousands of programmed variants within a single population of cells, a method we call "saturation genome editing". Furthermore, we show that despite the relatively low fraction of edited cells, massively parallel functional analysis of such populations is possible by performing selective amplification of RNA or DNA from edited genomes followed by high-throughput sequencing. As a proof-of-concept, in exon 18 of *BRCA1*, we replace a six base-pair (bp) genomic region with all possible hexamers, or the full exon with all possible single nucleotide substitutions, and measure strong and reproducible effects on transcript abundance attributable to nonsense-mediated decay and exonic splicing elements. We next perform saturation genome editing in haploid human cells, targeting a well-conserved 75 bp coding region of an essential gene, *DBR1*. Measurement of mutations' relative effects on growth span over three orders of magnitude for missense SNVs, correlate with predictive models of functional impact, and reveal only few conservative amino acid substitutions are well-tolerated in the enzyme's active site. These proof-of-concept experiments show that saturation genome editing enables the multiplex functional analysis of mutations within the context of the genome itself. This approach facilitates the precise measurement of the consequences of large numbers of genomic mutations and may potentially aid in the interpretation of variants of uncertain significance observed in clinical sequencing.

## 330

**Metagenomic deconvolution and species discovery in microbiomes using contact probability maps.** *J.N. Burton, I. Liachko, M.J. Dunham, J. Shendure.* Department of Genome Sciences, University of Washington, Seattle, WA.

Every surface of the human body and digestive tract hosts a microbial community, i.e., a microbiome, consisting of bacterial, archaeal, and eukaryotic species in varying abundances. It has become increasingly clear that the human microbiomes play crucial roles in human health and disease. However, most microbial species cannot be cultured independently of their native communities, and thus are difficult to study even if they are abundant. Microbiomes are often analyzed through metagenomic shotgun sequencing. With this approach, the library construction step removes long-range contiguity information; thus shotgun sequencing and de novo assembly of a metagenomic sample typically yields a collection of contigs that derive from many different species in the sample and cannot readily be grouped by species. To enable whole-genome assembly from metagenomic samples, we have adapted the Hi-C technique, which creates a map of interactions between spatially adjacent DNA regions. When applied to a mixed cell population, Hi-C produces paired-end reads which connect intracellular sequences. Our method, which we call "MetaPhase", exploits this signal to reconstruct the individual genomes of microbial species in a mixed sample, including uncultivable and previously unknown species. MetaPhase can also use the Hi-C signal to create scaffolded genome assemblies of individual eukaryotic species within the microbial community and model the 3-dimensional architecture of these newly assembled genomes. We have applied MetaPhase to several synthetic metagenome samples containing up to 18 species of bacteria, archaea, and fungi, successfully clustering their genome content with over 99% agreement with published reference genomes. We are also applying it to samples of vaginal microbiomes from individuals with bacterial vaginosis (BV), a disease characterized by an increase in the species diversity of the vaginal microbiome. Lastly, we are applying MetaPhase to the microbial communities living inside sputum from patients with cystic fibrosis, with the goal of understanding the species responsible for opportunistic lung infections in these patients.

## 331

**Deep whole-genome sequencing based analysis of mosaic transposable mobile element insertions in adult human tissue.** *X. Zhu[1], A. Fiston-Lavier[2], D. Petrov[3], M. Snyder[4], D. Levinson[1], A. Urban[1,4].* 1) Psychiatry, Stanford University, Palo Alto, CA; 2) Computer Science, University of Montpellier2, Montpellier, Hérault, France; 3) Biology, Stanford University, Palo Alto, CA; 4) Genetics, Stanford University, Palo Alto, CA.

Mobile elements (MEs) comprise a large portion of the human genome and some retain their capacity for transposition. Somatic retrotranspositions have been associated with several nervous system diseases, suggesting they may have a significant impact on the developing brain. To analyze somatic ME variation in human brain, previous studies have used target capture- or single cell-sequencing methods. Results suggest that somatic cell ME mutation is very likely, but it is not known how frequently new mobile element insertions (MEIs) occur and which proportions of cells, cell types or brain regions are affected. We developed an unbiased deep whole-genome sequencing-based approach to detect and quantify somatic MEIs without using target capture or whole-genome amplification methods for discovery. Using paired-end sequencing we sequenced DNA from postmortem brain from a subject with schizophrenia and a control subject, at mean 80x sequencing depth for superior temporal gyrus and 60x depth for both cerebellum and liver. We extensively modified an existing ME calling algorithm to identify novel MEIs. To validate predicted MEIs we used custom oligonucleotide capture for targeted re-sequencing of predicted novel MEIs at >300x depth. We identified 796 high-support potential novel MEIs each of which is present in only one of the four brain regions tested. The targeted-capture sequencing for validation tiled oligonucleotides across the regions surrounding ~20,000 of the whole-genome based MEI calls. Initial analysis showed high mappability of the resulting captured sequence and a substantial validation rate. At a false discovery rate below 10%, 82% of high-support Alu insertions and 73% of high-support LINE-1 insertions were validated. These validation rates will be further evaluated with digital droplet PCR, also to quantify the degree of mosaicism for each novel MEI. This project uses whole-genome sequencing to address more comprehensively the extent and precise genomic localization of somatic retrotransposition of mobile elements in human brain cells. This phenomenon represents an additional mechanism by which genomic variation can influence brain development and disease. Methods for analysis of novel ME sequences in deep whole-genome sequencing will be needed to fully characterize this type of mutation in diverse tissues and in subpopulations of cells using cell sorting techniques. This study is supported by grant R01MH094740 (to D.F.L.) from the NIMH.

## 332

**Increased complexity of the human genome revealed by single-molecule sequencing.** *M.J.P. Chaisson[1], J. Huddleston[1], P.H. Sudmant[1], M. Malig[1], F. Hormozdiari[1], U. Surti[3], R. Wilson[4], M. Hunkapiller[2], J. Korlach[2], E.E. Eichler[1].* 1) Genome Sciences, University of Washington, Seattle, WA; 2) Pacific Biosciences, Menlo Park, CA; 3) Department of Pathology, University of Pittsburgh, Pittsburgh, PA; 4) Washington University School of Medicine, St. Louis, MO.

The human genome is arguably the highest quality mammalian reference assembly yet more than 150 interstitial gaps remain and aspects of its structural variation remain poorly understood 10 years after its completion. We generated and analyzed 40-fold sequence coverage of a haploid human genome (CHM1) using Pacific Biosciences' single molecule, real-time (SMRT) sequencing (average mapped read length 5.8 kbp). We developed methods to detect indels and structural variants from several bases up to 20 kbp. We closed or extended 55% of the remaining interstitial gaps in the human GRCh37 reference genome and found that 78% of closed gaps carry long polypyrimidine/purine tracts multiple kilobases in length. Comparing the single haplotype to the human reference, we resolved 34,000 indels and structural variants at the basepair level with 99.9% sequence accuracy. We find a 3:1 excess of simple tandem repeat (STR) insertion over deletion of which 393 STR and variable number tandem repeat insertions are greater than 1 kbp. We find that 51% of such sequences vary at least twofold in copy number, representing sites of potential genetic instability. Of the STR insertions, 1,566 correspond to likely deficiencies in the reference sequence. In addition, the analysis uncovers other categories of complex variation that have been difficult to assess, including mobile element insertions (e.g., SVA) as well as inversions mapping within more complex and GC-rich regions of the genome. Our results suggest a systematic bias against assembly of longer and more complex repetitive DNA that can now be partially resolved with the application of new sequencing technologies.

## 333

**In situ genome-wide expression profiling of individual cell types.** *C. Kodira[1], A. Sood[1], L. Newberg[1], A. Miller[2], F. Ginty[1], E. McDonough[1], Y. Sui[1], A. Bordwell[1], Q. Li[1], S. Kaanumalle[1], K. Desai[1], Z. Pang[1], E. Brogi[2], S. Larson[2], I. Mellinghoff[2].* 1) Diagnostics, Imaging & Biomedical Technologies, GE Global Research Center, Niskayuna, NY 12309, USA; 2) Memorial Sloan Kettering Cancer Center, 1275 York Ave, New York, NY 10065, USA.

Understanding the tumor heterogeneity via probing of the key molecular alterations at the DNA, RNA and protein levels is crucial for tumor classification and linking their molecular status to appropriate therapy options. While Next Generation Sequencing (NGS) technology is extremely powerful for global characterization of tumour heterogeneity in terms of large-scale mutational and mRNA expression profiles, it is still limited in providing spatial/cellular context. Traditional immunohistochemistry (IHC) on the other hand is extremely valuable in providing cell level characterization and spatial context but is limited by the amount of information it can provide about altered genes/pathways. To address this challenge, we describe here a novel approach in extracting genome-wide expression profiles of individual cell types within a tumor while retaining information about their spatial arrangement in the tissue. This approach combines the hyperplex capabilities of a new immunofluorescence platform called MultiOmyxTM with microarray gene expression and Next Generation Sequencing (NGS) data and novel methods to derive cell-specific expression profiles. Profiles of individual cell types with similar molecular alterations can then be correlated to clinical information to predict specific outcomes. We present here preliminary results from a proof of concept breast cancer study to illustrate the utility of such a multi-omics approach for better understanding of tumor heterogeneity. We believe our approach is a powerful research tool for enabling biomarker discovery, therapy matching and potentially cancer management.

## 334

**Whole-genome sequencing characterizes multiple mutational mechanisms resulting from off-target effects of CRISPR-Cas9 and TALEN treatments in human embryonic stem cells.** *R.L. Collins[1], A. Veres[8,9,7], H. Brand[1,4], A. Ragavendran[1], S. Erdin[1], Q. Ding[8,9], B.S. Gosis[8,9], K. Musunuru[8,9,7,4,5], M.E. Talkowski[1,2,3,4,6].* 1) Center for Human Genetic Research, Massachussets General Hospital, Boston, MA; 2) Molecular Neurogenetics Unit, Massachussets General Hospital, Boston, MA; 3) Psychiatric and Neurodevelopmental Genetics Unit, Massachussets General Hospital, Boston, MA; 4) Broad Institute, Cambridge, MA; 5) Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, MA; 6) Department of Neurology, Harvard Medical School, Boston, MA; 7) Department of Medicine, Harvard Medical School, Boston, MA; 8) Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA; 9) Harvard Stem Cell Institute, Cambridge, MA.

Recent innovations in genome editing technologies have opened access to revolutionary new approaches in genetic research. The use of targeted nucleases, such as CRISPR-Cas9 and TALENs, have become widespread in human biology. The specificity of these tools to the targeted editing site, and the confounds presented off-target effects, have been widely debated, with most studies to assess off-target effects focusing on single nucleotide variations (SNVs) or indels at predicted sites based on sequence specificity. Here, we performed direct comparison of off-target effects from two genome-editing technologies (CRISPR-Cas9 and TALENs) in *HUES9* embryonic stem cell clones using unbiased whole-genome deep sequencing (~60X coverage). We sequenced a parental, untreated line, and nine treated clones. Each treatment group contained two clones with successful on-target mutations and one clone without targeted edits, serving as a treatment control. We selected three clones each with TALEN and CRISPR/Cas9 editing of the *SORT1* locus, as well as three CRISPR-Cas9 clones from an independent locus for comparison (*LINC00116*). We investigated the full mutational spectrum of genomic variation accessible to short read sequencing in these clones. Of the clonal SNVs that differed from the parental line, none were within 100bp of a predicted off-target site. Of 28 small indels identified across all clones, only one (from a TALEN treatment) was in proximity to a predicted off-target site, and none were attributable to CRISPR-Cas treatment. We also surveyed all balanced and unbalanced structural variation (SV) using the convergence of three different paired-end and split read mapping methods, as well as focal read depth analysis. Two clonal SVs were detected: a 5.5kb deletion with no discernible homology to the forecasted on- or off-target sites, and an intriguing 564bp interchromsomal insertion at the on-target site of *LINC00116* that exhibited five bases of microhomology with the target locus, suggesting homology-directed repair of the double-stranded DNA break. Finally, we also find support for integration of the Cas9 vector into the host genome. These results suggest that off-target mutations related to TALEN/CRISPR genome editing technologies are relatively infrequent, though such effects could present confounds when modeling disease and warrant consideration.

## 335

**Constitutional *BRCA1Italic Text* methylation is a major predisposition factor for high-grade serous ovarian cancer.** *A. Dobrovic[1,2,3], T. Mikeska[2,3], K. Alsop[2], G.V. Zapparoli[1], I.L. Candiloro[2,3], J. George[2], G. Mitchell[2], D. Bowtell[2,3], Australian Ovarian Cancer Study.* 1) Translational Genomics & Epigenomics Lab, Ludwig Institute for Cancer Research, Olivia Newton-John Cancer and Wellness Centre, Heidelberg (Melbourne), Victoria, Australia; 2) Research Division, Peter MacCallum Cancer Centre, East Melbourne, Victoria, Australia; 3) Department of Pathology, University of Melbourne, Parkville, Victoria, Australia.

Constitutional methylation refers to specific promoter methylation present in the normal tissues of some individuals. This study aimed to determine whether constitutional methylation of the *BRCA1* promoter region was a predisposition factor for high-grade serous ovarian cancer. We had previously shown in a case-control study of early-onset breast cancer patients that the presence of *BRCA1* methylation in the blood was associated with tumors that phenocopied pathogenic germline *BRCA1* mutations. This indicated that *BRCA1* methylation predisposed to and drove the development of these tumors. Germline *BRCA1* mutations also predispose to ovarian cancer, in particular, high-grade serous ovarian cancer. In collaboration with the Australian Ovarian Cancer Study, we determined the presence of mutations and detectable *BRCA1* methylation in 154 high-grade serous cancers. Fifteen women had germline *BRCA1* mutations and 11 had germline *BRCA2* mutations. The germline *BRCA1* mutation carriers were predominantly younger at diagnosis and 11/15 pathogenic *BRCA1* mutations were seen in patients under 55 at diagnosis. Similarly, patients with *BRCA1* methylated tumors were predominantly (15/25) under 55 at diagnosis. *BRCA1* methylation was mutually exclusive with *BRCA1* and *BRCA2* mutation. Nineteen patients showed detectable levels of *BRCA1* methylation in DNA extracted from their peripheral blood. Remarkably, when the corresponding tumor samples were assessed for methylation, 13 of these 19 patients had high level *BRCA1* methylation in their tumor ($p<.0001$ for the association of blood and tumour methylation). When the 50 tumors occurring before the age of 55 were considered, 9 of the 10 patients with detectable peripheral blood mutation had a corresponding *BRCA1* methylated tumor ($p<.0001$). This data indicates that constitutional *BRCA1* methylation can drive high-grade serous ovarian cancer, in particular early onset cancer, and moreover is as important a predisposition factor as *BRCA1* mutation.

## 336

**Candidate causal variants from three independent genetic signals at the 5q11.2 breast cancer risk locus regulate *MAP3K1*.** *D.M. Glubb[1], K.A. Pooley[2], K. Michailidou[3], M.J. Maranian[3], K.B. Meyer[4], J.A. Betts[1], K.M. Hillman[1], S. Kaufmann[1], G. Chenevix-Trench[1], D.F. Easton[2,3], A.M. Dunning[2], S.L. Edwards[1,5], J.D. French[1,5], Breast Cancer Association Consortium.* 1) Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Australia; 2) Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK; 3) Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 4) Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK; 5) School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Australia.

A genome wide association study has previously found a variant, rs889312, at 5q11.2 to be associated with breast cancer risk. To identify the causal variant(s) underlying this association, we analysed 909 genetic variants across 5q11.2 in 103,991 breast cancer cases and controls from 52 studies. Three sets of independent, correlated, highly trait-associated variants (iCHAVs), spanning a 183 kb region at 5q11.2, were identified in Europeans (46,451 cases and 42,599 controls). This region encompasses *MAP3K1*, a frequently mutated gene in breast cancer, encoding a stress-induced serine/threonine kinase which regulates apoptosis and induces cell proliferation through a RAS/RAF/MEK/ERK pathway. Using ENCODE data and chromatin conformation studies, we found that variants from the three iCHAVs coincide with five putative regulatory elements (PREs). We show that all five PREs physically interact with the *MAP3K1* promoter through chromatin looping and have effects on *MAP3K1* promoter activity in reporter gene assays. Analysis of iCHAV variants in the PREs revealed four variants that significantly alter *MAP3K1* promoter activity: rs74345699 and rs62355900 (iCHAV1), rs16886397 (iCHAV2) and rs17432750 (iCHAV3). The risk (minor) alleles of iCHAV1 SNPs r74345699 and rs62355900 increase the effect of the PRE3 enhancer on *MAP3K1* promoter activity in MCF7 breast cancer cells after estrogen stimulation. The risk (minor) allele of the iCHAV2 SNP rs16886397 confers enhancer activity on PRE5 and upregulates *MAP3K1* promoter activity in MCF7 cells. The protective (minor) allele of the iCHAV3 SNP rs17432750 diminishes the effect of a PRE2 enhancer on *MAP3K1* promoter activity in normal mammary epithelial cells. Furthermore, the protective allele of rs17432750 was shown by chromatin immunoprecipation to reduce GATA-3 binding and was associated with allele-specific chromatin looping between PRE2 and the *MAP3K1* promoter. Notably, the risk associated alleles of all four functional iCHAV SNPs increased *MAP3K1* transcriptional activity. Therefore, we propose that these risk alleles contribute to breast cancer risk by increasing *MAP3K1* expression and promoting cancer cell survival and proliferation.

## 337

**A profile of inherited predisposition to breast cancer among Nigerian women.** *Y. Zheng[1], T. Walsh[2], F. Yoshimatsu[1], M. Lee[2], S. Gulsuner[2], S. Casadei[2], A. Rodriguez[3], T. Ogundiran[4], C. Babalola[5], O. Ojengbede[6], D. Sighoko[1], R. Madduri[3], M-C. King[2], O. Olopade[1]*. 1) Center for Clinical Cancer Genetics and Global Health, Department of Medicine, The University of Chicago, Chicago, IL, USA; 2) Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA; 3) Computation Institute, The University of Chicago, Chicago, IL, USA; 4) Department of Surgery, University of Ibadan, Ibadan, Oyo, Nigeria; 5) Institute for Advanced Medical Research and Training, College of Medicine & Faculty of Pharmacy, University of Ibadan, Ibadan, Oyo, Nigeria; 6) Centre for Population & Reproductive Health, College of Medicine, University of Ibadan, Ibadan, Oyo, Nigeria.

Women of the African Diaspora disproportionately experience early-onset, hormone-receptor negative, aggressive breast cancer. This burden may be due at least in part to differences in the distribution of inherited mutations in breast cancer genes. But little is known about inherited mutations among African women with breast cancer. To obtain a comprehensive mutational profile, we are using the targeted capture and multiplex sequencing panel, BROCA to evaluate known and emerging breast cancer genes in genomic DNA of 1000 breast cancer patients and 1000 female controls from Ibadan, Nigeria; 400 cases and 400 controls have been sequenced thus far. Our analysis includes nonsense and frameshift mutations, gene-disrupting CNVs, missenses if experimentally established as damaging, and mutations at highly conserved sites (GERP>5.0) that are predicted by at least two in silico tools to disrupt splicing. The mutational profile of the Nigerian patients is strikingly different than mutational profiles of patients of other ancestries. Of the 400 Nigerian patients, 73 (18%) carry a mutation in any of 10 different genes: 43 in *BRCA1*, 18 in *BRCA2*, 3 in *PALB2*, 2 each in *BRIP1*, *RAD51C*, and *TP53*, and 1 each in *CHEK2*, *CHEK1*, *GEN1*, and *NBN*. One patient carries both *BRCA2* and *BRIP1* mutations. This profile is distinctive in that *BRCA1* and *BRCA2* represent more of the mutational burden compared to other genes, *PALB2* is a relatively minor contributor, *CHEK2* is a very minor contributor, and *ATM* is completely absent. At the allelic level, the Nigerian profile is highly heterogeneous: the 43 *BRCA1* patients carry 27 different mutations and the 18 *BRCA2* patients carry 16 different mutations. Furthermore, 10 mutations (8 in *BRCA1* or *BRCA2*), at completely conserved sites with very high GERP scores, were predicted to disrupt splicing. Most of these mutations were new to us, represent a higher rate of splice mutations than we have observed in any other population, and call for experimental assessment. Of the 400 controls, 9 (2%) carried a mutation in 7 genes: 2 in *BRCA1* (ages 35 and 45), 2 in *BRCA2* (ages 39 and 40) and 1 each in *BARD1*, *ATR*, *FAM175A*, *MRE11A*, and *SLX4*. Relative risks were 21.5 (95% CI [5.2, 88,2]) for *BRCA1* and 9.0 (95% CI [2.1, 38.5]) for *BRCA2*. To understand the burden of aggressive breast cancer among women of African ancestry will require comprehensive comparative genetic and clinical analysis of large cohorts of patients. This profile is intended as a beginning.

## 338

**Estimates for inherited mutations in breast cancer susceptibility genes among triple-negative breast cancer patients.** *F. Couch[1], S.N. Hart[1], P. Sharma[2], A. Ewart Toland[3], X. Wang[1], P. Miron[4], J.E. Olson[1], A. Godwin[2], V.S. Pankratz[1], C. Olswold[1], M.W. Beckmann[5], W. Janni[6], B. Rack[7], A. Ekici[8], D.J. Slamon[9], I. Konstantopoulou[10], F. Fostira[10], G. Fountzilas[11], L. Pelttari[12], S. Yao[13], J. Garber[4], A. Cox[14], H. Brauch[15], C. Ambrosone[13], H. Nevanlinna[12], D. Yannoukakos[10], S.L. Slager[1], C.M. Vachon[1], D.M. Eccles[16], P.A. Fasching[5]*. 1) Prof/Dept Lab Med, Mayo Clinic Col Med, Rochester, MN; 2) Department of Medicine, University of Kansas Medical Center, Kansas City, KS; 3) Departments of Internal Medicine and Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH; 4) Dana Farber Cancer Institute, Boston, MA; 5) Department of Gynecology and Obstetrics, University Hospital Erlangen, Erlangen, Germany; 6) Department of Gynecology and Obstetrics, University Hospital Ulm, Ulm, Germany; 7) Department of Gynecology and Obstetrics, Ludwig-Maximilians-University Munich, Campus Innenstadt, Munich, Germany; 8) Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany; 9) University of California at Los Angeles, David Geffen School of Medicine, Department of Medicine, Division Hematology/Oncology, Los Angeles, CA; 10) National Centre for Scientific Research "Demokritos", Athens, Greece; 11) Department of Medical Oncology, “Papageorgiou” Hospital, Aristotle University of Thessaloniki School of Medicine, Greece; 12) Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland; 13) Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY; 14) Department of Neuroscience, University of Sheffield, Sheffield, UK; 15) Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany; 16) Faculty of Medicine, University of Southampton, Southampton, UK.

Guidelines recommend germline mutation testing of breast cancer predisposition genes in triple negative (TN) breast cancer cases with a family history of breast or ovarian cancer or when diagnosed under age 60. However, the prevalence of mutations in these genes among TN cases unselected for family history of breast or ovarian cancer is not known. To assess the frequency of mutations in 16 predisposition genes in TN cases we screened a large cohort of 1824 TN patients unselected for family history of breast or ovarian cancer from 12 centers for mutations using a panel-based sequencing approach. Deleterious mutations were identified in 15% of TN patients: 8.5% had BRCA1, 2.7% had BRCA2, and 3.6% had mutations in 12 other genes. Mutations in non-BRCA1/2 genes encoding proteins implicated in homologous recombination repair of DNA double strand breaks were detected at the same frequency as in breast cancer families. TN cases with mutations were diagnosed at an earlier age than non-mutated cases, although 10% of TN cases diagnosed at ≥60 years and 5% with no family history of cancer were also found to carry likely mutations. Frequency estimates for mutations in the predisposition genes based on age of diagnosis and family history of cancer were also developed. These will allow for selection of TN patients most likely to carry mutations in predisposition genes.

## 339

**Inherited mutations in ovarian cancer - *PALB2* and *BARD1* are likely ovarian cancer susceptibility genes.** *B. Norquist[1], M.I. Harrell[1], M.F. Brady[2], T. Walsh[3], M.K. Lee[3], S. Gulsuner[3], Q. Yi[3], S. Casadei[3], S. Bernards[1], S.A. Davidson[4], R.S. Mannel[5], P.A. DiSilvestro[6], M.C. King[3], M.J. Birrer[7], E.M. Swisher[1, 3].* 1) Department of Obstetrics and Gynecology, University of Washington, Seattle, WA; 2) Gynecologic Oncology Group Biostatistical Office, Roswell Park Cancer Institute, Buffalo, NY; 3) Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA; 4) Department of Obstetrics and Gynecology, University of Colorado, Denver, CO; 5) Department of Obstetrics and Gynecology, University of Oklahoma, Oklahoma City, OK; 6) Department of Obstetrics and Gynecology, Brown University, Providence, RI; 7) Dana-Farber/Harvard Cancer Center, Department of Medicine, Harvard Medical School, Boston, MA.

To determine the frequency and significance of germline mutations in cancer-associated genes, we evaluated 1889 women with epithelial ovarian, fallopian tube, and primary peritoneal cancer (OC). We sequenced germline DNA using the targeted capture and multiplex sequencing panel BROCA. OCs were ascertained from three sources, two phase III clinical trials in newly diagnosed advanced stage ovarian cancer (GOG 218, N=773; GOG 262, N=573) and a university-based gynecologic oncology tissue bank (N= 543). Cases were not selected for age or family history. Mutation rates excluding CNVs were compared to population rates for European-Americans from the NHLBI GO Exome Sequencing Project (ESP), for which insertions and deletions of targeted genes were re-evaluated by hand. Only clearly damaging mutations were included. Of 1889 subjects, 381 (20%) had 390 germline mutations in cancer-associated genes. Of the 381 mutations carriers, 8 (2.1%) had more than one mutation, 278 (14.7%) had mutations in *BRCA1* (182) or *BRCA2* (96), and 109 (5.8%) had mutations in the following genes: *BRIP1* (27), *CHEK2* (13), *RAD51D* (11), *PALB2* (11), *ATM* (11), *RAD51C* (10), *NBN* (9), *TP53* (6), *BARD1* (4), *MSH6* (4), *FAM175A* (3), *PMS2* (2), and *MLH1* (1). Consistent with their role as ovarian cancer susceptibility genes, *BRIP1*, *RAD51C*, and *RAD51D* were significantly more frequently mutated in OC patients than in the ESP (all p<0.001). *PALB2* and *BARD1*, which are not proven ovarian cancer genes, were also significantly more frequently mutated in OC (*PALB2*: OR=11.0, 95% CI [2.4, 50], p= 0.0003; *BARD1*: OR=31.0, 95% CI [1.7, 577], p=0.02. *ATM*, *NBN*, *CHEK2*, and *FAM175A* mutations were not significantly more common in OC. Patients with *BRCA2* mutations had significantly better progression-free survival (p=0.009) and overall survival (p=0.0005). *BRCA2* carriers had more grade 4 neutropenia (p=0.04). Histologic subtype and primary site did not predict mutation status. Additional clinical data will be available at time of presentation regarding treatment response by mutation status. In summary, mutations in *PALB2*, *BARD1*, *BRIP1*, *RAD51C*, and *RAD51D* were more frequent in OC than in the ESP EA population. Mutation status affects treatment response and should be accounted for when designing and reporting clinical trials.

## 340

**Implementing *PALB2* gene testing in breast and ovarian cancer patients in UK.** *N. Rahman[1,2], E. Ruark[1], S. Seal[1], A. Renwick[1], E. Ramsay[1], S. Powell[1], M. Warren-Perry[1], H. Hanson[1], C. Lord[3], C. Turnbull[1].* 1) Division of Genetics & Epidemiology, The Institute of Cancer Research, London, United Kingdom; 2) Cancer Genetics Unit, Royal Marsden NHS Foundation Trust, London, United Kingdom; 3) Breast Cancer Research Division, The Institute of Cancer Research, London, United Kingdom.

In 2007 we reported that *PALB2* mutations confer moderately increased risks of breast cancer. The cancer association was confirmed by many other groups, but the risks conferred have been variable, with some studies suggesting much higher risks, potentially warranting interventions similar to those offered to *BRCA1/2* mutation carriers. Together with the advent of affordable gene panel testing this has led to *PALB2* testing entering the clinical arena. To explore clinical utility of *PALB2* testing in UK we performed *PALB2* mutation analysis in 5982 samples detecting mutations in 58/3906 cancer cases vs 1/2076 population controls (P= $4\times10^{-10}$).Via segregation analysis of the full pedigree information from 3200 families ascertained on the basis of personal ± family history of breast and ovarian cancer through Genetic clinics, and 2076 controls we found a 2.1 fold increased risk of breast cancer (95% CI=1.7-2.6, P=$3.35\times10^{-12}$), consistent with our previous estimate, and a 2.6 fold increased risk of ovarian cancer (95% CI=1.5-4.7, P=0.001). We also confirmed that PALB2 defects generate sensitization to PARP inhibitors and show that this is as extensive as that caused by gene silencing of BRCA1 or BRCA2. These data confirm *PALB2* mutations predispose to breast cancer and provide the first robust evidence of an association with ovarian cancer. Testing for PALB2 mutations in cancer patients that are having BRCA mutation analysis anyway is justifiable (if it incurs no additional cost), as it may aid treatment decisions, for example eligibility to PARP inhibitor trials. However, caution is required in using *PALB2* mutations in a predictive capacity in unaffected relatives. Overall, our large analysis suggest the cancer risks in UK families are moderate and not of the level to warrant major interventions such as mastectomy, unless additional factors present (e.g. a strong family history). Larger prospective studies will be invaluable in clarifying cancer risks further.

## 341

**Functional variant assays (FVAs) for predicting breast cancer risks of genetic variants in the DNA double-stranded break repair pathway.** *H. Ostrer[1], A. Pearlman[1], K. Upadhyay[1], Y. Shao[2], J. Loke[1].* 1) Albert Einstein College of Medicine, Bronx, NY; 2) NYU School of Medicine, New York, NY.

Introduction. Among the 5-10% of women from high-risk breast cancer families, 20% have a mutation in the genes, *BRCA1* and *2*, and others have mutations in related genes that play a role in repair of DNA double-stranded breaks (DSB). In response to DNA breaks, the proteins encoded by these genes bind to each other, are transported into the nucleus and initiate non-homologous recombination. Mutations in these genes may disrupt any of these processes. Methods and Methods. Lymphoblastoid cells (LCLs) were collected from individuals with three different sets of *BRCA1* variants -- known pathogenic mutations, benign variants, and variants of uncertain significance (VUS) and from individuals with known *BRCA2*, *FANCC,* and *NBN* mutations. Functional variant assays (FVAs) were developed to determine whether variants in *BRCA1* or in the other DSB repair genes disrupted normal functions, such as binding of BRCA1 to its protein partners, BARD1, PALB2, BRCA2 and FANCD2, the phosphorylation of p53, or the transport of BRCA1 into the nucleus in response to crosslinking drugs, diepoxybutane (DEB) and mitomycin C (MMC), and the DNA breakage drug, bleomycin (Bleo). LCLs for the benign variants and VUS were sequenced via the 1000 Genomes Project and analyzed for the presence of possible genetic modifiers. Results. Mutations in *BRCA1* decrease nuclear localization of BRCA1 in response to the DEB (p=$9.8\times10^{-32}$), MMC (p=$4.8\times10^{-16}$), Bleo (p= $6.4\times10^{-44}$), or drug combination (p=$4.8\times10^{-20}$). Mutations in *BRCA1* reduce binding to co-factors, PALB2 (p=$2.2\times10^{-18}$), BRCA2 (p=$3.0\times10^{-7}$), and FANCD2 (p=$3.1\times10^{-7}$). Mutations in *BRCA1* decrease phosphorylation of p53 (p=$4.5\times10^{-23}$), as do VUS in *BRCA1* (p=$3.0\times10^{-18}$). Mutations in *BRCA2*, *FANCC* and *NBN* decrease nuclear localization of BRCA1 in response to the DEB (p=$7.7\times10^{-23}$), MMC (p=$4.0\times10^{-18}$), Bleo (p=$7.8\times10^{-43}$), or drug combination (p=$9.2\times10^{-14}$), reduce binding to co-factors, PALB2 (p=$5.6\times10^{-15}$) and FANCD2 (p=$5.8\times10^{-5}$), and decrease phosphorylation of p53 (p= $3.0\times10^{-18}$). Unsupervised analysis of all of these assays demonstrated two apparent clusters, high-risk *BRCA1* mutations and phenocopies and low-risk *BRCA1* controls and VUS. Conclusion. FVA assays distinguish *BRCA1* mutations from benign variants and categorize most VUS as benign. Mutations in other DSB repair genes produce molecular phenocopies with these assays. FVA assays represent an adjunct to sequencing for categorizing VUS and a possible stand-alone test for assessing breast cancer risk.

## 342

**Genome-wide association study of progression-free survival in ovarian cancer patients treated with carboplatin and paclitaxel identifies an enhancer that regulates three nearby genes.** *G. Chenevix-Trench[1], Y. Lu[1], J. Beesley[1], K. Hillman[1], S. Edwards[1], S. Johnatty[1], S. Macgregor[1], B. Gao[2], J. French[1], A. deFazio[2] on behalf of the Ovarian Cancer Association Consortium.* 1) QIMR Berghofer Medical Research Institute; 2) Department of Gynaecological Oncology and Westmead Institute for Cancer Research, University of Sydney at the Westmead Millennium Institute, Westmead Hospital, Sydney, Australia.

Women diagnosed with advanced epithelial ovarian cancer are commonly treated with cytoreductive surgery followed by platinum/taxane chemotherapy, but there is considerable inter-individual variation in response. We have carried out a three-phase genome-wide association study of progression-free survival in a total of 1,053 serous ovarian cancer patients, from ten independent cohorts, with inclusion criteria that included uniform treatment with carboplatin and paclitaxel as first-line therapy. Fine mapping of our top region identified two rare SNPs in tight linkage disequilibrium in the TTC39B gene associated with progression free survival (P=1.2×10-7, Hazard Ratio = 2.5, 95%CI = 1.8 to 3.6). The minor allele of rs7874043 was associated with worse progression-free survival in all ten contributing sites, and there was no significant heterogeneity between sites. The minor allele was also associated with worse overall survival (P=5.6×10-3, Hazard Ratio = 1.75, 95%CI = 1.18 to 2.59). Through chromosome conformation capture and luciferase assays, we showed that the rare alleles (minor allele frequency = 0.02) that were associated with progression-free survival lie in an enhancer, and differentially interact with an alternative, but not the canonical, promoter of TTC39B. Common SNPs in TTC39B have previously been found to be associated with high density lipoprotein-cholesterol levels, so our findings suggest a possible role of cholesterol metabolism in serous ovarian cancer outcome. However, additional chromosome conformation capture experiments showed that the enhancer also interacts with the promoters of CCDC171 and PSIP1, which encodes a transcriptional co-activator. Luciferase assays are currently underway to confirm the role of the enhancer in regulating CCDC171 and PSIP1, and to determine which is the functional SNP.

## 343

**Emerging patterns of schizophrenia risk conferred by de novo mutation.** *D. Howrigan[1,2], B. Neale[1,2], K. Samocha[1,2], J. Moran[2], K. Chambert[2], S. Rose[2], M. Fromer[2,3], S. Chandler[4], N. Laird[5], H.G. Hwu[6], W.J. Chen[6], S. Faraone[7], S. Glatt[7], M. Tsuang[4], S. McCarroll[8].* 1) Massachusetts General Hospital, Boston, MA; 2) Broad Institute, Cambridge, MA; 3) Mount Sinai School of Medicine, New York, NY; 4) University of California San Diego, La Jolla, CA; 5) Harvard School of Public Health, Boston, MA; 6) National Taiwan University, Taiwan; 7) SUNY Upstate Medical University, Syracuse, NY; 8) Harvard University, Cambridge, MA.

Increased rates of deleterious de novo mutations have emerged as significant genetic risk factors among developmental disorders such as autism, intellectual disability, and epilepsy. In contrast, only modest effects of de novo mutation have been discovered so far among schizophrenia cohorts. In the current study, whole-exome sequencing has been performed on 1,110 complete trios from Taiwan with a sporadic schizophrenia diagnosis in the offspring. Exome sequencing data were generated using the Illumina HiSeq sequencing with the Agilent SureSelect exome capture platform, and validation of candidate de novo signals was analyzed using targeted high-throughput genotyping on Illumina HiSeq and Illumina MiSeq platforms. Confirmed de novo mutations were annotated using the NCBI RefSeq database. De novo mutation rates per trio and across the exome fall in line with the expected mutation rate. Using models that incorporate gene size and site-specific mutation rates into expectations of de novo mutation rates, we do not observe any single gene that surpasses exome-wide correction for multiple testing (set at p=1e-6). We do, however, see a significant enrichment of genes with multiple non-synonymous de novo mutations (empirical p= 7e-4). When we combine our results with published de novo studies of schizophrenia, we observe nine genes with multiple loss-of-function events (empirical p<1e-4), and 87 genes with multiple missense events (empirical p=0.01). Gene set analyses also indicate that both loss-of-function and missense de novo mutations are enriched among targets of the Fragile X mental retardation protein (p=0.004 and p=6e-5, respectively) and among genes under evolutionary constraint (p=0.001 and p=2e-5, respectively). The current findings do not identify any single gene as an unequivocal risk factor for schizophrenia when disrupted by de novo mutation; however, aggregate analyses of genes hit with multiple damaging mutations and among well characterized gene sets in the literature indicate that significant patterns of de novo risk for schizophrenia are clearly emerging. We firmly believe larger cohorts and accumulation of de novo mutations in the literature will soon lead to specific genes being unequivocal risk factors, but remain aware that the increased liability due to de novo mutations comprises a modest fraction of the overall genetic liability in schizophrenia.

## 344

**Discoveries from a Genome-Wide Analysis of CNVs in the PGC Study of Schizophrenia.** *J. Sebat[1], C.R. Marshall[2], D. Howrigan[3], D. Merico[1], B. Thiruvahindrapuram[2], W. Wu[1], M. O'Donovan[4], S. Scherer[2], B. Neale[3], Schizophrenia and CNV analysis groups of the Psychiatric Genomics Consortium (PGC).* 1) Department of Psychiatry & the Institute for Genomic Medicine, University of California San Diego, La Jolla, CA; 2) The Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto Canada; 3) Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 4) Cardiff University, Cardiff, UK.

Copy number variants (CNVs) throughout the genome contribute to schizophrenia, and evidence has been obtained for several CNV loci. However, the full extent of the CNV contribution to risk is unknown. Previous studies have lacked adequate power to detect genetic association for CNVs with low frequencies (MAF<0.001) or intermediate effect sizes (OR = 2-10). Identification of such risk factors requires sample sizes achievable only through large scale collaboration. To this end we developed a centralized CNV analysis pipeline, composed of multiple available calling tools, and applied it to the investigation of CNVs in a cohort of 46,288 subjects (23,650 cases and 22,638 controls) from the Psychiatric Genomics Consortium study of Schizophrenia. Following the processing of raw data, samples were filtered within datasets based on array QC metrics (probe variance, GC bias and aneuploidy). A consensus CNV call set was generated from the intersection of multiple callers and CNVs were filtered within each dataset based on frequency (<1%), probe density, size, overlap with segmental duplications or Immunoglobulin & TCR regions. A set of appropriate covariates for analysis was identified by examining the correlation of QC metrics with case status and CNV burden across datasets. Analysis of CNV burden was performed genome-wide and within functional gene sets. Genetic association was carried out as single marker (breakpoint) and gene-based tests (collapsing rare variants). Association was tested by logistic regression with covariates, and empirical P-values were estimated by permutation and converted to Z-scores to refine the accuracy of empirical significance. Appropriate thresholds for genome-wide significance were estimated by permutation. Results reveal a robust contribution of CNV to disease risk that is consistent across a wide range of microarray platforms and studies. CNV burden was enriched among gene sets involved in neurological function including the postsynaptic density (PSD) and genes associated with neurological phenotypes in animal models. Genome-wide significant evidence was obtained for CNVs at 2p16.3, 3q29, 16p11.2, 15q13.3, 22q11.2 and additional novel loci. Furthermore, the centralized CNV calling pipeline enables fine-scale delineation of select loci to the level of sinle genes. Our findings suggest that analysis of CNV in large GWAS datasets can advance our knowledge of rare genetic variants that contribute to risk for schizophrenia.

## 345

**A functional role for non-coding variation in schizophrenia genome-wide significant loci.** *P. Sklar[1], A. Mitchell[1], G. Voloudakis[1], V. Pothula[1], E. Stahl[1], A. Georgakopoulos[1], D. Ruderfer[1], J. Fullard[1], A. Charney[1], Y. Okada[2], K. Siminovitch[3], J. Worthington[4], L. Padyukov[5], L. Klareskog[5], P. Gregersen[6], R. Plenge[7], S. Raychaudhuri[7], M. Fromer[1], S. Purcell[1], K. Brennand[1], M. Fromer[1], N. Robakis[1], E. Schadt[1], S. Akbarian[1], P. Roussos[1].* 1) Icahn School of Medicine at Mount Sinai, New York, NY; 2) Department of Human Genetics and Disease Diversity, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan; 3) Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario; 4) Arthritis Research UK Centre for Genetics and Genomics, Musculoskeletal Research Centre, Institute for Inflammation and Repair, University of Manchester, Manchester Academic Health Science Centre, Manchester; 5) Rheumatology Unit, Department of Medicine (Solna), Karolinska Institutet, Stockholm; 6) The Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York; 7) Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.

Our recent GWAS in schizophrenia (SCZ) identified 22 loci that reached genome-wide significance (Ripke et al. *Nat Genet* 2013). The majority of identified SNPs reside within non-coding regions. In order to to understand these associations mechanistically, it is important to develop strategies for honing in on regions and SNPs more likely to have functional effects. Functional annotations were developed in a variety of ways. Brain eQTLs were generated in 8 datasets. Brain cis-regulatory elements (CRE) (active promoter, active enhancer, poised promoter, repressed enhancer and open chromatin regions) were generated based on ChIP-seq of histone modifications. Next, GWAS SCZ SNPs were classified into categories: eQTL, CRE, eQTL in a cis regulatory element (creQTL) or functionally unannotated variants. Relative enrichment for the categories was calculated using an empirical cumulative distribution of the GWAS P values after controlling for genomic inflation. We mapped the physical interaction of enhancers in two genes (*CACNA1C* and *NGEF*) with the transcription start site of each gene in human prefrontal cortex (n=6) and hiPSC derived-neurons by chromosome conformation capture (3C) assay. The largest enrichment of GWAS SNPs occurs in eQTLs, active promoters and enhancers. Enrichment is greater when the combined creQTL functional category is analyzed for all types of CREs (CRE range: 1.58-7.08 fold; creQTL range: 4.06-29.51 fold). We detected overlapping eQTL and GWAS signals using the regulatory trait concordance score for 10 of 22 intervals, four times the number expected by chance ($P=2\times10^{-5}$). The SCZ-related eQTLs are associated with expression of 17 genes, 5 of which are associated with loci within enhancers. For *CACNA1C* and *NGEF* genes, we identified enhancer regions that demonstrate increased interaction with the promoter and affect transcriptional activity of each gene. Our findings point to a functional link between SCZ-associated non-coding SNPs and 3-dimensional genome architecture associated with chromosomal loopings and transcriptional regulation in the brain.

## 346

**Comprehensive, integrative and hypothesis-free pathway analysis of genome-wide association data highlights synaptic transmission, dendritic spines and the post-synaptic density in schizophrenia.** *T.H. Pers[1,2], S. Ripke[3,4], L. Franke[5], J.N. Hirschhorn[1,2,6] for the Psychiatric Genetics Consortium.* 1) Division of Endocrinology, Children's Hospital Boston, Boston, Massachusetts, USA; 2) Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 3) Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA; 4) Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; 5) Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands; 6) Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.

Copy number variation studies and exome sequencing studies have supported hypotheses for the involvement of calcium channels, glutamatergic signaling, synapse plasticity and postsynaptic signaling pathways in schizophrenia. However, analysis of these genetic results with gene set enrichment analysis has been restricted to gene sets corresponding to these a priori hypotheses. Recently the Psychiatric Genomics Consortium identified 108 independent genome-wide significant loci for schizophrenia (based on 34,241 cases and 45,604 controls), but comprehensive, hypothesis-free gene set enrichment approaches failed to identify significantly enriched pathways. Therefore, additional etiologic pathways and the likely causal gene(s) at many associated loci have yet to be discovered. Recently we developed an integrative computational framework - Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT) - which uses gene function predictions derived from 14,461 gene sets and 77,840 microarray samples to systematically identify the most likely causal gene at associated loci, pathways enriched in enriched in genetic associations, and tissues and cell types in which genes from associated loci are highly expressed. DEPICT is already being widely used; here, we applied it to the 108 genome-wide significant schizophrenia loci to prioritize genes and highlight enriched gene sets. We validate our findings with a data set of rare disruptive variants from exome sequencing studies of 5,079 Swedish schizophrenia cases and controls. Based on microarray expression data from 37,427 samples spanning 209 tissue/cell types (a part of the DEPICT framework), we show that genes near associated loci are highly expressed in the brain and mononuclear leukocytes cells at false discovery rates (FDR) below 5%). Next, we show that DEPICT can identify 73 overlapping gene sets (13 after pruning) that are enriched in the associated loci at FDR below 5%). In particular, in this unbiased, hypothesis-free analysis, we highlight pathways related to the function of postsynaptic structures, including the postsynaptic density, and also dendritic spine formation and behavioral mouse phenotypes (all at FDR below 1%). Finally, we report 20 genes that are prioritized across 16 genome-wide significant loci (at FDR below 5%), and show that these genes are more likely to carry rare disruptive mutations in schizophrenia cases compared to controls ($P < 0.016$).

## 347

**Integrating network analyses and genetics with large-scale RNA-sequencing of schizophrenia brains.** *M. Fromer[1,2] for the CommonMind Consortium, Swedish Schizophrenia Consortium, Schizophrenia Working Group of PGC.* 1) Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY; 2) Psychiatric Genomics, Mount Sinai School of Medicine, New York, NY.

The most recent schizophrenia GWAS reported >100 associated loci, implying a high degree of polygenicity. To better understand the pathology of neuropsychiatric disease, we formed the CommonMind Consortium (commonmind.org) to generate large-scale data (RNA-seq, ChIP-seq, DNA-seq/genotyping) from human post-mortem brain samples. Here, we identify functional changes in gene expression using RNA-seq of 554 samples (265 schizophrenia cases and 289 controls) from the dorsolateral prefrontal cortex (BA9/46). Clinical (gender, age of death, medications) and technical (brain bank, post-mortem interval, RNA quality, sequencing batch) covariates, as well as hidden confounders, were controlled using surrogate variable analysis (SVA). Analyses are ongoing, but initial application of linear models implemented in voom/limma identified ~15% of expressed genes as differential between cases and controls (FDR 5%). These genes were nominally enriched for DNA variants associated with schizophrenia, including rare (frequency < 0.1%) nonsynonymous variants in Swedish case-control exome sequencing (p=0.012) and common GWAS loci (p=0.045). De novo loss-of-function (nonsense, frameshift, essential splice site) mutations in autism, intellectual disability, and epilepsy affected the differential genes (p=0.0053, 0.00058, 0.018), though this did not hold for schizophrenia de novos (p= 0.12). Preliminary gene coexpression networks constructed using WGCNA (Weighted Gene Co-expression Network Analysis) identified ~40 modules, 11 differentially expressed (FDR 5%). A differential module of ~1000 genes involved in synaptic transmission was seemingly enriched in rare nonsynonymous variants in case-control exomes (p=0.013), and a differential module of ~200 postsynaptic genes related to mitochondrial energy production showed enrichment of common GWAS loci (p=0.0093). The synaptic transmission module also tended to be enriched for loss-of-function mutations in autism, intellectual disability, and epilepsy (p=0.0015, 0.06, 0.035). Genes impacted by loss-of-function mutations in schizophrenia were enriched (p=0.00014), as were common GWAS loci (p=0.0044), in a non-differential glutamatergic signaling module. This large dataset will be made public in early 2015 and will include a catalogue of brain-expressed genes and isoforms, as well as eQTL, from cases and controls. This resource will facilitate novel discoveries relating neurobiology to disease risk and advance therapies.

## 348

**RNAseq Transcriptome Study Implicates Immune-Related Genes in Schizophrenia.** *A.R. Sanders[1,2], E.I. Drigalenko[3], J. Duan[1,2], W. Moy[1], J. Freda[1], M.G. S.[4], H.H.H. Göring[3], P.V. Gejman[1,2].* 1) Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, IL; 2) Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL; 3) Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX; 4) Molecular Genetics of Schizophrenia (MGS) Collaboration.

GWAS have implicated over 100 loci as being associated at genome-wide significant (GWS) levels with risk for schizophrenia (SZ), a common and severe psychotic disorder. Regulation of mRNA expression may be involved in the etiology for some of these loci. In a previous transcriptomic profiling study (Sanders et al., 2013), using arrays on lymphoblastoid cell lines from 268 SZ cases and 446 controls, we found 89 genes to be differentially expressed by affection status (FDR<0.05) and enriched for immune-related genes, consistent with hypothesized immune contributions to SZ risk. For the current study, we assayed expression by RNAseq on another set of 490 SZ cases and 662 controls, also of European ancestry from the Molecular Genetics of SZ dataset. After inverse normalization of expression data, we used regression analysis to identify genes differentially expressed by affection status, while simultaneously controlling for confounding effects (sex, age, ancestry, viral load, growth rate, energy status, RNAseq batch) as before, focusing here on the 8,466 genes detected well with both array and RNAseq. We found 807 genes to be differentially expressed by affection status (Bonferroni p<0.05) in this RNAseq dataset. These differentially expressed genes were somewhat enriched for genes involved in autoimmunity (Fisher p=0.01), and their pathway analyses showed gene ontology term enrichment (FDR<0.05) for categories such as response (immune, to virus), endoplasmic reticulum categories, activation (cell, leukocyte, lymphocyte), regulation of cell proliferation, and membrane (integral, intrinsic). Of the 89 genes differentially expressed in our previous array study, 80 had appreciable expression in the RNAseq data, and 18 of those were differentially expressed by affection status (Bonferroni p<0.05, all with the same direction of effect). Notable examples thereof included DBNDD2 (dysbindin domain containing 2, which mediates neural differentiation and apoptosis), B3GNT2 (UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 2, GWS association with rheumatoid arthritis), and SYT11 (synaptotagmin XI, GWS association with Parkinson's Disease). In future work, we will identify the expression quantitative trait nucleotides (eQTNs) for these differentially expressed genes and examine the relationship of these eQTNs to GWAS findings. This work was supported NIH grants RC2MH090030, R01MH094091, and R01MH094116.

## 349

**A rare regulatory noncoding variant in GWAS-implicated MIR137/MIR2682 locus potentially confers risk to both schizophrenia and bipolar disorder.** *J. Duan[1,2], J. Shi[3], A. Fiorentino[4], C. Leites[1], J. Chen[6], W. Moy[1], B. Alexandrov[7,8], D. He[1], J. Freda[1], A. Bishop[7,8], N.L. O'Brien[4], . MGS[9], . GPC[10], X. Chen[6], A. Usheva[7], A. McQuillin[4], A.R. Sanders[1,2], H.M.D. Gurling[4], M.T. Pato[5], K.S. Kendler[6], C.N. Pato[5], P.V. Gejman[1,2].* 1) Dept Psychiatry, Northshore Univ Healthsystem/Univ of Chicago, Evanston, IL; 2) Department of Psychiatry and Behavioral Sciences, The University of Chicago, Chicago, IL; 3) Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Insititute, Bethesda, MD; 4) Molecular Psychiatry Laboratory, Division of Psychiatry, University College of London, UK; 5) Department of Psychiatry and the Behavioral Sciences, Keck School of Medicine at USC, Los Angeles, CA; 6) Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA; 7) Department of Medicine, Endocrinology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA; 8) Los Alamos National Laboratory, Los Alamos, NM; 9) Molecular Genetics of Schizophrenia collaboration; 10) Genomic Psychiatric Cohort collaboration.

Genome-wide association studies (GWAS) have identified common genetic risk variants in >100 susceptibility loci for the risk developing schizophrenia (SZ); however, the contribution of rare variants at these loci remains to be explored. We sequenced one of the most strongly associated loci spanning two microRNAs, *MIR137* and *MIR2682*, in 2,610 SZ cases & 2,611 controls of European ancestry from the Molecular Genetics of Schizophrenia (MGS) study, covering ~6.9 kb of both microRNAs and the upstream ENCODE-annotated enhancers, promoters, and insulators. The burden test aggregating all the rare variants (MAF<0.5%) did not yield significant association with SZ; however, those within ENCODE-annotated promoters/enhancers were associated with SZ (P=0.021), and the association was improved (P=0.0007) when restricting the rare variants to those with MAF<0.1%;. We found an enhancer SNP, chr1_98515539, in 11 cases that was absent in controls (P=0.00048; Fisher's exact test). We further genotyped this SNP (or extracted genotypes for it from whole genome sequencing datasets) in 3 additional SZ samples (N=2,434 cases), 3 bipolar disorder (BP) samples (N=4,713 cases), SZ/BP study controls (N=3,572), and population controls (UK10K-TwinsUK, N=1,603; 1000Genome, N=85). In these additional datasets, we identified the rare allele in 2 SZ cases, 11 BP cases, and 3 controls (combined P for SZ, BP, and SZ/BP was 0.0004, 0.004, and 0.00029, respectively). Our reporter gene assay showed that the risk allele of chr1_98515539 reduced enhancer activity of its flanking sequence by >50%; in neuronal cells. Both an electrophoresis mobility shift assay and a local DNA breathing dynamics analysis confirmed the weakened transcription factor Ying-Yang 1 (YY1) binding by the risk allele of chr1_98515539. A chromatin conformation capture (3C) assay further indicated that the functional SNP influences M*IR137/MIR2682*, but not the nearby *DPYD* or *LOC729987*. Our results suggest the existence of rare SZ risk variants at the GWAS-implicated *MIR137/MIR2682* locus, with risk alleles decreasing *MIR137/MIR2682* expression. Furthermore, for the first time, we have shown that SZ and BP may share rare risk variants at the same locus.

## 350

**GWAS of Bipolar 1 Disorder in a Multi-ethnic Cohort of 72,823 Identifies Four Novel Loci.** *C. Schaefer[1], L. Shen[1], K. Kearney[1], M. McCormick[2], S.P. Hamilton[3], L.A. McInnes[3], V. Reus[4], J. Wall[5], P-Y. Kwok[5], M. Kvale[5], T.J. Hoffmann[5], E. Jorgenson[1], N. Risch[1,5].* 1) RPGEH, Division of Research, Kaiser Permanente, Oakland, CA; 2) Dept. of Psychiatry, Kaiser Permanente Santa Clara, Santa Clara, CA; 3) Dept. of Psychiatry, Kaiser Permanente San Francisco, San Francisco, CA; 4) Dept. of Psychiatry, UCSF School of Medicine, San Francisco, CA,; 5) Institute for Human Genetics, UCSF, San Francisco, CA.

Genome-wide association studies (GWAS) of bipolar disorder (BP) have identified susceptibility variants, and some have been replicated in the analysis of the Psychiatric Genetics Consortium (PGC), but the variation explained in BP is small and studies are considered under-powered. Most analyses have been performed in non-Hispanic whites. To examine potential associations with bipolar 1 disorder (BP1) in whites and other groups, we performed a GWAS and meta-analysis in a multi-ethnic sample of BP1 cases and screened controls, and examined consistency with results from the PGC. Cases and controls are participants in the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH). Cases were initially identified with multiple treatment episodes and/or hospitalization with BP1 from electronic medical records (EMR) with diagnosis confirmed by telephone interview. Individuals were excluded with diagnoses of schizophrenia spectrum disorder (SSD), schizoaffective disorder, or nonaffective psychoses. The resulting sample was 68% female; average age at study entry was 47 years, and 25% were minority or mixed race - ethnicity. Potential controls were drawn from the RPGEH Genetic Epidemiology Research on Adult Health and Aging general cohort. Controls were screened using EMR data to exclude those with one or more diagnosis of Axis 1 or other selected neuropsychiatric disorders. Cases and controls were genotyped using the Affymetrix Axiom system; four custom ancestry-specific arrays with 675,000 to 891,000 SNP markers were used to genotype the non-Hispanic white, Asian, African-American, and Latino participants separately. Following quality control and filtering, case-control analyses of non-Hispanic whites (3726 cases; 54620 controls) identified multiple variants in four regions with genome-wide significance ($p < 5.0E-08$): FAM160A1 on chr 4; SLC44A4 and CFB on chr 6; ENOX1 on chr 13; and RBFOX1 on chr 16. Analyses controlled for age, gender, and genetic ancestry. In analysis of the other race-ethnicity groups, SNPs on chr 13 were consistently associated with BP1 in the other groups, despite relatively small sample sizes. Our results generally do not replicate the significant associations found in the PGC meta-analysis, although two of the SNPs we identified on Chr 13 were associated with BPD in the PGC study.

## 351

**Partitioning heritability by functional category using summary statistics.** *H. Finucane[1,2,6], B. Bulik-Sullivan[3,9,11], A. Gusev[1,2], G. Trynka[3,7], P. Loh[1,2], H. Xu[2,8], C. Zang[2,8], S. Ripke[3,9], S. Purcell[3,4,5,9], M. Daly[3,9], E. Stahl[3,4], S. Raychaudhuri[3,7], S. Lindstrom[1], N. Patterson[3,10], B. Neale[3,9], A. Price[1,2,3], Schizophrenia Working Group of the Psychiatric Genetics Consortium.* 1) Dept of Epidemiology, Harvard School of Public Health, Boston, MA; 2) Dept of Biostatistics, Harvard School of Public Health, Boston, MA; 3) Broad Institute of MIT and Harvard, Cambridge, MA; 4) Dept of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY; 5) Dept of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY; 6) Dept of Mathematics, Massachusetts Institute of Technology, Cambridge, MA; 7) Division of Genetics and Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 8) Dana Farber Cancer Institute, Boston, MA; 9) Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA; 10) Department of Genetics, Harvard Medical School, Boston, MA; 11) Center for Neurogenomics and Cognitive Research, VU University Amsterdam, Amsterdam, The Netherlands.

Recent work has demonstrated that some functional categories of the genome contribute disproportionately to trait heritability. Partitioning heritability is traditionally done using a variance components approach; however this approach is not feasible at large sample sizes, and there are many datasets for which only summary statistics are available. Here, we introduce a new method for partitioning heritability that requires only GWAS summary statistics and LD information from a reference panel. Our method takes advantage of the fact that, under standard assumptions about genetic architecture, the expected chi-square statistic at SNP i is linear in the LD Score of SNP i, defined as the sum over SNPs j of r^2(i,j). In previous work (Bulik-Sullivan et al. 2014 bioRxiv), we used this relationship to differentiate between chi-square statistic inflation due to sample structure, which affects the intercept of a regression of chi-square statistic on LD Score, and inflation due to polygenicity, which affects the slope. Here, we use a multivariate regression of chi-square statistic on LD Scores specific to functional categories to partition heritability. Our method is robust to multiple causal variants at a locus, and obtains accurate estimates in simulations. On real WTCCC data across seven diseases it obtains results similar to the variance components approach and in a meta-analysis of these traits, infers significant enrichments for DNaseI Hypersensitivity Sites (DHS), histone marks and other functional categories. FANTOM5 enhancers (Andersson et al. 2014 Nature) were the most enriched, with 0.4% of the genome explaining 11.4% (s.e. 2.9%) of heritability (30x enrichment; P=7e-5). We applied the method to summary statistics from a schizophrenia (SCZ) dataset with 70,100 samples, and from a study of type two diabetes (T2D) with 69,033 samples. We found significant enrichment for many functional categories for both diseases; estimated enrichments tended to be much larger for T2D than for SCZ. Additionally, SNPs in fetal DHS regions were 4x enriched over SNPs in non-fetal DHS regions for SCZ (p =0.001), but not for T2D, which had a non-significant trend in the opposite direction (p = 0.058). Of the ten most enriched cell types for SCZ, three were brain cell types, and CD34 mobilized primary cells and embryonic stem cells were among the significantly enriched.

## 352

**Identification of multiple regulatory variants at the *GALNT2* human high-density lipoprotein cholesterol locus.** *T.S. Roman[1], A.F. Marvelle[1], M.P. Fogarty[1], S. Vadlamudi[1], M.L. Buchkovich[1], J.R. Huyghe[2], C. Fuchsberger[2], A.U. Jackson[2], K.J. Gaulton[1,3], A.J. Gonzalez[1], P. Soininen[4,5], A.J. Kangas[4], J. Kuusisto[6], M. Ala-Korpela[4,5], M. Laakso[6], M. Boehnke[2], K.L. Mohlke[1].* 1) Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI; 3) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 4) Computational Medicine, Institute of Health Sciences, University of Oulu, Oulu, Finland; 5) Nuclear Magnetic Resonance Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland; 6) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.

Genome-wide association studies have identified >150 loci associated with blood lipid and cholesterol levels; however, many of the underlying molecular and biological mechanisms are unknown. We sought to identify functional variants at the HDL-C-associated locus in *GALNT2* intron 1, and we evaluated the use of ENCODE open chromatin, histone modification and transcription factor ChIP-seq data to detect variants with allele-specific regulatory activity. To characterize the locus, we performed fine-mapping and conditional analyses using 9,797 individuals from the Metabolic Syndrome in Men (METSIM) study. We confirmed a single signal and observed the strongest evidence of association with total cholesterol in medium HDL ($P = 1.7\times10^{-12}$) among 228 analyzed traits. To identify a target gene, we performed allelic expression imbalance experiments in 36 human hepatocyte samples. Alleles associated with increased HDL-C showed a 7% increase in *GALNT2* expression ($P = 5.4\times10^{-7}$). All 23 intronic HDL-C-associated variants (index $r^2>.8$) overlap ≥1 mark of HepG2 or liver open chromatin, histone modification, or transcription factor (TF) ChIP-seq data from ENCODE or the Epigenomics Roadmap. We tested all 23 for allelic differences in HepG2 transcriptional reporter assays and observed strong allele-specific enhancer activity for the SNP rs2281721 (75- vs 26-fold) and for a 3-variant haplotype of rs4846913, rs2144300, and rs6143660 (49- vs 14-fold). Of these 4 variants, 3 overlap ≥22 open chromatin, histone ChIP-seq, and/or TF ChIP-seq peaks. No allelic differences were observed for 19 variants in histone ChIP-seq peaks (4 of the 19 also overlap open chromatin or TF ChIP-seq peaks). Reporter assays using site-directed alterations of the haplotype showed that rs4846913 and rs2144300 act additively to increase transcriptional activity. To examine differential transcription factor binding, we performed electrophoretic mobility shift assays. Supershifts identified allele-specific USF-1 binding to rs2281721 and FOXO3 binding to rs6143660. We also observed differential C/EBPβ binding to rs4846913, which we confirmed by allele-specific ChIP assays in cell lines of differing genotypes. These data show evidence that this HDL-C GWAS association signal is driven by at least three intronic variants in *GALNT2*. Open chromatin, histone modification, and TF ChIP-seq data aided in the detection of, but did not perfectly predict, the multiple functional regulatory variants.

## 353

**A PAX1 Enhancer Locus Increases Risk of Idiopathic Scoliosis in Females.** *C. Wise[1,2,3,4], S. Sharma[1,13], D. Londono[5], W. Eckalbar[6,12], X. Gao[1], I. Kou[7], A. Takahashi[7,8], M. Matsumoto[9], J.A. Herring[2,10], D.K. Burns[11], S. Ikegawa[7,8], N. Ahituv[6,12], D. Gordon[5].* 1) Seay Center for Musculoskeletal Research, Texas Scottish Rite Hosp, Dallas, TX; 2) Department of Orthopaedic Surgery, University of Texas Southwestern Medical Center at Dallas, Dallas, TX; 3) Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center at Dallas, Dallas, TX; 4) Department of Pediatrics, University of Texas Southwestern Medical Center at Dallas, Dallas, TX; 5) Department of Genetics and Human Genetics Institute, Rutgers University, Piscataway, NJ; 6) Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA; 7) Laboratory for Statistical Analysis, Center for Genomic Medicine, RIKEN, Tokyo, Japan; 8) Laboratory of Bone and Joint Diseases, Center for Genomic Medicine, RIKEN, Tokyo, Japan; 9) Department of Orthopaedic Surgery, School of Medicine, Keio University, Tokyo, Japan; 10) Department of Orthopaedics, Texas Scottish Rite Hospital for Children, Dallas, TX; 11) Department of Pathology, University of Texas Southwestern Medical Center at Dallas, Dallas, TX; 12) Institute for Human Genetics, University of California San Francisco, San Francisco, CA; 13) School of Biotechnology, Shri Mata Vaishno Devi University, Katra, India.

Idiopathic scoliosis (IS) is the most common pediatric spinal disorder, affecting more than thirty million children worldwide. IS exhibits a striking ten-fold greater risk of progressive disease in females for reasons that are unknown. Although IS is highly heritable, few replicated risk loci have been identified. To discover and characterize new IS genetic risk factors, we performed a two-stage GWAS that combined a prior family-based GWAS with a new case-control study (N= 3,102 individuals total). Our analyses identified an association with 20p11 SNPs (combined $P=1.33\times10^{-8}$) clustering ~186 kb distal to the 5' end of the PAX1 homeobox gene (OMIM #167411). Surprisingly, several of the 20p11 SNPs were previously associated with protection against early-onset male pattern baldness, suggesting that the region participates in sexually dimorphic gene expression. Accordingly, stratification by sex yielded IS association in females but not males for both GWAS stages (region combined $P=6.89\times10^{-9}$ in females and P= 0.71 in males). We replicated the IS association with top SNP rs6137473 in an independent NHW female cohort (OR=1.67, $P = 2.4\times10^{-4}$), and in a Japanese female cohort (OR = 1.19, $P =3.7\times10^{-3}$). Combined female results across all four studies yielded $P=2.15\times10^{-10}$, OR=1.30 for this marker. Spontaneous mouse pax1 deletion mutants develop scoliosis and other spinal anomalies. Consequently we hypothesized that the Chr 20p11 IS susceptibility locus functions through cis-acting regulation of PAX1 expression. To identify potential enhancers in the region we selected highly conserved sequences and tested activity in zebrafish transgene assays, revealing two candidate enhancers. One enhancer, SE7, drove expression in somitic muscle and spinal cord and harbored top disease-associated SNP alleles that are predicted to disrupt transcription factor binding. Re-sequencing SE7 in IS cases revealed a common haplotype that when tested in zebrafish abolished the enhancer activity of this sequence. In addition, immunohistochemistry in mouse spinal tissues revealed most persistent Pax1 protein expression in myogenic cells beginning at E16.5 to day P84. This study is the first to yield insights into the well-known and puzzling sex bias in IS. Our data suggest that Chr 20p11 variants increase susceptibility to IS by altering cis-acting regulatory sequences that may participate in sexually dimorphic PAX1 expression in spinal muscle and/or nerve cells.

## 354

**Characterization of the Type 2 Diabetes associated *KLF14* trans-regulatory network.** *K. Small[1], L. Quaye[1], A. Hough[2], M. Todorcevic[3], A. Mahajan[4], M. Horikoshi[4], A. Buil[5], A. Vinuela[1], C. Glastonbury[1], A. Brown[6], J. Bell[1], A. Gloyn[3], R. Cox[2], F. Karpe[3], M. McCarthy[3,4].* 1) Department of Twin Research and Genetic Epidemiology, King's College London, London, UK; 2) MRC Harwell, Oxford, UK,; 3) Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford, UK; 4) Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK; 5) Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; 6) Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK.

The region upstream of the maternally expressed transcription factor *KLF14* is associated with Type 2 Diabetes (T2D) and HDL-cholesterol. The T2D association at *KLF14* is complex - it is maternal-specific and shows evidence of female-specificity, which is supported by female-specific lipid associations to this region. We have previously shown that the T2D and HDL-associated variants at *KLF14* mediate a maternal-specific *trans*-regulatory network in adipose tissue via *cis*-regulation of *KLF14*. We have extended our initial results from microarrays to multi-tissue RNAseq data from ~800 female twins from the TwinsUK cohort and identified an expanded network of 140 genes (FDR 5%) associated in *trans* to *KLF14*. *KLF14 cis*-regulation and the *trans*-regulatory network are adipose-specific: replicating in two adipose cohorts but not present in other key insulin action tissues such as muscle and liver, nor in skin, whole blood or brain tissues. In addition to regulation of mean expression levels, the *KLF14* variants show evidence of *trans*-association with variance in expression levels (vQTL). There is a genome-wide enrichment for *trans*-vQTL ($\pi1$ =0.13) in adipose, but no evidence of *cis*-vQTL with *KLF14* expression itself (p= 0.3). We are investigating if the *trans*-vQTL effects are driven by gene-gene interaction between the *KLF14* variants and SNPs in *KLF14* binding sites of downstream genes. The *KLF14* variants are associated with methylation of an Illumina HumanMethylation450 probe ~3KB upstream of *KLF14* in adipose tissue (N = 603, p = 2.2e-7) but not whole blood (N = 217, p = 0.69) in the same cohort, indicating the *cis*-regulatory effect may be mediated by an adipose-specific epigenetic mechanism. The expanded set of *trans*-regulated genes show no enrichment for known pathways or biological processes, but are highly correlated with concurrently measured metabolic traits and enriched for variants directly associated to metabolic traits in large GWAS studies, including insulin levels, lipids and T2D. These associations highlight a subset of the trans-network that can independently influence disease and provide a novel biological link between disparate metabolic GWAS loci. Metabolic phenotyping of a *KLF14* knockout mouse is ongoing, together with cellular phenotyping of adipocytes these studies will help inform how *KLF14*-dependent transcription events in adipose tissue play a critical role in the development of insulin resistance and predisposition to T2D.

## 355

**mRNA-seq of 278 diverse skeletal muscle biopsies reveals mechanistic insights about type 2 diabetes genetic risk and identifies disease state specific eQTLs.** *J.R. Huyghe[1], S.C.J. Parker[2], M.R. Erdos[2], H. Koistinen[3], P.S. Chines[2], R. Welch[1], X. Wen[1], H. Jiang[1], N. Narisu[2], L. Taylor[2], B. Wolford[2], L.J. Scott[1], H. Stringham[1], L. Kinnunen[3], T. Blackwell[1], A.U. Jackson[1], Y. Lee[1], A.J. Swift[2], L. Bonnycastle[2], M.L. Stitzel[4], R.M. Watanabe[5,6], K. Mohlke[7], T. Lakka[8], M. Laakso[8], J. Tuomilehto[3], F.S. Collins[2], M. Boehnke[1].* 1) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA; 2) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; 3) National Institute for Health and Wellfare, Helsinki, Finland; 4) The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA; 5) Department of Preventive Medicine, University of Southern California (USC) Keck School of Medicine, Los Angeles, CA, USA; 6) Department of Physiology and Biophysics, Keck School of Medicine of USC, Los Angeles, CA, USA; 7) Department of Genetics, University of North Carolina, Chapel Hill, NC, CA, USA; 8) University of Eastern Finland, Kuopio, Finland.

Type 2 diabetes (T2D) is a complex disease caused by an interplay between genes, environment, and behavioral factors, acting over time and across multiple tissues. Genome-wide association studies (GWAS) have identified >80 loci associated with T2D risk. For most identified loci, the causal gene and functional variant(s) remain elusive because the associated region resides in noncoding DNA, suggesting a major contribution of transcriptional regulatory elements to disease risk. Regulatory element usage is often tissue-specific. Therefore, a crucial next step to guide the functional follow-up of GWAS is to determine the relationship between single nucleotide polymorphisms (SNPs) associated with T2D or related traits, and gene expression in disease-relevant tissues and across disease progression. As part of the Finland-United States Investigation of NIDDM Genetics (FUSION) study, we obtained *vastus lateralis* skeletal muscle biopsies from 278 clinically well-characterized Finns with normal and impaired glucose tolerance, and with newly diagnosed T2D without antihyperglycemic medication. Skeletal muscle is a major insulin target tissue and accounts for ~25-30% of postprandial glucose uptake. We performed dense genotyping, phasing, and imputation, constructed strand-specific mRNA-seq libraries, and sequenced 15.3 billion fragments (mean depth 55 million 101 base read pairs). We identified >8000 genes with expression and/or splicing quantitative trait loci (e/sQTL) (5% FDR). Some of these eQTL (e.g., for the genes *TTNT3* and *SDCCAG8*) appear disease state specific. Multiple eQTL in our catalog are in high linkage disequilibrium with GWAS SNPs for T2D or related traits, highlighting genes at these loci as probable candidates for a role in T2D risk. Interestingly, for a subset of these GWAS SNP overlapping eQTL, gene expression is also significantly associated with T2D or a glycemic trait. E.g., T2D GWAS index SNP rs516946 is the most significant eQTL SNP for the *ANK1* gene, which is differentially expressed between normal glucose tolerant vs. T2D individuals. Similarly, our eQTL analysis points out *CCHCR1*, associated to glycemic traits and BMI, as a candidate gene for the T2D GWAS index SNP rs3130501. This rich data resource enables identification of diverse molecular processes involved in skeletal-muscle-based insulin resistance and changes in gene transcription with progression towards T2D, and reveals mechanistic insights about T2D risk.

## 356

**Genetic analyses of hepatic steatosis GWAS associated loci.** *E.K. Speliotes, Y. Chen, A.W. Tai.* University of Michigan, Ann Arbor, MI.

Nonalcoholic Fatty Liver Disease (NAFLD) is a heritable and prevalent disease, affecting about 30% of the population. A characteristic feature of NAFLD is hepatic steatosis, the presence of excess fat (mostly triglycerides (TG)) in the liver. Using genome wide association analysis (GWAS) we identified genetic variants in PNPLA3 and GCKR, and near LYPLAL1 that associate with population based hepatic steatosis. How these variants result in increased liver steatosis is not known. Here we aim to characterize the genetic mechanism by which genetic variants at these loci may affect nearby genes to result in hepatic triglyceride accumulation. HuH-7 and HepG2 liver cell lines were infected with lentiviruses expressing wildtype PNPLA3, GCKR, and LYPLAL1 as well as the mutants PPP1R3B(I148M) and GCKR(P446L) or with shRNAs to PNPLA3, GCKR, and LYPLAL1 and stably expressing cell lines were selected. Overexpression/knockdown was quantified using Western/Northern blotting analysis. Stable cell lines were loaded with oleic acid, hepatic steatosis was measured using LipidTOXTM (Life Technology), and total cellular triglyceride was quantified using a Triglyceride Determination Kit (Sigma-Aldrich). Overexpression of wildtype and to a larger extent mutant PNPLA3 but not knockdown of PNPLA3 resulted in increased steatosis/TG accumulation. Overexpression of wildtype GCKR but not mutant GCKR resulted in increased steatosis/TG accumulation whereas knockdown of GCKR also resulted in increased steatosis/TG accumulation. Overexpression of LYPLAL1 resulted in decreased steatosis/TG accumulation and knockdown in increased steatosis/TG accumulation. These results suggest that variants in PNPLA3 exert their effect through an increase/gain-of-function mechanisms, those in GCKR and near LYPLAL1 likely through a loss-of-function mechanism.

## 357

**FOXO3 Regulates Fetal Hemoglobin Levels in Sickle Cell Anemia.** *V. Sheehan[1], Y. Zhang[1], J. Crosby[2,3], R. Ware[5], E. Boerwinkle[3,4].* 1) Pediatrics, Baylor College of Medicine, Houston, TX; 2) The University of Texas Graduate School of Biomedical Sciences at Houston; Department of Biostatistics, Bioinformatics, and Systems Biology, University of Texas, Houston, TX; 3) Human Genetics Center, University of Texas, Houston, TX; 4) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 5) Division of Hematology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH.

Although they ostensibly have a monogenetic disease, individuals with sickle cell anemia (SCA) exhibit wide variability in their degree of clinical severity. One of the most powerful and reproducible predictors of disease severity is the level of endogenous fetal hemoglobin (HbF), composed of two $\gamma$-globin and two $\alpha$-globin chains. We have used whole exome sequencing (WES) to find new genetic variants associated with baseline HbF in SCA, using 171 pediatric SCA genomes from participants in two clinical trials, HUSTLE (NCT NCT00305175) and SWiTCH (NCT 00122980). WES allows identification of both common and rare exonic variants; in order to capture the association between the phenotype variants with a minor allele frequency (MAF) below 1%, burden tests maximize power by grouping low frequency variants together by gene. Burden analysis (T1), found seven unique non-synonymous variations in a Forkhead box O transcription factor, FOXO3, to be significantly associated with lower HbF ($p=5.6\times10^{-4}$, $\beta$-value ln HbF -0.66). All variants produced the same effect, a lowering of HbF. HbF values were normalized using natural log transformation to permit analysis, and adjusted for age, *BCL11A*, and *XmnI* variant status. Each individual was heterozygous for a variant. In order to verify the association between FOXO3 and endogenous HbF levels, we performed functional studies in K562 cells, an erythroid leukemia cell line that expresses $\gamma$-globin. We knocked down *FOXO3* in K562 cells using silencing RNA (siRNA). RT-qPCR and Western blot analysis detected a substantial decrease in $\gamma$-globin expression with *FOXO3* knockdown. This knockdown of *FOXO3* recapitulated the patient phenotype of lower HbF levels with the inheritance of a *FOXO3* mutation. We then examined the effect of overexpression of FOXO3 on $\gamma$-globin expression. K562 cells were transfected with expression plasmids encoding wild-type FOXO3 (FOXO3-WT) or a plasmid containing a constitutively active FOXO3 with three serine residues altered, so it could not be phosphorylated and sent to the cytoplasm where it is inactive (FOXO3-TM). A significant increase in $\gamma$-globin expression was observed in K562 cells transfected with FOXO3-WT. FOXO3-TM overexpressing cell lines had an even greater increase in $\gamma$-globin expression compared to FOXO3-WT. Our results indicate that FOXO3 participates in regulating $\gamma$-globin gene expression in K562 cells, and corroborate the association found through WES of sickle cell patient samples.

## 358

**Susceptibility to tuberculosis is associated with the ASAP1 gene that regulates dendritic cell migration.** *Y. Luo[1], J. Curtis[2], H.L. Zenner[2], D. Cuchet-Lourenco[2], C. Wu[2], K. Lo[3], M. Maes[2], A. Alisaac[2], E. Stebbings[2], J.Z. Liu[1], O. Ignatyeva[4,5], Y. Balabanova[4,5], V. Nikolayevskyy[5], P. Nürnberg[6], R. Horstmann[7], F. Drobniewski[4], V. Plagnol[3], J.C. Barrett[1], S. Nejentsev[2].* 1) Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; 2) Department of Medicine, University of Cambridge, Cambridge, UK; 3) UCL Genetics Institute, UCL, London, UK; 4) Samara Oblast Tuberculosis Dispensary, Samara, Russia; 5) Health Protection Agency National Mycobacterium Reference Laboratory and Clinical TB and HIV Group, Institute for Cell and Molecular Sciences, Barts and the London School of Medicine, Queen Mary, University of London, London, UK; 6) Cologne Center for Genomics, University of Cologne, Cologne, Germany; 7) Department of Molecular Medicine, Bernhard Nocht Institute for Topical Medicine, Hamburg, Germany.

Tuberculosis (TB), caused by the pathogen *Mycobacterium tuberculosis*, remains a major threat to public health, with two billion people estimated to be infected with the pathogen worldwide. Despite its high infection rate, only ~10% of infected individuals eventually develop active TB. Early case observations, twin studies and mouse models indicate that host genetic factors are important in determining susceptibilities to *M. tuberculosis*. Recent Genome-wide association studies (GWAS) have identified two loci on chromosomes 11p13 and 18q11 in two African populations. However, pathophysiological mechanisms underpinning these associations, as well as their role in other TB endemic regions, remain unknown. In this study, we conducted the largest TB GWAS to date of 5,914 patients with active pulmonary TB and 6,022 healthy adult controls from Russia using Affymetrix Genome-Wide Human SNP array 6.0, and then imputed genotypes at ~7.6 million SNPs using the 1000 Genome reference panel. Our study identified a new TB-associated locus on chromosome 8q24, where 11 SNPs located in introns of the ASAP1 gene reached the genome-wide significant level ($P < 5 \times 10^{-8}$). Among the most significantly associated SNPs, 7 were selected for replication in the combined cohort of 15,087 subjects and showed strong association with TB ($P = 2.6 \times 10^{-11}$). We also see association in one of the two previously reported TB GWAS loci at 11p13 (*rs2057178*, P = 0.00068), but did not detect any signal at the 18q11 locus, suggesting that it may be population-specific. To advance our understanding of the pathophysiological mechanisms underpinning the ASAP1 associations, we further studied the expression of ASAP1 in dendritic cells (DCs). DCs are essential in the formation and maintenance of granulomas -- structures that are characteristic of human TB. DCs had high ASAP1 expression, which was reduced after *M. tuberculosis* infection. We found that the TB-associated SNP was also associated with the level of reduction of ASAP1 expression after *M. tuberculosis* infection ($P = 4.6 \times 10^{-3}$). After knockdown of ASAP1 expression, DCs showed impaired matrix degradation and migration. Therefore, genetically determined excessive reduction of ASAP1 expression in *M. tuberculosis*-infected DCs may lead to their impaired migration and predispose to TB. Our study highlights a novel functional mechanism, indicating that the ASAP1-mediated pathways are involved in mycobacterial infection and TB pathogenesis.

## 359

**Targeting Calpains: A Therapeutic Strategy for the Treatment of TGFβ Mediated Mesenchymal Transition and Associated Pathologies.** *D. Kim[1], R. Gould[1,2], J. Butcher[2], H. Dietz[1].* 1) Inst Gen Med, Johns Hopkins Sch Med, Baltimore, MD; 2) Department of Biomedical Engineering, Cornell University, Ithaca, NY.

Expression profiling studies associate enhanced expression/activity of the calpain family of cysteine proteases with multiple genetic or environmentally-induced TGFβ-related disease processes including fibrosis and tumor metastasis. The underlying mechanistic connection (if any) remains unknown. We reasoned that this association might relate to TGFβ-induced mesenchymal transition, the process by which cells of epithelial or endothelial origin lose polarity and cell adhesion and adopt an invasive character and fibrotic synthetic repertoire (EMT). This hypothesis was tested in NMuMG epithelial cells which show striking EMT within 2 days of TGFβ administration, as evidenced by downregulation of E-cadherin, transition from a cortical to a stress fiber distribution of F-actin, and upregulation of α-smooth muscle actin, collagen, and matrix metalloproteinases. As expected, concomitant treatment with a broad inhibitor of TGFβ signaling (SB431542) prevented EMT in association with attenuation of intracellular TGFβ signal propagation (phosphorylation of Smad2/3). In contrast, inhibition of calpain activity (as evidenced by failed cleavage of the natural calpain substrate FLNA) with the broad-spectrum calpain inhibitors MDL-28170 or calpeptin abrogated EMT despite maintenance of the pSmad2/3 response. Furthermore, robust EMT inhibition was achieved using either 2-ABP (which is reported to prevent the TRPM7-mediated calcium influx needed for calpain activation) or overexpression of calpastatin, a naturally-occurring and highly specific dimeric calpain inhibitor. Among dimeric calpains, the CAPN1 and CAPN2 large subunits and CAPNS4 small subunit show broad expression, however siRNA-mediated silencing had no effect on EMT. In contrast, we show that the relatively obscure CAPN9 and CAPNS2 subunits only show physiologic expression in the GI tract and skin, respectively, but are potently induced by TGFβ in both epithelial and endothelial cells; siRNA-mediated silencing of either abrogated EMT in culture systems. Identical provocations prevented EMT in all epithelial and endothelial cell lines tested and showed the capacity to reverse an established mesenchymal phenotype (MET). Taken together, these data suggest that calpain inhibition is an attractive therapeutic strategy for multiple TGFβ pathologies and lend optimism that CAPN9/S2 inhibition will have a greater influence on pathologic vs. physiologic events and will therefore exhibit a favorable tolerance profile.

## 360

**Metabolic regulation by the MeCP2/HDAC3 transcriptional corepressor complex points to new therapeutic targets in Rett syndrome.** *S.M. Kyle[1,2], C.M. Buchovecky[1], M.J. Justice[2].* 1) Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario.

Metabolic dysregulation can lead to downstream pathogenesis in nearly all tissues and organ systems. In recent decades, a large body of data has implicated metabolic perturbations in neurological development and degeneration. In particular, dysregulation of cholesterol trafficking and biosynthesis are responsible for the onset of Neimann-Pick type C and Smith-Lemli Opitz syndrome, respectively. Furthermore, Fragile X syndrome, Alzheimer, Parkinson, and Huntington diseases have all been linked to aberrant cholesterol homeostasis. Rett syndrome (RTT) is a progressive neurodevelopmental disorder of females primarily caused by mutations in the X-linked gene encoding methyl-CpG binding protein 2 (*MECP2*). To identify pathways in disease pathology for therapeutic intervention, we carried out a dominant random mutagenesis suppressor screen in *Mecp2* null mice. One suppressor identifies a stop codon mutation in a rate-limiting enzyme in cholesterol biosynthesis, which ameliorates RTT-like symptoms and increases longevity in *Mecp2* null mice by altering brain cholesterol homeostasis. Although RTT has been classically labeled a neurological disorder, these studies suggest that a metabolic component contributes to pathology. Here we show that the *Mecp2* mutation induces metabolic defects in mice including fatty liver, increased lipolysis, and insulin resistance in muscle and adipose. These metabolic phenotypes are strikingly similar to that in mice with a liver-specific knockout of histone deacetylase 3 (*Hdac3*), a potent regulator of lipogenesis and cholesterol biosynthesis. Consistently, we show that MeCP2 and HDAC3 work in a complex to suppress expression of the rate-limiting cholesterol enzyme identified in our screen. Our data suggest a novel metabolic component present in RTT arising from loss of interaction between MeCP2 and HDAC3. These studies inform highly targetable therapeutic pathways relevant to treating RTT; remarkably, statin drug administration improves motor symptoms and confers increased longevity in *Mecp2* null mice. The suppressor mutation also suggests that symptoms may be modified in patients by mutations in genes that affect metabolism. In support of this idea, a subset of RTT patients has increased serum cholesterol and triglycerides, independent of body mass index. Our ongoing studies point to additional metabolic pathways that are prime targets in the pursuit of preventing morbidities associated with Rett syndrome.

## 361

**Increasing IKAP expression by mRNA splicing modification improves phenotype in a mouse model of Familial Dysautonomia.** *E. Morini[1], P. Dietrich[2], M. Salani[1], F. Urbina[1], M. Nilbratt[1], I. Dragatsis[2], S. Slaugenhaupt[1].* 1) Center for Human Genetic Research, Massachusetts General Hospital/Harvard Medical School, Boston, MA; 2) Department of Physiology, The University of Tennessee Health Science Center.

Familial dysautonomia (FD) is a recessive neurodegenerative disease caused by a splice mutation in the IKBKAP gene which leads to variable skipping of exon 20. We found that kinetin can correct the IKBKAP splicing defect and increase the amount of normal mRNA and protein in FD cell lines. We have also shown that kinetin can increase the level of functional IKAP protein in mice following oral dosing in all tissues tested, including brain. Despite these remarkable advances we lacked an animal model in which to test the effect of increasing IKAP protein on the FD phenotype. In order to create a phenotypic model of FD in which we could also manipulate mRNA splicing, we introduced an FD transgene (TgFD9), which contains the human IKBKAP gene with the major FD splice mutation, into the Ikbkap$^{delta20/flox}$ mouse model by sequential mating. The introduction of the human IKBKAP transgene attenuates the severe FD phenotype that we observed in the Ikbkap$^{delta20/flox}$ mouse and recreates the same tissue-specific mis-splicing defect seen in FD patients. Characterization of this new mouse model, FD9/ Ikbkap$^{delta20/flox}$, recapitulates several phenotypic features observed in FD patients, including reduced growth rate, reduction in the number of fungiform papillae on the tongue, spinal abnormalities, and reduction in the volume of the sympathetic stellate and dorsal root ganglia. Our results demonstrate that the new TgFD9/ Ikbkap$^{delta20/flox}$ mouse accurately models both the disease phenotype and the tissue-specific mRNA mis-splicing defect seen in FD patients. The creation of this new model has allowed us to initiate a preclinical trial of kinetin and will permit testing of other strategies aimed at either targeting mRNA splicing or increasing expression of IKBKAP. We have initiated in parallel three different preclinical trials to evaluate the efficacy of kinetin: 1) administration of kinetin to dams prior to mating and throughout gestation; 2) administration of kinetin to dams immediately following birth and pups weaned onto kinetin chow; 3) administration of kinetin to mice beginning at weaning age. Our preliminary results show that kinetin leads to significantly improved IKBKAP splicing and has beneficial effects on the overall growth of FD embryos and adult mice. The completion of this study will allow us to fully evaluate the spectrum of benefits that modification of IKAP protein levels by kinetin and other drugs and/or mechanisms may offer FD patients.

## 362

**Impact of Early Hormonal Therapy (EHT) on the Neurobehavioral Profile of Boys with 47, XXY (Klinefelter Syndrome) at 9 Years of Age.** *C. Samango-Sprouse[1,2,3,4], D.C. Gibbs[3], E. Stapelton[3], T. Sadeghin[1], A.L. Gropman[2,4].* 1) Neurodevelopmental Diagnostic Center for Young Children, Davidsonville, MD; 2) George Washington University School of Medicine, Washington, DC; 3) The Focus Foundation, Davidsonville, MD; 4) Children's National Medical Center, Washington, DC.

47, XXY is associated with frontal lobe dysfunction and language-based learning difficulties contributing to a complex behavioral phenotype that may include ADHD and atypical social skills. Recent studies have shown the positive effects of early hormonal treatment (EHT) on the neurodevelopmental outcome of boys with 47, XXY in early childhood, but the effects of EHT on behavioral and social development have not been explored at later ages. 59 prenatally diagnosed 47, XXY boys [22 who received EHT (3 injections of 25 mg testosterone enanthate) and 37 who received no treatment] were evaluated at 9 years of age using the Behavior Rating Inventory of Executive Function (BRIEF), Social Responsiveness Scale (SRS-2) and Child Behavior Checklist (CBCL). Significant differences between group scores were tested using appropriate biostatistics. The EHT treatment group had significantly improved Global Executive Functioning (p=0.038), Monitoring (p=0.027) and Initiation (p=0.0023) on the BRIEF, significantly fewer Aggressive Behaviors (p=0.039) Affective Problems (p=0.002) and Total Problems (p=0.006) on the CBCL and improved social cognition (p=0.0045), social communication (p=0.0175) and fewer autistic features (p=0.0005) on the SRS-2. These results provide further evidence of the sustained and positive effects of EHT on the neurodevelopmental outcome and phenotypic presentation of boys with 47, XXY. The significant improvements in social and behavioral skills and executive functioning after EHT presented in this study support the need for continued research and earlier biological treatment interventions for 47, XXY boys.

## 363

**A Causative Role for Oxytocin in Pregnancy-Induced Aortic Dissection in Marfan Syndrome Mouse Models.** *J.P. Habashi[1], E.M. Gallo[2], N. Huso[2], Y. Chen[2], D. Bedja[2], D. Huso[3], H.C. Dietz[2,4].* 1) Pediatrics, Johns Hopkins University, Baltimore, MD; 2) Genetics, Johns Hopkins University, Baltimore, MD; 3) Comparative Medicine, Johns Hopkins University, Baltimore, MD; 4) Howard Hughes Medical Institute.

Marfan syndrome (MFS) is an autosomal dominant connective tissue disorder caused by mutations in fibrillin-1 (FBN1) that includes a strong predisposition for aortic aneurysm and dissection. Studies in both mouse models and people with MFS suggest that excessive TGFβ signaling in the aortic wall contributes to disease pathogenesis, with particular relevance of the noncanonical activation of extracellular signal-regulated kinase (ERK). Both aortic aneurysm progression and aberrant ERK activation are abrogated with TGFβ antagonists including the angiotensin II type 1 receptor blocker (ARB) losartan. Pregnant women with MFS show a high risk of aortic dissection in the immediate peripartum period. We hypothesized a mechanistic role for oxytocin, since it peaks at the end of pregnancy, is sustained during lactation and stimulates peripheral tissues through ERK activation. Using the mgR/mgR mouse model of MFS, we demonstrated a 90% incidence of dissection in the 4 weeks following delivery. Simply removing the mothers from their pups on the day of delivery, thereby eliminating lactation-induced oxytocin release, decreased the incidence of death from 90% to 26%. This was associated with a significant decrease in ascending aortic growth over the 7 week period of pregnancy and lactation (1.40 ± 0.15 vs 0.69 ± 0.36 mm/7weeks; respectively, $p<0.01$). Treatment of pregnant mgR/mgR mice with a continuous infusion of a highly specific oxytocin receptor antagonist (ORA; desGly-NH2-d(CH2)5[D-Tyr2,Thr4]OVT), beginning in the 3rd trimester, reduced the incidence of pregnancy-associated aortic dissection from 90% to 6.7%, despite a sustained ability of treated animals to deliver spontaneously and to nurse. Ascending aortic growth was significantly reduced with ORA treatment as compared to placebo-treated pregnant mgR/mgR mice (0.81 ± 0.54 vs. 1.36 ± 0.40 mm/7wks, respectively; $p<0.01$). The risk of aortic dissection is directly proportional to the level of phosphorylated ERK1/2 in the aortic wall, with elevated levels in MFS mice compared to wild-type littermates, a substantial further elevation in the peripartum period, and normalization of levels through either pup removal or treatment with the oxytocin receptor antagonist ($p<0.05$). This therapeutic strategy has the strong potential to modify vascular risk in woman with MFS and perhaps other heritable connective tissue disorders including Loeys-Dietz syndrome and vascular Ehlers-Danlos syndrome.

## 364

**Most participants in the agalsidase beta phase 3 clinical trial in patients with classic Fabry disease experienced no severe clinical events during a 10-year follow-up period.** *D.P. Germain[1], J. Charrow[2], R.J. Desnick[3], J.T. Ebels[4], N. Guffon[5], J. Kempf[4], R.H. Lachmann[6], R. Lemay[4], G.E. Linthorst[7], S. Packman[8], C.R. Scott[9], S. Waldek[10], D.G. Warnock[11], N.J. Weinreb[12], W.R. Wilcox[13].* 1) Division of Medical Genetics, University of Versailles - St Quentin en Yvelines, Versailles, France; Assistance Publique - Hôpitaux de Paris, Garches, France; 2) Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; 3) Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 4) Genzyme, a Sanofi company, Cambridge, MA, USA; 5) Centre de Référence des Maladies Héréditaires du Métabolisme, Hospices Civils de Lyon, Hôpital Femme-Mère-Enfant, Bron, France; 6) Charles Dent Metabolic Unit, National Hospital for Neurology and Neurosurgery, London, United Kingdom; 7) Department of Endocrinology and Metabolism, Academic Medical Center, Amsterdam, Netherlands; 8) Division of Medical Genetics, Department of Pediatrics, University of California San Francisco, CA, USA; 9) Department of Pediatrics, University of Washington, Seattle, WA, USA; 10) Salford Royal NHS Foundation Trust (retired Oct. 2011); 11) Division of Nephrology, University of Alabama at Birmingham, Birmingham, AL, USA; 12) University Research Foundation for Lysosomal Storage Diseases, Coral Springs, FL, USA; 13) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA.

Fabry disease (OMIM 301500) is an X-linked disorder caused by deficiency of the lysosomal enzyme α-galactosidase A that results in accumulation of globotriaosylceramide (GL-3) and other glycosphingolipids in cells throughout the body. Signs and symptoms can begin in early childhood and lead to progressive life-threatening renal, cardiac, and cerebrovascular complications. In this study, the outcomes of 52 of 58 classic Fabry disease patients enrolled in the 1999 randomized, placebo-controlled, double-blind phase 3 agalsidase beta clinical trial were analyzed. Data from the phase 3 clinical trial, the phase 3 extension study (NCT00074971), and the Fabry Registry (NCT00196742), all sponsored by Genzyme, a Sanofi company, were used. During a median of 10 years of treatment, 81% of patients (42/52) did not experience any severe clinical event and 94% (49/52) were alive. Severe clinical events were defined as dialysis, kidney transplant, atrial fibrillation, congestive heart failure, myocardial infarction, major cardiac procedures, stroke, and death. Patients with a urine protein-to-creatinine ratio (UPCR) ≤0.5 g/g and <50% sclerotic glomeruli at baseline were classified as low renal involvement (LRI, n=32); patients with UPCR >0.5 g/g or ≥50% sclerotic glomeruli at baseline were classified as high renal involvement (HRI, n=20). The mean slope for estimated glomerular filtration rate (eGFR) for LRI was -1.88 ml/min/1.73 $m^2$/year; the mean slope for eGFR for HRI was -6.82 ml/min/1.73 $m^2$/year. In addition, the mean slopes for eGFR for patients who either maintained (n=20) or did not maintain UPCR values ≤0.5 g/g (n=12) within the LRI group throughout the treatment period were -1.48 and -2.6 mL/min/1.73$m^2$/year, respectively. There were no significant changes in mean inter-ventricular septum thickness and left posterior wall thickness slopes. Patients classified as LRI started therapy a mean of 13 years younger than HRI (25 vs. 38 years of age). Mean plasma GL-3 levels decreased to normal levels within 6 months of treatment and remained normal. This 10-year study documents the effectiveness of treatment with agalsidase beta (1 mg/kg/two weeks) in patients with classic Fabry disease. Most patients remained event-free (81%) and alive (94%) during the 10-year follow-up period of the phase 3 study. Patients who initiated treatment at a younger age and with less severe kidney involvement benefited the most from agalsidase beta therapy.

## 365

**ENGAGE: A phase 3, randomized, double blind, placebo controlled, multi center study to investigate the efficacy and safety of eliglustat in adults with Gaucher disease type 1: 18-month results.** *M. Balwani[1], D. Amato[2], M. Dasouki[3], G. Pastores[4], S. Packman[5], S. Assouline[6], P. Mistry[7], A. Ortega[8], S. Shankar[9], M. Solano[10], J. Angell[11], L. Ross[11], J. Peterschmitt[11].* 1) Mt. Sinai Hospital, New York, NY; 2) Mount Sinai Hospital, Toronto, Canada; 3) University of Kansas Hospital, Kansas City, KS, USA; 4) New York University School of Medicine, New York, NY, USA; 5) UCSF School of Medicine, San Francisco, CA, USA; 6) Jewish General Hospital, Montreal, Quebec, Canada; 7) Yale University School of Medicine, New Haven, Connecticut, USA; 8) OCA Hospital, Monterrey, Mexico; 9) Emory University, Atlanta, GA, USA; 10) Hospital San Jose, Bogota, Colombia; 11) Genzyme, a Sanofi company, Cambridge, MA, USA.

**Introduction**: Gaucher disease is an autosomal recessive disorder caused by deficient activity of the lysosomal enzyme acid β-glucosidase, resulting in progressive substrate accumulation and a spectrum of debilitating visceral, hematologic, and skeletal manifestations. Eliglustat, a ceramide analogue, is a novel oral substrate-reduction therapy in development for Gaucher disease type 1 (GD1). We present 18-month results from ENGAGE (NCT00891202), a randomized, double-blind, placebo-controlled, Phase-3 trial (sponsored by Genzyme, a Sanofi company) investigating the efficacy and safety of eliglustat in untreated adults with GD1.

**Methods**: Forty patients (mean age: 31.8 years; 20 males) with splenomegaly and thrombocytopenia and/or anemia were randomized 1:1 to receive eliglustat (50 or 100 mg BID depending on plasma levels) or placebo for 9 months and then entered a 9-month open-label extension phase in which all patients received eliglustat. The primary efficacy endpoint was percent change in spleen volume (multiples of normal). Other efficacy measures included hemoglobin, liver volume, and platelets. Results: In the 9-month primary analysis period, eliglustat was superior to placebo in all primary and secondary endpoints; no patients discontinued due to an adverse event. For 18/20 patients who have now received 18 months of eliglustat, mean improvements from baseline continue (spleen: -45%, hemoglobin: +1.02 g/dL; liver: -11%; platelets: +58%). For 20/20 patients previously receiving placebo, mean improvements after 9 months of eliglustat were consistent with what was seen in the primary analysis period in the eliglustat-treated patients: spleen: - 33%; hemoglobin: +0.79 g/dL; liver: -7.3%; platelets: +40%. No new safety concerns were identified. The adverse event profile for all patients after 18 months is similar to that of eliglustat patients in the primary analysis period and that of patients who switched from placebo to eliglustat after 9 months.

**Conclusion**: ENGAGE met its primary and secondary efficacy endpoints. Patients from both treatment arms have showed continued improvements in the first 9 months of the extension phase.

## 366

**Effective treatment of mitochondrial myopathy by nicotinamide riboside, a vitamin B3.** *N. Khan[1], M. Auranen[1], I. Paetau[1], E. Pirinen[2], L. Euro[1], C. Carroll[1], J. Auwerx[2], A. Suomalainen[1].* 1) Molecular Neurology, Biomedicum, Helsinki, FI., Finland; 2) Laboratory of Integrative Systems Physiology, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland.

Mitochondrial disorders are the most common group of inherited metabolic diseases. These disorders manifest with respiratory chain deficiency (RCD), but lead to a multitude of clinical manifestations, not explained purely by ATP-deficiency. Despite their progressive and often fatal outcome, no curative treatment is available. Therapy trials have been hampered by the absence of patient groups with homogenous genetic and clinical presentations. Therefore, development of mouse models, replicating the disease phenotype of patients, are essential for understanding the molecular basis of the metabolic consequences of the patients for enabling disease intervention. We have previously generated the mouse model with adult onset mitochondrial myopathy. Molecular data from patients and mice suggested a role for nutrient signaling in the pathogenesis, with disease-induced pseudostarvation response. True fasting increases NAD+:NADH ratio, which activates SIRT1 deacetylase, mitochondrial biogenesis, lipid oxidation and ATP-production. We hypothesized that RCD reduces NADH utilization, decreasing NAD+/NADH and signaling for high nutrient availability, leaving Sirt1 inactive and attenuating mitochondrial biogenesis. However, RCD also decreases ATP production, increasing AMP/ATP, signaling for low nutrition availability, and leading to a potential conflict in nutrient sensor activation and a partial pseudo-starvation response.We report that per-oral nicotinamide riboside (NR), a vitamin-B3-form and NAD+ precursor, boosts NAD+ levels and effectively delayed mouse MM progression. NR robustly induced mitochondrial biogenesis in muscle and brown fat, cured mitochondrial ultrastructure, and decreased mtDNA deletion load -hallmarks of MM. The MM-mice displayed a mitochondrial unfolded protein response (UPRmt), with induction of fasting-cytokine FGF21. NR further enhanced UPRmt, supporting a protective role of UPRmt upon MM. These results indicate that vitamin cofactors modify metabolism and that treatment strategies increasing NAD+ are warranted for MM.

## 367

**Transition from clinically fully validated panels to medically relevant exome.** *L. Wong, V.W Zhang, E.S. Schmitt, J. Wang.* Molecular and Human Genetics, Baylor College of Medecine.

Introduction: Whole exome next generation sequencing (NGS) technology has been widely applied to clinical diagnosis. However, the diagnostic yield is only about 25%. Fully validated gene panels focusing on specific diseases with consistently deep coverage for individual exons and the ability to detect copy number changes (CNVs) on the other hand provides much higher diagnostic yields. Methods: Thirty NGS based panels have been developed using SeqCap EZ capture for the enrichment of target genes followed by NGS on Illumina HiSeq2000. All coding exons and 20bp flanking intron regions were sequenced at an average depth of ~1000X and validated by Sanger sequencing. Any exons containing any insufficiently covered bases (<20X) are sequenced separately by PCR/Sanger method. Results: The diagnostic yields of pathway driven panels are usually high due to the indication of abnormal metabolites. These panels include cobalamin and related pathway, MSUD, fatty acid oxidation, congenital deficiency of glycosylation, and glycogen storage disease (GSD). Using GSD as an example, all exons in this panel are sufficiently covered at >20X. The diagnostic yield is about 64%, which is less than expected due to the phenotype overlap between GSD and other metabolic disorders. The clinically defined Usher syndrome panel contains 9 known large genes and has a high diagnostic yield of 92%. Analysis of 66 genes associated with nonsyndromic RP was able to make confirmatory diagnosis in about 84% of RP patients. On the contrary, analysis of 200 genes known to cause the most genetically and clinically heterogeneous mitochondrial disorders only reaches a diagnostic yield of about 25%. Further analysis indicates that many patients labeled with mitochondrial disorders harbor deleterious mutations in genes unrelated to mitochondrial structure, function, or energy metabolism. Conclusion: Our experience in NGS based target gene analysis suggests that unbiased capture and enrichment of the target genes with 100% consistently deep coverage of individual exons are essential to high diagnostic yields, which are achieved by the simultaneous detection of copy number variations (CNVs) using the same set of NGS data. With the clinical availability of NGS based panels of several other diseases, including hereditary cancer, severe combined immunodeficiency, Leigh disease, bone disorders, and all metabolic disorders, transition to a medically relevant exome is anticipated.

## 368

**How Well Do Whole Exome Sequencing Results Correlate with Clinical Findings? A Study of 89 Mayo Clinic Biobank samples.** *S. Middha[1], N.M. Lindor[2], S.K. McDonnell[1], K.J. Johnson[3], J.E. Olson[1], E.D. Wieben[4], G. Farrugia[3], J.R. Cerhan[1], S.N. Thibodeau[1,5].* 1) Department of Health Sciences Research, Mayo Clinic, Rochester, MN; 2) Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ; 3) Center for Individualized Medicine, Mayo Clinic, Rochester, MN; 4) Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, MN; 5) Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, MN.

Whole Exome Sequencing (WES) is increasingly used for diagnosis of known Mendelian disorders and for gene discovery. Yet there is inadequate information on the number and type of clinically reportable variants typically found from WES on any given individual and their correlation with clinical phenotype. WES was performed on 89 deceased individuals (mean age at death 74 years) from the Mayo Clinic Biobank. Significant clinical diagnoses were abstracted from each individual's electronic medical record (EMR) via chart review. Variants (SNV, INDEL) were filtered based on quality (accuracy>99%, read-depth>20, alternate-allele read-depth>5, allele-freq<0.1) and available HGMD/OMIM phenotype information. Variants were defined as Tier 1 (stop-gain, splice or frame-shifting) and Tier 2 (missense, predicted damaging by SIFT or PolyPhen). The number and type of variants reported were evaluated in a list of 56 ACMG-reportable genes and 58 cancer predisposition genes, as well as examining overall genotype-phenotype correlations. Following the filtering, we found a total of 6992 variants (79 per person, 623 Tier 1 and 6369 Tier 2). Of these, 161 variants (1.8 per person, 13 Tier 1 and 148 Tier 2) were found among the 56 ACMG-reportable genes, and 115 variants (1.3 per person, 3 Tier 1 and 112 Tier 2) were found among 58 cancer predisposition genes. The number and type of variants in the 58 cancer predisposition genes did not significantly differ between individuals with a history of invasive cancer and those without. For the broader genotype-phenotype correlation, only nine Tier 1 variants were found in autosomal-dominant (AD) inherited genes known to cause a phenotype that correlated with an observation in the individual's EMR. However potential phenotype from 115 additional Tier 1 variants identified in AD genes did not correlate with data in the individuals' EMR. We used WES to evaluate potential phenotypes from identified genetic variants and their correlation with the individuals' medical records. The list of genes with no evident phenotype correlation was more than 10 times longer than the list of genes with phenotype correlation to EMR. These data highlight challenges that need to be addressed including both phenotype issues (e.g., disease penetrance, uncertainty about what is clinically reportable) and sequencing issues (e.g., incomplete sequencing coverage, thresholds for data filtering, lack of high quality databases to determine functional annotation).

## 369

**Clinical Whole Exome Sequencing Reveals Contribution of Rare Genetic Events to Undiagnosed Disease.** *C.M. Eng[1], D. Muzny[2], F. Xia[1], Z. Niu[1], R. Person[1], Y. Ding[2], P. Ward[1], A. Braxton[1], M. Wang[2], C. Buhay[2], N. Veeraraghavan[2], A. Hawes[2], T. Chiang[2], M. Leduc[1], J. Beuten[1], J. Zhang[1], W. He[1], J. Scull[1], A. Willis[1], M. Landsverk[1], W. Craigen[1], M. Bekeirnia[1], P. Liu[1], S. Wen[1], W. Alcaraz[1], H. Cui[1], M. Walkiewicz[1], M. Bainbridge[2], E. Boerwinkle[2,3], A.L. Beaudet[1], J.R. Lupski[1,2], S.E. Plon[1,2], R.A. Gibbs[1,2], Y Yang[1].* 1) Dept Molec & Human Genetics, Baylor Col Medicine, Houston, TX; 2) Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3) University of Texas Health Sciences Branch, Houston, TX.

We developed and optimized technical, bioinformatic and interpretive whole exome sequencing (WES) pipelines in a CAP and CLIA certified lab to identify causative mutations underlying disease phenotypes in undiagnosed patients being evaluated clinically for genetic disorders. We previously reported (Yang, et al NEJM 2013;369:1502) implementation of a pilot program to provide WES on a clinical basis and a summary of the first 250 cases. The current study examines the subsequent 2000 patients referred for clinical WES yielding insight into the underlying disease mechanisms in patients with Mendelian disorders. Approximately 10 Gb of data were generated for each clinical sample with a mean coverage of 100X and >95% of the targeted bases covered at >20X. The mitochondrial genome was sequenced concurrently. Approximately 87% were pediatric patients with neurologic phenotypes. Diagnoses were based on mutation severity according to ACMG guidelines for variant interpretation, appropriate inheritance patterns, phenotypic fit and feedback from referring physicians. An overall molecular diagnosis was reported for 504 patients (25%). Patients with specific neurologic findings e.g., seizures had the highest diagnostic rate (36.1%). Mendelian disease patterns included: 280 (53.1%) autosomal dominant (of which at least 74% were shown to have arisen *de novo*), 181 (34.3%) autosomal recessive, 65 (12.3%) X-linked and 1 (0.2%) mitochondrial. Twenty-three patients (4.6% of those with diagnoses) had blended phenotypes resulting from two underlying single gene defects. In five patients, uniparental isodisomy (UPD) "unmasked" autosomal recessive mutations resulting in markedly different recurrence risks for these families. Of the solved cases, 30% harbored causative mutations in disease genes reported since 2011. Medically actionable incidental findings were reported in 95 subjects (4.8%), including 60 patients (3%) with variants in genes recommended for reporting by the ACMG. Referred cases without a diagnosis have the option of entering a research protocol at our institution potentially leading to novel gene discovery, such as that recently reported for AHDC1. WES is an important diagnostic tool that provides a molecular diagnosis for 25% of previously undiagnosed patients. Further examination of the results of this large cohort reveals data on the mechanisms of genetic disorders such as blended phenotypes, *de novo* mutations, and rare genetic events such as UPD.

## 370

**Clinical Exome Sequencing at UCLA: diagnosis rate, variant spectrum and novel gene discoveries.** *H. Lee[1], J.L. Deignan[1], N. Dorrani[3], S. Strom[1], N. Ghahramani[1], S. Kantarci[1], F. Quintero-Rivera[1], K. Das[1], M. Fox[3], W.W. Grody[1,2,3], E. Vilain[2,3], S.F. Nelson[1,2].* 1) Pathology and Laboratory Medicine, UCLA, Los Angeles, CA; 2) Human Genetics, UCLA, Los Angeles, CA; 3) Pediatrics, UCLA, Los Angeles, CA.

The CLIA-certified/CAP-accredited UCLA Clinical Genomics Center launched Clinical Exome Sequencing (CES) in 2012 to improve the genetic diagnosis of rare Mendelian disorders. CES has substantial clinical utility with a higher diagnostic yield than most other single-gene or panel-based molecular diagnostics. Here, we report on the first sequential 500 cases performed at UCLA to explore broad-based usage of CES and outcomes. The most frequent clinical indication for CES was developmental delay (DD) presenting as part of a complex syndrome including co-morbid diagnosis of seizures, hypotonia or dysmorphic features. Most causative genes were observed as mutated in only one pedigree. However, mutations in *KMT2A*, *ZEB2*, *DYRK1A* and *SCN2A* were determined to be causative in multiple independent cases, suggesting these genes are more commonly altered in individuals with syndromic DD. Pathogenic variants were identified and reported in 25% of all cases. Of these 126 diagnoses, 62 were based on heterozygous variants revealing an autosomal dominant disease (35 de novo, 3 inherited from an affected parent, 24 in proband-only cases), 22 were based on homozygosity of a recessive allele, 35 were based on compound heterozygosity, and 10 were based on X-linked recessive inheritance. The conclusive diagnosis rate was higher when CES was performed on a trio (30%, vs 21% for proband-only cases, $p$=0.046). New case reports with clear causal variants are serving to expand and clarify the full spectrum of phenotypes for some disease genes such as *KMT2A* (identified as the disease gene for Wiedemann-Steiner Syndrome only two years ago by whole exome sequencing). Gene discovery in rare Mendelian disorders is continuing at a fast pace and CES is well-suited to convert these new findings into meaningful diagnostic tests for patients. For instance, we observed two independent *de novo* variants in *TUBB2A*, recently demonstrated to be pathogenic for brain malformations, permitting specific molecular diagnosis in both cases. In 19 patients, novel variants were identified in genes that are not yet associated with any human disorder but have animal models available with significant phenotypic overlap with the case. These genes are now being investigated as novel disease genes. As more patients are having CES performed as the first-line diagnostic tool, we expect the diagnostic yield to increase, the phenotype spectrum of the known genes to expand, and the novel gene discoveries to continue.

## 371

**Medical Exome: Towards achieving complete coverage of disease related genes.** *A. Santani[2,3], K. McDonald[1], D. Mandelkar[2], A. Ankala[4], C. da Silva[4], Z. Yu[1], K. Cao[1], H. Sharma[2], R. Shakbatyan[2], M. Lebo[2], B. Funke[1], M. Hegde[4].* 1) Path & Lab Med, Philadelphia, PA; 2) Laboratory for Molecular Medicine, Partners HealthCare Personalized Medicine, Cambridge, MA; 3) Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Bosto; 4) Emory Genetics Laboratory, Emory University School of Medicine, Atlanta, GA 30047.

Next Generation Sequencing (NGS) is being rapidly adopted by clinical laboratories and has allowed the development of disease targeted gene panels and more recently exome and genome sequencing. One drawback of gene panels is their suboptimal clinical sensitivity, which is often only marginally increased by adding novel genes. In contrast, exome and genome sequencing (ES/GS) interrogates all genes, but is flawed by incomplete coverage and interrogates a significant number of genes with no clinical relevance. The Medical Exome Project was launched to bridge the gap between gene panels and ES/GS by curating the genes of clinical relevance and improving the technical performance of exome capture assays by enhancing their coverage. In addition, exons with conserved paralogous sequences were excluded to avoid technical and analytical challenges. Over 4,499 genes with known or suspected roles in disease were extracted from literature and public databases. This list has been made available to the community through public resources such as the ICCG consortium and resources maintained at NCBI. A pilot curation was performed to evaluate the evidence for their role in disease by applying a scoring grid from 0 (no evidence) to 3 (definitely associated with disease). Of over 500 genes evaluated, approximately 20% did not have any evidence. An enhanced exome capture assay was developed using Agilent's version 5 as a backbone and enhancing coverage of the 4,499 medically relevant genes. Intra- and inter laboratory clinical validation analyses of this enhanced exome (V5-PLUS) at a mean coverage of 100X indicated complete coverage of all exons was obtained for 3034 genes, whereas >95% coverage was obtained for the remaining 1465 genes. Most gene panels were covered to >99% and the rest of the exome was covered at >97%, which is an increase 5% over the Agilent V5. The performance was stable across 3 laboratories using 3 different data analysis pipelines. Alternate approaches are being employed to enhance the coverage of suboptimal genes. A collaborative, multi-center working group has been established to provide a detailed curation of the available literature to further evaluate the role of genes in disease. The availability of a curated medical exome will improve sensitivity and specificity of NGS based exome and genome sequencing and will provide a stepping stone for standardizing interpretation of genetic test results by clinical testing laboratories.

## 372

**Look before you leap, and list before you look: the use of *a priori* curated gene lists to guide exome analysis.** *B.C. Powell[1], A.K.M. Foreman[1], J.M. O'Daniel[1], K. Lee[1], L. Boshe[1], K.R. Crooks[2], M. Lu[2], Z. Fan[3], J.K. Booker[2], K.E. Weck[2], J.P. Evans[1], J.S. Berg[1].* 1) Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC; 3) Department of Neurology, University of North Carolina at Chapel Hill, Chapel Hill, NC.

The application of genome-scale clinical genetic testing inevitably generates large numbers of variants of uncertain clinical significance. Evaluating these numerous variants can greatly complicate and prolong variant adjudication. To avoid undertaking substantial duplication of effort in an already time-intensive endeavor, molecular analysts need a structured way to store and query their evaluation of literature regarding the clinical impact of variants on specific genes. The NCGENES project (North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing) mitigates these challenges through systematic curation of *a priori*-determined sets of genes associated with the clinical signs or symptoms that comprise the indication for testing. Constraining the hypothesis of which genes are regarded as clinically-relevant for a presenting phenotype accepts the possibility of a mild reduction in sensitivity but is expected to increase positive predictive value of diagnostic analysis.

Generation of such lists is a critical task that will facilitate the general analysis of clinically relevant variants. The described ongoing effort has cataloged the phenotypic association, inheritance pattern, and strength of evidence for 1620 genes in 25 phenotypic classes (ranging from narrow scope such as hypodontia to broad phenotypic categories such as central-nervous system disorders). The number of variants examined in an analysis depends on the size of the *a priori* gene list; among over 200 exomes sequenced in NCGENES, for each 100 genes included in a diagnostic list, we have analyzed a median of 5.0 variants previously classified as disease-associated in HGMD, 1.5 rare truncating variants, and 12.9 rare missense variants per individual.

These diagnostic gene lists and the provenance of information used to create them represent an important resource and provide annotations to aid in ongoing adjudication of variants by focusing analysis on the genes with most likely diagnostic relevance in a variety of clinical contexts. Such lists must evolve with burgeoning knowledge; thus the provenance of information and feedback from molecular analysts and clinicians are essential for periodic updates and maintenance of candidate diagnostic gene lists.

## 373

**Validation of small-molecule metabolomic profiling for the clinical screening of inborn errors of metabolism.** *M.J. Miller[1], A.D. Kennedy[2], A.D. Eckhart[2], J.E. Wulff[2], M.V. Milburn[2], J.A. Ryals[2], A.L. Beaudet[1], Q. Sun[1], V.R. Sutton[1], S.H. Elsea[1].* 1) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 2) Metabolon, Research Triangle Park, North Carolina.

Advances in chromatography, mass spectrometry, and informatic technologies have made possible the rapid identification of hundreds of metabolites in a single analysis, but many questions remain about the practical applications of metabolomic profiling in clinical testing. As an initial proof-of-concept, we employed a rapid and scalable metabolomic workflow to analyze 129 plasma samples collected from patients with a confirmed inborn error of metabolism (IEM). In total, 25 different IEMs were represented within our sample set including amino acid, organic acid, fatty acid oxidation, vitamin cofactor, pyrimidine biosynthesis, creatine biosynthesis, and urea cycle disorders. Analysis was completed using a state-of-the-art MS platform, and the resulting spectra were compared against a library of ~2,500 human metabolites. On average, 886 small molecules were detected in a given sample with a core group of 404 analytes found in all specimens. The analytes detected encompass numerous classes of important small molecule biomarkers such as acylcarnitines, amino acids, bile acids, carbohydrates, lipids and nucleotides. For the majority of IEM samples studied, classic pathognomonic compounds were among the most significantly elevated analytes detected. In many cases, metabolomic data afforded a much richer view of a patient's metabolic disturbance by identifying: (1) elevated metabolites located far upstream of the genetic defect, (2) treatment related compounds, and (3) spectrally unique analytes that are not yet associated with a biochemical. In total, metabolic profiling was able to correctly diagnose 19 of the 20 disorders in our panel for which plasma analysis is informative. Importantly, to achieve a similar diagnostic outcome in our laboratory, we estimate ten different biochemical tests would be required. As a negative control, we analyzed plasma specimens from 71 patients who had non-diagnostic testing within our laboratory. In a subset of these cases, metabolomic analysis uncovered disturbances that pointed to a genetic disorder (e.g., sarcosinemia and trimethyllysine hydroxylase deficiency) or assisted in the interpretation of concurrent molecular genetic testing. This proof-of-concept study demonstrates that metabolomic profiling is ready for use in the initial detection of a wide range of IEMs and represents an attractive screening option for phenotypically undifferentiated cases with a suspected biochemical genetic etiology.

## 374

**Free the Data: EmBase and EmVClass Facilitate Storage, Interpretation, Curation, and Sharing of Over 11,000 Sequence Variants Identified Through Clinical Testing.** *L.J.H. Bean, S.W. Tinker, C. da Silva, M.R. Hedge.* Human Genetics, Emory University, Decatur, GA.

Current technology allows clinical laboratories to generate large amounts of sequence data from single genes, gene panels or whole exomes through clinical testing. It is critical that clinical laboratories recognize the importance of the data they hold and share this data with the medical community. To better manage and share our data, Emory Genetics Laboratory (EGL) developed the two components of our data management suite: EmBase and EmVClass. EmBase is EGL's highly-curated clinical grade sequence variant database, maintained at the gene level, the variant level, and the patient level to manage internal workflow and reporting processes. The EmBase data structure is designed to facilitate open sharing of variants identified in samples tested at EGL. To date, EmBase contains over 11,000 variants classified as either pathogenic (n=2670), likely pathogenic (n=89), variant of unknown significance (n=3740), likely benign (n=24), or benign (n=4632). Of these variant classifications, over half (n=5982) have been reviewed and validated since the launch of EmBase in July, 2012. The remaining variant classifications (n=5181) were assigned between 2005 and July, 2012. Importantly, this system tracks changes in variant classifications. Also documented in EmBase are other reportable variants (e.g. pseudodeficiency alleles; n=10). The EmBase data structure was designed to easily transfer data to an electronic medical record or publically available database. To date over 5,300 variants with validated classifications were submitted to the NCBI ClinVar database with an approximately 99% successful submission rate. Because efforts such as the ClinVar project are still in the earliest phase, we developed EmVClass, web-based tool that allows any user access to variants seen at EGL and their current classification. A review of classification for a particular variant can be requested through a simple request form. To date, EmVClass has received over 7000 searches from over 700 distinct users in 50 countries and 41 US states. In addition, data from EmVClass have been absorbed into locus specific databases, such as the Leiden Muscular Dystrophy Pages (dmd.nl). The ease with which EGL sequence variant data can be browsed, searched, and transferred to other databases underscores the need to invest in bioinformatics personnel and infrastructure so that large numbers of curated sequence variants can be stored in a highly structured environment.

## 375

**The role of TET1-mediated demethylation in gene regulation and memory formation.** *A.J. Towers[1], XL. Li[2], A.L. Bey[3], P. Wang[2], YH. Jiang[1,2,3].* 1) Program in Genetics and Genomics, Duke University, Durham, NC; 2) Pediatrics Dept, Duke University, Durham, NC; 3) Neurobiology Dept, Duke University, Durham, NC.

Dynamic regulation of gene expression is implicated in memory formation, although the molecular mechanism remains poorly understood. Accumulating evidence suggests epigenetic modifications are part of the mechanism. Activity-induced DNA demethylation occurs in the hippocampus after learning or electrical stimulation, but the enzymes responsible for it have been elusive. Recently, the oxygenases of the Ten-eleven translocation (TET) family, including TET1, were suggested to play a role in DNA demethylation in the postnatal brain, particularly the hippocampus, a region important for memory formation. The function of TET1 and 5hmC in the postnatal brain, however, is unclear. We hypothesized that TET1 plays an important role in memory formation by regulating activity-dependent gene expression via DNA demethylation.

We have successfully obtained *Tet1* exon3 deletion mice. To determine if activity-dependent genes are dysregulated in *Tet1*-/- hippocampi after stimulation, we performed electroconvulsive shock in *Tet1*-/- and *Tet1*+/+ mice and compared the expression profile by RNA-seq. We used TopHat to align reads and Cufflinks to find differentially expressed genes. The expression of a master, memory gene regulator, Npas4, was significantly downregulated in *Tet1*-/- samples. We then used bisulfite sequencing to examine the promoter region of Npas4 and found it to have increased DNA methylation in *Tet1*-/- samples, suggesting a role for TET1 in regulating the methylation state of Npas4. Gene Ontology analyses of the dysregulated genes using DAVID revealed a significant enrichment for genes involved in the extracellular matrix (ECM), a structure which has been implicated in synaptic plasticity.

We then performed a battery of behavioral tests to address whether hippocampus-dependent learning and memory formation are affected in *Tet1*-/- mice. We discovered both short-term (1 hr) and long-term (24 hr) episodic memory deficits in *Tet1*-/- mice in the novel object recognition and the social transmission of food preference behavioral paradigms.

Our study suggests a role for TET1 in regulating gene expression important for memory formation. A better understanding of the molecular mechanisms of memory formation will be critical for finding future therapies for human memory disorders.

## 376

**DNA methylation in the central nucleus of the amygdala contributes to anxious temperament in young primates.** *R.S. Alisch[1], P. Chopra[1], A.S. Fox[1], K. Chen[2], A.T.J. White[1], P.H. Roseboom[1], S. Keles[2], N.H. Kalin[1].* 1) Dept. of Psychiatry, Univ. of Wisconsin SMPH, Madison, WI; 2) Dept. of Statistics, Univ. of Wisconsin, Madison, WI.

Considerable evidence demonstrates that children with an anxious temperament (AT) are at a substantially increased risk to develop anxiety and depression. A validated non-human primate model of AT revealed that the critical neural components underlying AT (e.g. the central nucleus of the amygdala, CeA) differ in their level of heritability, suggesting a role for epigenetics in the expression of this at risk phenotype. Here, we profiled 5-methylcytosine in the CeAs of rhesus monkeys (N = 23) fully phenotyped for AT and identified 5,489 CpG sites (1,363 genes) with methylation levels predictive of individual differences in AT (FDR p-value < 0.05), including genes previously implicated in psychiatric-related disorders (e.g. GRIN1, GRM5, HTT, ADCYAP1, and SHANK3). To further link these genes to function, we examined the CeA gene expression data from these same monkeys and found that AT-associated methylation patterns in twenty-two genes were also correlated with gene expression (p-value < 0.05), including the glutamate receptors, GRIN1 and GRM5. Of these twenty-two genes, BCL11A and JAG1 had expression levels that also significantly predicted AT (p-value < 0.05). These transcripts have well-defined roles in neurodevelopmental processes, including axon branching and dendrite outgrowth (BCL11A) and the regulation of astrogenesis and neurogenesis (JAG1). Together, these data highlight genes of importance in the expression of the early life risk to develop anxiety and depression and implicate epigenetic mechanisms involved in modifying the function of the CeA, a core neural substrate of AT.

## 377

**MicroRNA-486 overexpression delays the disease pathology of muscular dystrophy.** *M.S. Alexander[1,2], J.C. Casar[3], N. Motohashi[4], N.M. Vieira[1,2], I. Eisenberg[5], J.L. Marshall[1,2], M.J. Gasperini[1], A. Lek[1,2], J.A. Myers[1], E.A. Estrella[1,6], P.B. Kang[1,6,7], F. Shapiro[8], F. Rahimov[1,2], G. Kawahara[1,2], J.J. Widrick[1], L.M. Kunkel[1,2,9,10].* 1) Division of Genetics and Genomics at Boston Children's Hospital, Boston, MA 02115; 2) Department of Pediatrics and Genetics at Harvard Medical School, Boston, MA 02115; 3) Departamento de Neurología, Escuela de Medicina, Pontificia Universidad Católica de Chile. Santiago, RM, Chile; 4) Stem Cell Institute, Paul and Sheila Wellstone Muscular Dystrophy Center, Department of Neurology, University of Minnesota Medical School, Minneapolis, MN 55455; 5) Center for Human Placenta Research, Department of Obstetrics and Gynecology, Hadassah-Hebrew University Medical Center- Mt. Scopus Jerusalem, Israel; 6) Department of Neurology, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115; 7) Present Address: Division of Pediatric Neurology, University of Florida College of Medicine, Gainsville, FL 32610; 8) Departments of Orthopedic Surgery at Boston Children's Hospital and Harvard Medical School, Boston MA 02115; 9) The Manton Center for Orphan Disease Research at Boston Children's Hospital, Boston, MA 02115; 10) Harvard Stem Cell Institute, Cambridge, MA 02138.

Duchenne muscular dystrophy (DMD) is caused by mutations in the dystrophin gene that result in the dysregulation of many signaling pathways that interact directly or indirectly with the dystrophin protein. Previously, we identified miR-486 as being strongly reduced in its expression levels in the dystrophin-deficient mouse and muscle biopsies of human DMD patients. Here we report that transgenic overexpression of the muscle-enriched microRNA, miR-486, in mdx5cv (dystrophin-mutant) mice resulted in improved serum biochemistry, reduced apoptosis, increased myofiber size, and improved muscle physiological force output. Using a bioinformatic approach, we identified DOCK3, dedicator-of-cytokinesis-3, as being a direct downstream target of miR-486 in skeletal muscle. Manipulation of DOCK3 expression in myoblast cell culture had strong effects on normal and DMD myoblast apoptosis, and on the RAC1/RHOA signaling pathway. Overexpression of DOCK3 resulted in myoblast apoptosis and reduced myoblast fusion, and activated RAC1/RHOA signaling in normal muscle cells. Conversely, while overexpression of DOCK3 in DMD muscle cells induced myoblast apoptosis and reduced myoblast fusion, the RAC1/RHOA signaling pathway was not activated. Overexpression of miR-486 in DMD myoblasts resulted in a restoration of normal RAC1/RHOA signaling most likely due increased muscle membrane stability and reduced muscle apoptosis. Together, these studies demonstrate that stable overexpression of miR-486 ameliorates many of the signs of the disease pathology of dystrophin-deficient muscle.

## 378

**Mutations in nuclear envelope change myogenic epigenomic programs and normal cell fate.** *J. Perovanovic[1,2], S. Dell"orso[3], V. Sartorelli[3], K. Mamchaoui[4,5], V. Mouly[4,5], C. Vigouroux[6,8,9,10], G. Bonne[4,5,7], E.P. Hoffman[1,2].* 1) Center for Genetic Medicine Research Children's National Medical Center, Washington, DC, 20010, USA; 2) Department of Integrative Systems Biology, The George Washington University School of Medicine and Health Sciences, Washington, DC, 20010, USA; 3) Laboratory of Muscle Stem Cells and Gene Regulation, National Institute of Arthritis, Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20852, USA; 4) INSERM U974, F-75013, Paris, France; 5) Sorbonne Universités, UPMC Univ Paris 06, Myology Center of Research, UMRS974; Institut de Myologie, F-75013, Paris, France; 6) Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Tenon, Service de Biochimie et Hormonologie, F-75020, Paris, France; 7) Assistance Publique-Hôpitaux de Paris, Groupe Hospitalier Pitié-Salpêtrière, U.F. Cardiogénétique et Myogénétique, Service de Biochimie Métabolique, Paris F-75013, France; 8) INSERM UMR_S938, Centre de Recherche Saint-Antoine, F-75012, Paris, France; 9) Sorbonne Universités, UPMC Univ Paris 06, UMR_S938, F-75005, Paris, France; 10) ICAN, Institute of Cardiometabolism and Nutrition, Paris, France.

Mutations of *LMNA* cause wide spectrum disorders and exhibit allelic heterogeneity of post mitotic tissue with mutation specific phenotypes. One of the phenotypes is Emery-Dreifuss Muscular Dystrophy (EDMD (MIM 300200 and MIM 181350) where *LMNA* mutations and loss-of-function mutations in *EMD* specifically affect muscle tissue. Here we show that *EMD* and *LMNA* mutations alter epigenomic programming during myogenesis via shared locus-specific perturbations of chromatin-nuclear envelope interactions. ChIP-seq (H3K9me3) showed that loss of emerin from myogenic cells leads to decrease in heterochromatin enrichment on key cell fate regulators and re-activation of signaling pathways involved in stem cell differentiation. Specifically, ChIP-seq show differential silencing of the *Sox2* locus and its downstream targets in emerin null cells, suggesting that commitment to myogenic lineages was perturbed. To determine if EDMD-AD (MIM 181350) LMNA mutations showed allele-specific perturbations of myogenic *SOX2* silencing and downstream targets, patient-derived normal, EDMD-AD (MIM 181350) (p.H222P), and Familial Partial Lipodystrophy (FPLD (MIM 151660)) (p.R482W) cells were studied by ChIP and qRT-PCR. Analysis showed allele-specific epigenomic perturbation of myogenic silencing at the *SOX2* locus. Furthermore, EDMD patient muscle biopsies showed disease-specific overexpression of *SOX2* pathways relative to normal and disease-control (FKRP (MIM 607155)) muscles. Direct interaction between wild-type LMNA protein and the *SOX2* locus was shown by DamID methods, and these interactions were disrupted by LMNA disease mutations. These findings suggest that nuclear envelope disorders cause allele-specific alterations in cell commitment via inadequate epigenomic programming.

## 379

**Aberrant DNA hypermethylation of *SDHC*: A novel mechanism of tumor development in Carney Triad.** *F. Faucz[1], F. Haller[2], E.A. Moskalev[2], S. Batthelmeb[2], S. Wiemann[3], M. Bieg[4], G. Assie[5,6], J. Bertherat[5,6], I-M. Schaefer[7,12], C. Otto[8], E. Rattenberry[9], E.R. Maher[9,10], P. Strobel[7], M. Werner[8], J.A. Carney[11], A. Hartmann[2], A. Agamy[2], C.A. Stratakis[1].* 1) Program on Developmental Endocrinology and Genetics, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, Bethesda, Maryland, USA; 2) Institute of Pathology, University Hospital, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany; 3) Genomics and Proteomics Core Facility, German Cancer Research Center, Heidelberg, Germany; 4) Division of Theoretical Bioinformatics, German Cancer Research Center, Heidelberg, Germany; 5) Institut Cochin, INSERM U1016, CNRS UMR 8104, Université Paris Descartes, Sorbonne Paris Cité, Paris, France; 6) Department of Endocrinology, Referal Center for Rare Adrenal Diseases, Assistance Publique Hôpitaux de Paris, Hôpital Cochin, Paris, France; 7) Institute of Pathology, University Hospital, Georg-August University, Göttingen, Germany; 8) Institute of Pathology, University Hospital, Albert-Ludwigs University Freiburg, Freiburg, Germany; 9) Centre for Rare Diseases and Personalised Medicine, School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom; 10) Department of Medical Genetics, University of Cambridge, Cambridge CB2 0QQ, UK; 11) Laboratory Medicine and Pathology, Emeritus Staff, Mayo Clinic, Rochester, Minnesota, USA; 12) Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

Carney triad (CT) is a rare condition with synchronous or metachronous occurrence of gastrointestinal stromal tumors (GISTs), paragangliomas (PGLs) and pulmonary chondromas in a patient. The disease has a striking predilection for young females, for reasons that remain unknown. In contrast to Carney-Stratakis Syndrome (CSS) and familial PGL syndromes, no germline or somatic mutations in the succinate dehydrogenase complex (SDH) subunits A, B, C or D have been found in most tumors and/or patients with CT. Nonetheless, the tumors arising among patients with CT, CSS or familial PGL share a similar morphology with loss of the SDHB subunit on the protein level. Given the shared loss of SDHB protein in tumors among all three syndromes, we hypothesized epigenetic silencing of SDH genes as an alternate mechanism in tumors occurring in CT patients. For the current study we performed broad, high-resolution and quantitative assessment of DNA methylation to evaluate CpG islands pattern in the proximity of transcriptional start sites of all four genes encoding the SDH subunits in tumors from four CT patients who did not have any SDH coding sequence abnormalities. The DNA methylation patterns were compared to tumors from a patient with CSS, a patient with PGL1 as well as sporadic GISTs harboring mutations in the KIT receptor. For the first time, we report on a recurrent aberrant dense DNA methylation at the gene locus of the *SDHC* in all tumors from the CT patients, while virtually no methylation was detectable in tumors of patients with CSS or PGL, or in sporadic GISTs with KIT mutations. mRNA expression of all four SDH subunits A, B, C and D was determined by qPCR, and a significant downregulation of SDHC on mRNA level in the CT tumors was observed what was in contrast to a virtually equal expression of all four SDH subunits in the other tumor samples. Both SDHB and SDHC subunits were absent at the protein level in the tumors from the CT patients. Collectively, these data demonstrate that DNA methylation of the *SDHC* gene locus is a recurrent and specific event in tumors of CT patients, and suggests SDHC downregulation by epigenetic inactivation as a plausible tumor-initiating event.

## 380

**A screen-informed candidate gene approach identifies a large human telomere maintenance network.** *B. Holohan, W. Wright, J. Shay.* Cell Biology, UT Southwestern Medical Center, Dallas, TX.

Telomeres are nucleoprotein structures on the ends of chromosomes that shield the termini of linear chromosomes from recognition by the DNA double strand break repair machinery. Because of the end replication problem, telomeres shorten with every cell division, thus telomere shortening is an inherent limitation in the number of times that a cell can divide before triggering a DNA damage response. Germ cells and certain stem cells utilize telomerase, a ribonucleoprotein reverse transcriptase, to partially maintain their telomere lengths. This allows them to have extended proliferative potential, but telomeres still get progressively shorter with increased age. In contrast, roughly 90% of human tumors use telomerase to maintain their telomeres, and they can fully maintain their telomere lengths although most cancer cells have very short telomeres. Since telomerase is not expressed in most somatic cells but is expressed in the vast majority of cancer cells, telomerase is a near-universal target for cancer therapy. In order to identify ways to target telomerase for cancer therapy, we constructed a mutant telomerase RNA (TERC) component that templates non-canonical telomere repeats when utilized by telomerase. These mutant repeats are toxic via telomere upcapping and activation of the DNA double strand break repair machinery, causing senescence, genomic catastrophe or apoptosis. The introduction of the mutant TERC into telomerase silent cells had no effect on cell growth. However, using cells with an inducible telomerase (TERT) allowed us to utilize shRNA tools to identify positive regulators of telomerase; shRNAs that reduced the incorporation of mutant sequences upon telomerase induction would allow cell survival. Using this system and a screen-informed candidate gene approach, we have identified and validated 50 new genes that induce telomere shortening in human cells upon knockdown. Based on the upstream regulators of some of the genes identified, we examined the effects of Perifosine, an AKT inhibitor, on telomere biology in vitro and in human xenograft studies. Perifosine induced telomere shortening in a majority; of tumor cell lines investigated and reduced telomere length and tumor size in xenografts after prolonged exposure. Our work has revealed a large and surprisingly malleable telomerase regulation network that suggests a number of existing drugs may be useful telomerase modulators.

## 381

**Association between telomere length SNPs and five cancer types: a Mendelian randomization study from the GAME-ON post-GWAS consortium.** *C. Zhang[1], S. Burgess[4], P. Kraft[5], S. Lindstrom[5], R. Hung[6], U. Peters[7], H. Ahsan[1,2,3], T. Sellers[8], A. Monteiro[8], G. Trench[9], J. Doherty[10], B. Pierce[1,2] on behalf of the CORECT, DRIVE, ELLIPSE, OCAC, and TRICL studies and the GAME-ON Network.* 1) Department of Health Studies, The University of Chicago, Chicago, IL; 2) Comprehensive Cancer Center, The University of Chicago, Chicago, IL; 3) Departments of Medicine and Human Genetics, The University of Chicago, Chicago, IL; 4) Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK; 5) Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard School of Public Health, Boston, MA; 6) Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto; 7) Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA; 8) Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center, Tampa, FL; 9) Department of Genetics, Queensland Institute of Medical Research, Brisbane, Queensland, Australia; 10) Section of Biostatistics & Epidemiology, The Geisel School of Medicine at Dartmouth, Lebanon, NH.

Epidemiological studies have reported associations between telomere length (measured in peripheral blood cells) and risk for various cancer types, but results have been inconsistent. These inconsistencies have been attributed in part to methodological issues related to the use of cross-sectional or retrospective study designs, which may be biased due to reverse causality or the effects of cancer progression and treatment. One way to address this issue is to indirectly estimate the association between telomere length and cancer risk using single nucleotide polymorphisms (SNPs) associated with telomere length. We estimated the association between a multi-SNP score consisting of nine telomere length-associated SNPs and risk for five cancer types (breast, lung, colorectal, ovarian and prostate cancer) including various subtypes, using data from the Genetic Association Mechanisms in Oncology genome-wide association study consortium (GAME-ON). The association estimates correspond to the effects of telomere length on the cancer risk under Mendelian randomization assumptions. We found the multi-SNP score for short telomeres to be significantly associated with decreased risk of lung adenocarcinoma (OR = 0.31, 95% CI: 0.23-0.41, P = $6.1\times10^{-15}$), and suggestively associated with decreased risk of colorectal (OR = 0.73, 95% CI: 0.52-1.03, P = 0.08) and prostate (OR = 0.82, 95% CI: 0.65-1.03, P = 0.09) cancer. These estimates can be interpreted as the reduction in odds of cancer risk per 1000 base pair reduction in telomere length. Our results suggest that short telomere length is associated with decreased risk of lung adenocarcinoma, possibly associated with decreased risk of prostate and colorectal cancer, and not associated with breast or ovarian cancer. Findings from this study provide additional information to help interpret relationships between telomere length and specific cancer types, and may be used in the development of cancer prediction strategies.

## 382

**Imputation and subset based association analysis across different cancer types identifies multiple independent risk loci in the TERT-CLPTM1L region on chromosome 5p15.33.** *Z. Wang[1,2], M. Zhang[1], B. Zhu[1], H. Parikh[1], J. Jia[1], C.C Chung[1,2], J.N. Sampson[1], J.W. Hoskins[1], A. Hutchinson[1,2], L. Burdette[1,2], L. Mirabello[1], S.A. Savage[1], P. Kraft[3,4], S.J. Chanock[1], M. Yeager[1,2], M.T. Landi[1], J. Shi[1], N. Chatterjee[1], L.T. Amundadottir[1] For AsianLung,EurLung,AALung,PanScan,ChinaPC,TGCT, GliomaScan,BladderNCI,Pegasus,CGEMS PrCa, AdvPrca.* 1) Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America; 2) Cancer Genomics Research Laboratory, National Cancer Institute, Division of Cancer Epidemiology and Genetics, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States of America; 3) Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America; 4) Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America.

Genome-wide association studies (GWAS) have mapped risk alleles for at least ten distinct cancers to a small region of 63,000 bp on chromosome 5p15.33. This region harbors the *TERT* and *CLPTM1L* genes; the former encodes the catalytic subunit of telomerase reverse transcriptase and the latter may play a role in apoptosis. To investigate further the pleiotropy in this region, we conducted an agnostic subset-based meta-analysis (ASSET) across six distinct cancers in 34,248 cases and 45,036 controls. Based on sequential conditional analysis, we identified as many as six independent risk loci marked by common single nucleotide polymorphisms (SNPs): five in the *TERT* gene (region 1: rs7726159, P=$2.10\times10^{-39}$; region 3: rs2853677, P=$3.30\times10^{-36}$ and $P_{Conditional}=2.36\times10^{-8}$; region 4: rs2736098, P=$3.87\times10^{-12}$ and $P_{Conditional}=5.19\times10^{-6}$, region 5: rs13172201, P=0.041 and $P_{Conditional}=2.04\times10^{-6}$; and region 6: rs10069690, P=$7.49\times10^{-15}$ and $P_{Conditional}=5.35\times10^{-7}$) and one in the neighboring *CLPTM1L* gene (region 2: rs451360; P=$1.90\times10^{-18}$ and $P_{Conditional}=7.06\times10^{-16}$). Between three and five cancers mapped to each independent locus with both risk-enhancing and protective effects. Allele specific effects on methylation were seen for a subset of risk loci indicating that methylation and subsequent effects on gene expression may contribute to the biology of risk variants on 5p15.33. Our results provide strong support for extensive pleiotropy across this region of 5p15.33, to an extent not previously observed in other cancer susceptibility loci. This project has been funded in whole or in part with federal funds from the National Cancer Institute National Institutes of Health, under contract HHSN261200800001E.The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

## 383

**Genome-wide analysis of mitochondrial single nucleotide polymorphism (mtSNP)-nuclear SNP interaction in age-related macular degeneration (AMD).** *P.J. Persad[1], M.D. Courtenay[1], G. Wang[1], P.W. Gay[1], W. Cade[1], A. Agarwal[3], S.G. Schwartz[2], J.L. Kovach[2], M.A. Brantley[3], R.J. Sardell[1], J.N. Cooke Bailey[4], J.L. Haines[4], M.A. Pericak-Vance[1], W.K. Scott[1].* 1) Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL; 2) Bascom Palmer Eye Institute, Naples, FL, University of Miami Miller School of Medicine, Miami, FL; 3) Vanderbilt University, Nashville, TN; 4) Department of Epidemiology & Biostatistics, Case Western Reserve University, Cleveland, OH.

AMD has 19 identified genetic risk factors, and variations in *CFH* are among the strongest. Deletion of related genes *CFHR1* and *CFHR3* is inversely associated with AMD. Mitochondrial (mt) dysfunction contributes to retinal pigment epithelium degeneration, which characterizes AMD. Prior studies suggested that mt haplogroup H was inversely associated with AMD and was associated with increased *CFH* expression (Kenney et al., 2013). Since this implied that mt variants might alter effects of nuclear SNPs (nSNPs), we examined joint effects of mtSNPs and nSNPs in a genome-wide interaction analysis of 668,238 SNPs on the Affymetrix 6.0 GWAS chip (Naj et al., 2013). In total, 1,419 people with smoking history were included: 894 AMD cases with varying grades of AMD and 525 controls. A two-degree of freedom joint test of gene-gene interaction was conducted using PLINK. A model including the nSNP (coded additively), the mtSNP (coded 0,1), and the nSNP-mtSNP interaction term was compared to a second model excluding the nSNP and interaction terms. Both models were adjusted for age, sex, and smoking. Interaction was assessed for 26 mtSNPs and all nSNPs. Analyses stratified by mt allele for statistically significant interactions utilized logistic regression models containing the nSNP, age, smoking, and sex. Joint effects of mtSNP rs3928306 in the 16S rRNA gene and nSNPs rs13375236, rs6428369, and rs2336503 in the *CFH* region achieved genome-wide statistical significance ($p<5.00\times10^{-8}$) with nominally statistically significant ($p<0.01$) interaction effects. Analyses stratified by rs3928306 allele revealed genome-wide significant association in rs3928306-G carriers for the three nSNPs, which were in complete linkage disequilibrium (LD) [OR=0.31, 95% CI: (0.22-0.44), $p=2.91\times10^{-11}$]; there was no association in the rs3928306-A carriers [OR= 0.87, CI: (0.47-1.61); p=0.66]. The three nSNPs were in moderate LD ($r^2=0.52$) with rs7542235 (HapMap phase II data), a tag SNP for the *CFHR1-CFHR3* deletion (Raychaudhuri et al., 2010). Thus, the mtSNP rs3928306 may modify the association of the *CFHR1-CFHR3* deletion with AMD. This deletion's protective effect is not present in rs3928306-A carriers but is found in rs3928306-G carriers. These results demonstrate the importance of considering the joint effects of mtSNPs and nSNPs in AMD.

## 384

**Unravelling the complex genetics of age-related macular degeneration — The International AMD Genomics Consortium (IAMDGC).** *V. Cipriani[1,2,3] on behalf of the International AMD Genomics Consortium (IAMDGC).* 1) UCL Institute of Ophthalmology, University College London, London, United Kingdom; 2) Moorfields Eye Hospital, London, United Kingdom; 3) UCL Genetics Institute, London, United Kingdom.

**Background** Age-related macular degeneration (AMD) is a leading cause of blindness in the elderly. AMD susceptibility is influenced by multiple environmental and genetic factors. To accelerate the discovery in the complex genetics of AMD, the International AMD Genomics Consortium (IAMDGC) has carried out the largest genotyping of rare and common variation ever conducted on AMD. **Methods** A large set of predominantly late AMD cases (N≈26K) and age-matched controls (N≈22K) were gathered from 26 studies and centrally genotyped on the same custom genome-wide array (>250K tag-SNPs + >40K AMD variants) enriched for >225K common and rare exonic variants. Single-variant, forward conditional, haplotype-based tests and pathway analyses were performed on genotyped and 1000 Genomes-based imputed variants. We report results from primary analysis on unrelated late cases and controls of European ancestry. **Results** Single-variant association tests on 16,144 late AMD cases and 17,832 controls and >12M genotyped and well-imputed variants confirmed 18 of the 19 established AMD loci (Fritsche et al., *Nat Genet*, 2013). Genome-wide significance ($P<5\times10^{-8}$) was observed for common variants at 16 novel loci (OR range=0.7-1.5; AUC score=0.58). Conditional tests culminated in a set of 57 common and rare independently associated variants (up to 36% of disease variance explained) with multiple signals within 10 known loci, e.g. *CFH*, *C2/CFB*, *C3*, *COL8A1* and *RAD51B*. Heterozygotes for the common Y402H haplotype had variable risk (OR range=0.2-7.1) compared to homozygotes, and rare *CFH* haplotypes without the R1210C variant showed risk that was not significantly different from the Y402H haplotype. Sub-phenotype analyses revealed loci with different effects on the dry and wet form of the disease. Enrichment of pathways beyond the known complement cascade was observed, e.g. lipid metabolism. **Discussion** The IAMDGC has carried out the largest discovery of rare and common variants for AMD that has already resulted in an outstanding yield of 16 novel loci. Ongoing fine mapping analyses suggest independent evolution of multiple functional entities underlying an association locus. Pathway analyses and available expression data have already indicated promising leads for subsequent functional studies to uncover the biological significance of these findings. The IAMDGC effort has enormous potential to guide development of future drug targets and inform genetically guided management of patients.

## 385

**Whole-Genome Sequencing Study of ~6,000 Samples for Age-related Macular Degeneration.** *A. Kwong[1,2], X. Zhan[1,2], L.G. Fritsche[1,2], J. Bragg-Gresham[1,2], K.E. Branham[3], M. Othman[3], A. Boleda[6], L. Gieser[6], R. Ratnapriya[6], D. Stambolian[4], E.Y. Chew[5], A. Swaroop[6], G. Abecasis[1,2].* 1) Department of Biostatistics, University of Michigan, Ann Arbor, MI; 2) Center for Statistical Genetics, University of Michigan, Ann Arbor, MI; 3) Department of Ophthalmology and Visual Sciences, University of Michigan Kellogg Eye Center, Ann Arbor, MI; 4) Department of Ophthalmology and Human Genetics, University of Pennsylvania Medical School, Philadelphia, PA; 5) Division of Epidemiology and Clinical Applications, National Eye Institute/National Institutes of Health, Bethesda, MD; 6) Neurobiology-Neurodegeneration and Repair Laboratory, National Eye Institute/National Institutes of Health, Bethesda, MD.

Purpose: Age-related Macular Degeneration (AMD) is a leading cause of blindness among the elderly. Over the past several years, genetic studies of common variation have provided many clues about disease biology. Due to assay limitations, these studies have typically either ignored rare variants or examined them only in a small set of candidate regions. Here, we set out to systematically study the contribution of rare variants to disease. Methods: We assembled a collection of ~3,000 cases and ~3,000 controls with advanced AMD (67% neovascularization, 33% geographic atrophy) from the Kellogg Eye Center at University of Michigan, Age-Related Eye Disease Study from the NEI, and the University of Pennsylvania. We matched cases and controls according to age, gender, and ethnicity. The large number of samples to be processed presented a computational challenge. Our data will enable a systematic assessment of coding and non-coding variation in previously associated loci as well as a genomewide search for new risk alleles. Results: As of today, ~3,000 samples have been sequenced, representing >55 Terabytes ($5.5 \times 10^{13}$ bytes) of sequence data. This corresponds to a total genomic coverage of ~18,000x and an average coverage of ~6x per sample. In an initial analysis of a subset of the data, we discovered and genotyped ~31 million variants. The set includes several previously-studied rare AMD risk variants that were found in complement genes (such as *CFH*:p.R1210C, *CFI*:p.G119R, *C9*:p.P167S, and *C3*:p:K155Q), but also many new functionally-interesting variants, such as 20 missense and 1 nonsense mutations in the *CFH* gene that are very rare (median minor allele frequency = 0.05%) and missing from previous studies. Among the variants in the current dataset, we found 172,971 non-synonymous SNPs and 7,601 loss-of-function SNPs. Conclusions: We provide a first detailed look at the genetics of AMD through whole-genome sequencing of ~6,000 individuals. Our data will enable a systematic genomewide search for rare risk alleles and should allow us to evaluate the effect of nonsense variants in many previously associated genes.

## 386

**Examining the Casual Role of Central Corneal Thickness in Glaucoma: A Mendelian Randomization Approach.** *C.Y. Cheng[1,2,3], T.H. Tham[1,2], J.M. Liao[2], T.Y. Wong[1,2,3], T. Aung[1,2].* 1) Singapore Eye Research Institute, Singapore National Eye Centre, Singapore; 2) Department of Ophthalmology, National University of Singapore and National University Health System, Singapore; 3) Duke-National University of Singapore Graduate Medical School, Singapore.

Thin central corneal thickness (CCT) has been previously reported to be associated with glaucoma. Nevertheless, its causal role in glaucoma remains controversial. In this study, we aimed to determine the causal relationship between CCT and glaucoma by using the Mendelian randomization (MR) approach with CCT-associated SNPs as instrumental variables. Participants across 3 ethnicity cohorts of the Singapore Epidemiology Eye Disease (SEED) Study were included in this analysis. We first identified SNPs which were most strongly associated with CCT in our cohorts, by examining 100kb within 15 previously established CCT genes. Based on the selected SNPs, we then constructed a multi-locus genetic risk score (GRS) by summing the number of alleles of each SNP, weighted by its effect size with CCT. We examined the association between CCT with primary open angle glaucoma (POAG) and all glaucoma using conventional logistic regression analyses, while adjusted for age, gender, intraocular pressure and axial length. We also performed MR analyses which used CCT GRS as instrument variables to test the associations between CCT and glaucoma. Each cohort was analyzed separately and the estimates were combined across cohorts with fixed effect meta-analysis. A total of 6,945 participants (2,529 Malay, 2,531 Indian, and 1,885 Chinese) across the 3 study cohorts were included in the analyses. There were a total of 222 glaucoma cases, of which 142 were POAG. Conventional logistic regression analyses showed that each 10 μm decrease in CCT was significantly associated with increased risk of glaucoma (odds ratio [OR] 1.08, 95% confidence interval [CI] 1.03 to 1.13, $P$ = 0.001) and POAG (OR 1.09, 95% CI 1.03 to 1.15, $P$ = 0.003). However, MR analyses revealed no evidence of causal relationships between CCT with glaucoma (OR 1.05, 95% CI 0.98 to 1.13, $P$ = 0.197) and POAG (OR 1.00, 95% CI 0.92 to 1.00, $P$ = 0.937). As opposed to findings from conventional analyses, MR approach showed that thinner CCT is not causally associated with glaucoma and POAG. These findings do not collectively provide consistent evidence to substantiate the causal role of CCT in glaucoma development.

## 387

**The role of rare *TIMP3* mutations in Age-Related Macular Degeneration.** *L.G. Fritsche, International AMD Genomics Consortium.* Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI.

Age-related macular degeneration (AMD) is the most common cause of blindness in the elderly. The disease typically has onset at >70 years of age. Disease progression results in loss of photo-receptors and ultimately leads to loss of central vision. Variants in ≥20 loci have been associated with susceptibility to AMD, including common non-coding variants near *SYN3* and *TIMP3*, an especially attractive candidate gene. Mutations in *TIMP3*, particularly those that result in unpaired cysteine residues in the C terminus of the protein, can result in Sorsby fundus dystrophy (SFD; Weber *et al.*, 1994). SFD is a rare autosomal-dominant retinal dystrophy that typically has onset <50 years of age. Disease progression is similar to that of age related macular degeneration, including choroidal neovascularization and pigment epithelium atrophy. To elucidate the potential role of rare *TIMP3* mutations in late onset degenerative disease, we used genotyping arrays to examine 10 reported *TIMP3* mutations and 44 other *TIMP3* variants that, if present, would result in unpaired cysteine residues in >24,000 AMD cases and >20,000 controls. We observed 8 of these variants at least once in 40,633 unrelated Europeans. On aggregate, we identified 28 heterozygotes among 16,144 advanced AMD cases (with geographic atrophy or choroidal neovascularization) and one heterozygote in a single control. The burden of *TIMP3* mutations was thus significantly enriched in advanced AMD patients (odds ratio = 30.1; P = 0.00081). Interestingly, the majority of this burden was accounted for by a "predicted" mutation TIMP3:p.(Ser38Cys) located in the N terminus of the protein. Mutations in the N terminus of TIMP3 have not been previously implicated in macular disease. The average age of onset for macular disease among mutation carriers was significantly younger (64.5 years vs. 76.8 years for non-carriers, P<.000001), but still later than for typical cases of SFD. Phenotype re-evaluation of mutation carriers and segregation analysis in available family data is ongoing and will help clarify if *TIMP3* mutation carrying AMD cases more closely resemble SFD cases than typical age-related disease. Our results illustrate how rare coding variants can contribute substantially to late onset disease and also illustrate an inexpensive approach for array based screening of 10,000s of individuals for interesting variants. The approach can be naturally extended to search for premature stop codons in many genes.

## 388

**High throughput screening of 51 known causative genes in families with congenital cataract.** *S. Javadiyan[1], J. Craig[1], S. Sharma[1], K. Lower[2], K. Burdon[1, 3].* 1) Ophthalmology , Flinders university, Adelaide, South Australia, Australia; 2) Haematology and Genetic Pathology,Flinders university , Adelaide, South Australia, Australia; 3) Menzies Research Institue, University of Tasmania, Hobart, Tasmania, Australia.

Congenital cataract is a leading cause of blindness in children. Approximately 200,000 children worldwide are blind from this condition. The incidence in Australia is 2.2 per 10,000 live births. Mutations in genes that encode enzymes, structural proteins, membrane proteins, transcription factors and signalling molecules are associated with hereditary forms of the disease. The overall aim of the study is to describe the spectrum of mutations in known congenital cataract genes and to determine the contribution of each gene to the disease in Australia. We screened 51 reported congenital cataract causing genes in 70 probands of families with congenital cataract. The study adhered to the tenets of the Declaration of Helsinki. Custom Ampliseq libraries were sequenced on the Ion Torrent Personal Genome Machine. Reads were mapped against human genome (hg19) and variants called with the Torrent Suite software. Variants were annotated to dbSNP 137 using Ion Reporter (IR 1.6.2). Variants were prioritised for validation if they were not in public databases or an in-house list of known artefacts and were predicted to be protein changing. The average coverage depth was 1014 x while 89.7 % of target bases were covered at least 100 x. A total of 41 novel variants were detected of which 29 have been validated with Sanger sequencing. Five were sequencing errors and 7 are remaining to be validated. Of the 70 probands sequenced, 41 did not have mutations in selected genes. The remaining 29 had at least one variant passing the filtering criteria out of which 3 were benign or non-segregating mutations. Segregating mutations were identified in 11 families. In addition, low penetrance but likely causative mutations were detected in 6 families and 4 further likely causative variants were identified but families were not available for segregation analysis. We are in the process of validation of further 5 candidate mutations for 5 families. We have identified the genetic cause of congenital cataract in ~ 32 % of cases which suggests that more causative genes are yet to be identified. The observed mutations are present in a range of genes including *CRYAA, CRYGS, BFSP2, GJA3*, *GJA8* and *GCNT2*. The most commonly mutated gene in our repository is *GJA8*.

## 389

**"Primary Cilia Mediate Retinal Development and Photoreceptor Homeostasis".** *C. Carter[1,2], A. Drack[3], Q. Zhang[1,2], N. Nuangchamnong[4], C. Searby[1,2], V.C. Sheffield[1,2,3].* 1) Howard Hughes Medical Institute, University of Iowa Carver College of Medicine, Iowa; 2) Department of Pediatrics, Division of Medical Genetics, University of Iowa Carver College of Medicine, Iowa; 3) Department of Ophthalmology and Visual Sciences, University of Iowa Carver College of Medicine, Iowa; 4) Obstetrics and Gynecology, Division of Maternal-Fetal Medicine, University of Iowa Carver College of Medicine, Iowa.

Retinal degenerative diseases such as retinitis pigmentosa (RP) and macular degeneration are the leading cause of incurable blindness in the western world affecting one in 2,000 individuals worldwide. A common finding shared among these diseases is the loss of photoreceptors leading to blindness. However, the lack of therapy for these diseases is due to a poor understanding of the disease mechanisms leading to the loss of photoreceptors. Recent research implicates cilia, tiny hair-like organelles protruding from the cell surface, in the pathophysiology of photoreceptor loss as forms of retinal degeneration show high penetrance in many human ciliopathies including, Bardet-Biedl syndrome (BBS), Meckel-Gruber syndrome and Leber congenital amaurosis. Animal models of retinal degeneration have been investigated for decades in the hope of understanding the cause of photoreceptor cell death, however, the disease mechanisms leading to photoreceptor loss remain unknown. Here we employ several mouse models with primary cilia dysfunction and we discover a novel cause of retinal degeneration. We find that conditional ablation of the cilia genes IFT88 and BBS1 and the growth factor receptor, PDGFRα in retinal glia (PDGFRα+ and GFAP+ Cre) leads to photoreceptor degeneration and blunted electroretinogram responses indicating impaired visual function. These unique and novel models allowed us to study the contribution of the inner retina to retinal degeneration for the first time. The blunted visual responses occur in the presence of normal rhodopsin localization and normal photoreceptor outer disk ultrastructure. Moreover, PDGFRαCKO mice, display impaired B-wave in the presence of a normal A-wave, indicating that the primary cause of visual deficits in these mice is a result of dysfunction within the inner layers of the retina. Importantly, we demonstrate the specificity of our gene knockouts in retinal Muller cells and other retinal glia with no expression observed in photoreceptors. Together these data introduce a novel, parallel and photoreceptor independent mechanism underlying impaired visual function in ciliopathies in addition to the well-characterized retinal degeneration present. These findings introduce a novel role of cilia and retinal glia in maintaining photoreceptor function and disease mechanism for retinal degeneration.

## 390

**Using zebrafish to assess novel therapeutics and model the eye disease of cblC disease.** *N.P. Achilly, J.L. Sloan, K. Bishop, M.S. Jones, R. Sood, C.P. Venditti.* National Human Genome Research Institute, National Institutes of Health, Bethesda, MD.

Cobalamin C disease (cblC) is the most common inborn error of intracellular cobalamin metabolism and is caused by mutations in MMACHC, a gene responsible for processing and trafficking intracellular cobalamin. Disease manifestations include growth failure, anemia, heart defects, and progressive blindness. At present, the pathological basis of these symptoms remains unknown, as no animal viable model exists. Using zinc-finger nucleases, we generated two independent lines with mutations in the zebrafish homolog of MMACHC. *mmachc*$^{hg12/hg12}$ and *mmachc*$^{hg13/hg13}$ fish display many of the classic phenotypes of cblC disease, including growth retardation, impaired survival, anemia, and metabolite perturbations. Hydroxocobalamin (OH-cbl) injections typically administered to the patients to ameliorate some of the disease-related complications. Although hypothetical, the effectiveness of other potential therapeutics, such as methylcobalamin (Me-cbl) and methionine, has not been evaluated. Growth parameters improved significantly when *mmachc*$^{hg13/hg13}$ fish were maintained in water supplemented with Me-cbl (200 μg/ml) or methionine (10 mM), Standard length (SL) increased by 21% and 16% ($p<0.0001$), respectively, and height at anterior of anal fin (HAA) increased by 63% and 61% ($p<0.0001$), respectively, compared to the untreated group. These changes represent a marked improvement compared to OH-cbl or betaine treatment (SL: 14% and 13%; HAA: 31% and 38%; $p<0.0001$). To visualize the cblC-related eye disease, we bred heterozygote *mmachc* fish with *Tg(rho:EGFP,gnat2:tdTomato)* fish, a double transgenic line expressing GFP in the rods and RFP in the cones. Using confocal microscopy, we observed a thinner photoreceptor layer in *mmachc*$^{hg13/hg13}$*;Tg(rho:EGFP,gnat2:tdTomato)* mutants as well as a reduction in outer segment material present in the retinal pigment epithelium. To understand the underlying gene expression changes associated with the eye phenotype, we performed microarray analysis on eyes dissected from *mmachc*$^{+/hg13}$ and *mmachc*$^{hg13/hg13}$ fish. 327 genes involved in cholesterol metabolism, phototransduction, cAMP signaling, and oxidant stress emerged as differentially regulated (>2-fold change and $p<0.05$). Our zebrafish model recapitulates the phenotypic and biochemical features of cblC disease and demonstrates a response to conventional and novel therapeutics, and will be important to further delineate the pathophysiological mechanism and assess additional therapies.

## 391

**A trans-ethnic genome-wide association study of 21,483 cases and 97,977 controls identifies 27 genetic susceptibility variants for atopic dermatitis.** *L. Paternoster[1], M. Standl[2], H. Baurecht[3,4], J. Waage[5], M. Hotze[3], J.A. Curtin[6], K. Bønnelykke[5], D. Glass[7], D.A. Hinds[8], E. Melen[9], P. Sleiman[10], B. Feenstra[11], M. Pino-Yanes[12], H.T. den Dekker[13], M. Bustamante[14], I. Marenholz[15], B. Jacobsson[16,17], A.D. Irvine[18], A.C. Alves[19], M.M. Groen-Blokhuis[20], A. Franke[21], M. Ferreira[22], M. Tamari[23], N. Probst-Hensch[24], K. Williams[25], D.P. Strachan[26], S.J. Brown[27], J. Heinrich[2], D.M. Evans[1,28], S. Weidinger[3] on behalf of the EAGLE Eczema Consortium.* 1) MRC IEU, University of Bristol, Bristol, UK; 2) Institue of Epidemiology I, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; 3) Department of Dermatology, Allergology, and Venereology, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany; 4) Institute of Genetic Epidemiology, Helmholtz Zentrum Munchen, German Research Center for Environmental Health, Neuherberg, Germany; 5) Copenhagen Prospective Studies on Asthma in Childhood, Faculty of Health and Medical Sciences, University of Copenhagen & Danish Pediatric Asthma Center, Gentofte Hospital, University of Copenhagen, Denmark; 6) Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University of Manchester and University Hospital of South Manchester, Manchester, UK; 7) Department of Twin Research and Genetic Epidemiology, King's College London, London, UK; 8) 23andMe, Inc., Mountain View, California, USA; 9) Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; 10) The Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pensylvania, USA; 11) Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark; 12) Department of Medicine, University of California, San Francisco, California, USA; 13) The Generation R Study Group; Department of Pediatrics, division of Respiratory Medicine; Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands; 14) Center for Research in Environmental Epidemiology (CREAL), Center for Genomic Regulation (CRG), Pompeu Fabra University (UPF), CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain; 15) Experimental Clinical Research Center, Charité Medical Faculty and Max-Delbrueck-Center for Molecular Medicine, Berlin, Germany; 16) Department of Genes and Environment, Division of Epidemiology, Norwegian Institute of Public Health, Oslo, Norway; 17) Department of Obstetrics and Gynecology, Sahlgrenska University Hospital, Sahlgrenska Academy, Göteburg University, Göteburg, Sweden; 18) Paediatric Dermatology, Our Lady's Children's Hospital Crumlin, Dublin, Ireland; National Children's Research Centre, Our Lady's Children's Hospital Crumlin, Dublin, Ireland; Clinical Medicine, Trinity College Dublin, Dublin, Ireland; 19) Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK; 20) Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands; 21) Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany; 22) Queensland Institute of Medical Research Berghofer, Brisbane, Queensland, Australia; 23) Laboratory for Respiratory Diseases, Center for Genomic Medicine, RIKEN, Yokohama, Japan; 24) Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute Swiss TPH, Basel, Switzerland; University of Basel, Basel, Switzerland; 25) Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, MI, USA; Department of Medicine, Henry Ford Health System, Detroit, MI, USA; 26) Population Health Research Institute, St George's, University of London, London, UK; 27) Dermatology and Genetic Medicine, Medical Research Institute, University of Dundee, Dundee, UK; 28) University of Queensland Diamantina Institute, Translational Research Institute, University of Queensland, Brisbane, Australia.

Atopic dermatitis (AD) is a highly heritable common chronic-inflammatory skin disease. Previous genome-wide association studies (GWAS) have identified 19 associated loci. The largest previous GWAS involved around 5,000 cases and 20,000 controls, included only Europeans and identified 3 associated loci. This puts AD some way behind many other complex common diseases that have been studied in much larger numbers of individuals and identified far greater numbers of loci. We conducted the world's largest GWAS of AD, including studies of different ethnicities and using data imputed to the 1000 genomes reference panel. Our discovery cohort consisted of 21,483 cases and 97,977 controls from 25 studies. In addition to carrying out a fixed effects genome-wide meta-analysis of European individuals, we also included Japanese, Latin-American and African-American individuals in a trans-ethnic meta-analysis (using MANTRA). We identified 27 loci associated with AD at genome-wide significance (11 novel). Due to differences in patterns of linkage disequilibrium between ethnicities, we were able to refine the regions of interest for some previously identified loci whereas other loci showed population-specific associations. One particularly interesting association involved a series of SNPs near *CD207*, a gene selectively expressed in dendritic cells of the epidermis (Langerhans cells), putatively involved in antigen uptake and processing. These SNPs also showed strong association with *CD207* expression in skin tissue from the MuTHER study ($p=2\times10^{-10}$). Thus, our results suggest a possible role for genetic factors influencing epithelial dendritic cell function in the aetiology of eczema. Our results also show a substantial genetic overlap of AD with immune-mediated diseases, particularly inflammatory bowel disease (IBD). 39 of 163 SNPs robustly associated with IBD in a recent GWAS were at least nominally associated ($p<0.05$) with AD in our sample (34 of these in the same direction). A gene-set enrichment analysis using MAGENTA identified 38 significantly enriched gene-sets (FDR<0.05) out of a total of >10,000 tested, and 9 additional SNPs (with $p<10^{-5}$) to include in the replication phase. These gene sets predominantly belong to T cell proliferation and differentiation, cytokine and chemokine signalling, and NF-kappaB signalling pathways. 24 polymorphisms newly associated with AD are currently being tested in over 200,000 individuals in a replication phase.

## 392

**Trans-ancestral ImmunoChip: SLE risk loci show enrichment for NK cytotoxicity and Cell Adhesion Pathways.** *D.S. Cunninghame Graham[1], J.A Kelly[2], C.D. Langefeld[3], R.R Graham[4], P.M. Gaffney[2], T.J. Vyse[1], SLE ImmunoChip Consortium.* 1) King's College, London, United Kingdom; 2) Oklahoma Medical Research Foundation, OA; 3) Wake Forest School of Public Health Genomes, NC; 4) Genentech Inc, CA.

**Background:** The prototypical autoimmune disease Systemic Lupus Erythematosus (SLE) shows variation in prevalence and disease severity across ancestries. We and others have shown differences in the identity and frequencies of susceptibility alleles at associated loci for different ancestries. Multiple genome-wide association analysis in European and in SE Asian samples have identified many susceptibility loci (OR>1.3). However low genotyping density has prevented pinpointing of risk alleles, so the vast majority of moderate risk alleles (OR 1.1-1.3) remain either undiscovered or unconfirmed. **Aim:** To identify novel susceptibility loci for SLE across multiple ancestries we genotyped large cohorts of affected individuals with shared controls from three different ethnicities on the Immunochip platform and undertook network analyses to discover underlying biological relevance of these loci in SLE pathogenesis. **Study Cohort:** All SLE cases met the American College of Rheumatology criteria. The ratios of cases and controls ($n_{SLE:nc}$) in each population were: African American (AA)(2970:2452), European (EA) (6748:11516)and Hispanic American (HA) (1872:2016) were genotyped and passed quality control - a total study cohort of 12147 cases and 16732 controls. **Analysis methods:** Association was modelled using a logistic regression adjusting for admixture proportions. Trans-racial non-parametric meta-analysis was computed using METAL. In each ancestry functional pathways were investigated using GSA-SNP. False discovery rate (FDR) adjusted p-values were computed to account for the actual number of tests. **Results:** Numerous regions met Immunochip-wide significance ($P<1\times10\text{-}8$) and FDR<0.05: AA (11), EA (52) and HA (25), including novel and fine-mapping of known loci. Heterogeneity of odds ratio plots between populations from the trans-ancestral meta-analysis indicated greatest heterogeneity in regions including the HLA region, chromosomes 8 and 17. Preliminary pathway analysis revealed an enrichment of risk loci implicated in natural killer cell cytoxicity (hsa04650) (including *BRAF*, *VAV2*, *KRAS*, *SHC4*, *GRB2*, *VAV1* and *IFNAR*) and in cell adhesion (hsa04514) (including *NTXN1*, *NCAM1*, *ITGA8*, *GLG1*, *PTPRM*, *CDH4*, *ICAM1*) ($P_{corr}<0.01$ in any of the three ethnicities). **Conclusion:** Trans-ancestral mapping of known autoimmune loci has revealed novel SLE-associated loci and signalling pathways important in immunological function, further extending our understanding of lupus pathogenesis.

## 393

**Eight amino acid positions in five HLA class I and II genes explain the MHC association to type 1 diabetes risk.** *X. Hu[1,2,3,4,5,6], B. Han[1,2,3,4,5], S. Onengut-Gumuscu[7], W. Chen[7], A.J. Deutsche[1,2,3,4,5,6], T.L. Lenz[2,5], P.I.W. de Bakker[8,9], S.S. Rich[7], S. Raychaudhuri[1,2,3,4,5,10].* 1) Division of Rheumatology, Immunology and Allergy, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA; 2) Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA; 3) Partners Center for Personalized Genetic Medicine, Boston, MA, USA; 4) 3.Partners Center for Personalized Genetic Medicine, Boston, MA, USA Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 5) Harvard Medical School, Boston, MA USA; 6) Harvard-MIT Division of Health Sciences and Technology, Boston, MA USA; 7) Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA; 8) Department of Medical Genetics, University Medical Center Utrecht, Utrecht, Netherlands; 9) Department of Epidemiology, University Medical Center Utrecht, Utrecht, Netherlands; 10) Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK.

Type 1 diabetes (T1D) is a highly heritable metabolic disorder caused by autoimmune destruction of pancreatic cells. Variation in the human leukocyte antigen (HLA) genes, encoding the major histocompatibility complex (MHC) molecules, explains approximately 50% of phenotypic variance for T1D. It is postulated that certain amino acids in the MHC peptide-binding grooves may alter antigen binding or presentation. However high levels of polymorphism and extensive linkage disequilibrium prohibited fine-mapping in the region; for decades, risk was not robustly attributed to specific amino acids other than DQβ1-57. Recent developments in statistical imputation enabled HLA fine-mapping through accurate in silico typing of classical allelic and amino acid polymorphisms. Here we imputed and tested HLA variants in 16,085 Caucasian T1D patients and controls. Method: We genotyped 6,670 T1D patients and 9,416 healthy controls collected through the Type 1 Diabetes Genetics Consortium on the ImmunoChip platform. Using SNP2HLA and a reference panel of individuals with typed classical alleles, we imputed 424 2- and 4-digit alleles and 399 amino acid residues in eight HLA class I and II genes. We performed logistic regression and conditional analyses to evaluate the effect of each variant. Results: We confirmed that the most statistically significant association was at DQβ1-57 (p=10-917), which explained most of the DQB1 association. Conditioning on this position, we identified a comparably strong independent association signal at DRβ1-13 (p=10-570). At this position, histidine conferred risk (OR=4.01) while arginine conferred protection (OR=0.08) against T1D. Conditioning on these two positions, we identified six additional amino acid positions that independently conferred risk: DRβ1-71 (p=10-54), B-158 (p=10-54), A-105 (p =10-39), DRβ1-86 (p=10-26), C-24 (p=10-18), and C-173 (p=10-11). Discussion: These eight positions almost completely explained the T1D association across the MHC. The two most significant positions, DQβ1-57 and DRβ1-13, showed stronger association than any of the respective classical alleles, underscoring the importance of comprehensive analysis including amino acid residues. DRβ1-13 was previously unreported in T1D, but known to confer risk to rheumatoid arthritis and follicular lymphoma, suggesting its potentially common role in autoimmune diseases. Our results may aid in discovering autoantigens and physiochemical basis of peptide-MHC-T cell binding.

## 394

**Increased Risk of Rheumatoid Arthritis (RA) among Shared Epitope-negative (SE-) Mothers with Shared Epitope-positive (SE+) children: Results from the Mother-Child Immunogenetic Study in Autoimmunity (MCIS).** *G.I. Cruz[1], L.A. Criswell[2], X. Shao[1], H. Quach[1], J.A. Noble[3], N.A. Patsopoulos[4], M.P. Busch[5], L.F. Barcellos[1,6].* 1) Genetic Epidemiology and Genomics Lab, Division of Epidemiology, School of Public Health, University of California Berkeley, Berkeley, CA; 2) Rosalind Russell / Ephraim P. Engleman Rheumatology Research Center, Department of Medicine, University of California San Francisco, San Francisco, CA; 3) Children's Hospital Oakland Research Institute, Oakland, CA; 4) Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Department of Neurology, Brigham & Women's Hospital, Boston, MA; Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston,; 5) Blood Systems Research Institute, San Francisco, CA; 6) California Institute for Quantitative Biosciences (QB3), University of California Berkeley, Berkeley, CA.

RA (RA [MIM 180300]) disproportionately affects women of reproductive age, implicating pregnancy-related factors. Fetal microchimerism (FMC), or the persistence of a small population of cells in the mother, is a natural consequence of pregnancy. FMC is present more often in RA cases than in controls. Mother-child histocompatibility could determine long-term FMC, possibly increasing risk of RA through exposure to fetal HLA-antigens. We hypothesized that RA cases are more likely to have histocompatible (HC) children compared to controls. The MCIS included +5,000 individuals; mothers with RA or SLE and controls (n=750), their children and fathers. RA cases with 1+ birth before diagnosis were recruited at UC San Francisco. Controls were primarily recruited from blood donors. Mothers provided information on their reproductive history, history of transfusion, transplant and infections. Comprehensive MHC region SNP genotyping was conducted using the Illumina MHC panel (n=1,783), ImmunoChip (n=8,842), and 660K (n=1,991) arrays. Four-digit genotype data for *HLA-A, B, C, DPA1, DPB1, DQA1, DQB1* and *DRB1* were imputed using the T1DGC reference panel and BEAGLE. We estimated ancestry proportions from 384 markers using STRUCTURE. A child was HC from the mother's perspective if the paternal allele did not differ from the non-inherited maternal allele. Carrier status (+ or -) of any *DRB1* allele associated with RA risk (Raychaudhuri, 2012) and corresponding to SE amino acid sequences QKRAA and QRRAA (01:01, 04:01, 04:04, 04:05, 04:08) and DERAA (01:03, 04:02, 11:02, 13:01, 13:02) was determined for mothers and children. We used logistic regression models to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for the association between RA and a) HC at each HLA locus and b) exposure to any SE+ and DERAA+ children, stratifying on maternal carrier status. Increased HC among cases was only evident at *DQB1* (25.3% vs. 17.4%, p=0.03). Having any SE+ children significantly increased risk of RA for SE- mothers (n=218) (OR 2.56; 95% CI, 1.43-4.58) but not SE+ mothers (n= 248) (OR 1.48; 95% CI, 0.83-2.62). No association was found with DERAA+ children, regardless of maternal carrier status. Ancestry, parity, and history of transfusion did not impact results. Exposure to SE+ children and *DQB1* HC may contribute to RA etiology and could contribute to RA's female-predominance. This is the largest study confirming the association between RA and SE+ children in SE- mothers.

## 395

**Steroid-responsive genes play a major role in the genetic basis of sexual dimorphism in complex human disease.** *L.A. Weiss[1,2], K.M. Tsang[1,2], B. Adviento[1,2], K.A. Aldinger[3], H. Lee[4], K. Kim[5], R.J. Schmidt[5,6], S.F. Nelson[4,7], P. Levitt[8,9], D.G. Amaral[6,10,11], I. Hertz-Picciotto[5,6], C. Ladd-Acosta[12], M.D. Fallin[13,14], L.A. Croen[15], N. Zaitlen[2,16].* 1) Psychiatry, University of California San Francisco, San Francisco, CA; 2) Institute for Human Genetics, University of California San Francisco, San Francisco, CA; 3) Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA; 4) Department of Pathology and Laboratory Medicine, University of California Los Angeles, Los Angeles, CA; 5) Public Health Sciences, University of California Davis Medical School, Davis, CA; 6) MIND Institute, University of California Davis, Davis, CA; 7) Human Genetics, University of California Los Angeles, Los Angeles, CA; 8) Institute for the Developing Mind, Children's Hospital Los Angeles, Los Angeles, CA; 9) Keck School of Medicine, University of Southern California, Los Angeles, CA; 10) on behalf of the Autism Phenome Project team; 11) Psychiatry and Behavioral Sciences, University of California Davis, Davis, CA; 12) Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; 13) Wendy Klag Center for Autism and Developmental Disabilities, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; 14) Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; 15) Autism Research Program, Kaiser Permanente Division of Research, Oakland, CA; 16) Medicine, University of California San Francisco, San Francisco, CA.

Most complex diseases show sexual dimorphism in prevalence or presentation, sometimes accounting for a large proportion of liability, such as in neurodevelopmental disorders (4-5 fold increased autism prevalence in males) and autoimmune disease (2-3 fold increased rheumatoid arthritis prevalence in females). There have been a multitude of theories to explain this phenomenon, including the contribution of sex chromosomes, the influence of steroid hormones, cultural biases in socialization, and sex-limited phenomena such as pregnancy. In this study, we leverage statistical approaches to examine genetic hypotheses utilizing autism and rheumatoid arthritis (RA) as examples of male-biased and female-biased conditions.

Our autism dataset consists of imputing from published GWAS datasets as well as uniquely genotyped samples, using family-based and case-control designs to examine 7,462 affected males and 1,503 females. We compared these results with the RA WTCCC dataset (473 affected males, 1,385 females). First, we utilized GCTA to assess SNP-based autosomal heritability ($h^2_g$) in males and females separately. Second, we leveraged a local $h^2_g$ approach (Gusev *et al* 2014 arXiv) to determine the phenotypic variance explained by estrogen- and androgen-responsive genes. To support both approaches, we also used an FDR-based test for enrichment of association signal. Finally, we use both GCTA and FDR-enrichment to examine the contribution of the X chromosome. Sex and gene set permutations tests were used to assess significance where appropriate.

For the first time, we show that autosomal SNP-based $h^2_g$ of several diseases is higher in the lower-prevalence sex ($P$<0.01). We observe excess $h^2_g$ for androgen-responsive genes in females across diagnoses ($P$<0.01). Further, we show an overabundance of high test statistics in androgen-responsive genes in male-prevalent conditions ($P$=0.05) and in estrogen-responsive genes in female-prevalent conditions ($P$<0.01). Finally, we quantify the contribution of the X chromosome to heritability by sex.

In summary, we support the liability threshold model for common variants, describe the contribution of the X chromosome, as well as implicate a specific and major role for hormone-responsive genes in complex disorders. Our results increase understanding of the underlying mechanism of sexually-dimorphic risk and carry important implications for future studies.

## 396

**Functional characterization of a multiple sclerosis associated variant in IL7Ra.** *S.G. Gregory[1], G. Galarza-Muñoz[2], F.B. Briggs[3], L. Bergamaschi[1], S. Arvai[1], X. Shao[4], L.F. Barcellos[4], M.A. Garcia-Blanco[2].* 1) Medicine, Duke Molecular Physiology Institute, Durham, NC; 2) Molecular Genetics and Microbiology, Duke University medical Center, Durham, NC; 3) Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH; 4) School of Public Heath, University of California Berkeley, Berkeley, CA.

PURPOSE: We previously identified a non-synonymous SNP, rs6897932, in the interleukin 7 receptor alpha chain (IL-7Rα) as genetically associated with multiple sclerosis (MS). In vitro functional analysis established that the "C" allele of rs6897932 increases exon 6 'skipping' altering the ratio of membrane and soluble receptor isoforms. We hypothesized that rs6897932 influences T cell IL-7 signaling in an allele specific manner, and that the trans-acting proteins that regulate IL-7Rα splicing could be MS susceptibility candidates. METHODS: Tobramycin affinity chromatography was used to identify exon 6 trans-acting proteins and their impact on splicing was validated by siRNA-based screening in depleted HeLa cells. These proteins were screened for genetic association in MS case/control cohorts using IMSGC data. Finally, we assessed allele specific signaling of rs6897932 by measuring phosphorylated STAT5 in primary T cell cultures from relapsing-remitting MS patients and in HEK293 cell lines transiently transfected with "C" or "T" alleles of rs6897932. RESULTS: We established that CPSF1, PTBP1, and DDX39B regulate alternative splicing of exon 6 of IL-7Rα and that SNPs within DDX39B are genetically associated with MS after adjusting for population stratification, gender, and known MS risk haplotypes. Analysis of soluble IL-7Rα in T-cells show dosage "C" allele dosage effects and IL-7 signaling was reduced in the presence of soluble IL-7Rα in "C" and "T" transiently transfected cell lines. Phospho-STAT5 analysis in T-cells and cell lines do not show differences in allele specific IL-7 signaling.

CONCLUSIONS: We established that it is the level of soluble IL-7Rα and not genotype that are important for IL-7 signaling. We therefore focused on the etiological relationship between on the trans-acting protein regulation of IL-7Rα exon 6 'skipping' and MS. Our in vivo studies have established new mechanistic relationships between three proteins that regulate IL-7Rα splicing. After adjusting for multiple factors we also identified DDX39B as a novel genetically associated MS candidate gene. These data emphasize the significance of mRNA splicing in the development of autoimmune disease and illustrate the importance of establishing functional pathways to identify novel MS genes that would not have been identified using traditional genetic approaches.

## 397

**UBE2L3 polymorphism amplifies NF-kB activation and promotes B cell development linking linear ubiquitination to multiple autoimmune diseases.** *M.J. Lewis[1], S. Vyse[1], A.M. Shields[2], S. Boeltz[2], D. Leirer[2], P.A. Gordon[6], T.D. Spector[3], P.J. Lehner[5], H. Walczak[4], T.J. Vyse[2].* 1) Experimental Medicine and Rheumatology, Queen Mary University of London, London, United Kingdom; 2) Medical and Molecular Genetics, King's College London, London, United Kingdom; 3) Twin Research Unit, King's College London, London, United Kingdom; 4) Centre for Cell Death, Cancer and Inflammation, UCL Cancer Institute, University College London, London, United Kingdom; 5) Cambridge Institute for Medical Research, University of Cambridge, Cambridge, United Kingdom; 6) Rheumatology Dept, King's College Hospital, London, United Kingdom.

**Background:** Genome-wide association studies have identified a strong association between a single risk haplotype of the *UBE2L3* gene and Systemic Lupus Erythematosus (SLE), as well as multiple autoimmune diseases (rheumatoid arthritis, juvenile idiopathic arthritis, ulcerative colitis, Crohn's disease, coeliac disease, psoriasis). UBE2L3 is a highly specific E2 ubiquitin-conjugating enzyme. Linear ubiquitination is a newly described form of ubiquitination, whose only known function is controlling activation of NF-kB, mediated by the linear ubiquitination chain assembly complex (LUBAC). **Results:** Data from SLE GWAS, imputed to 1000 Genomes level identified rs140490 as the most strongly associated *UBE2L3* SNP, located at -270bp of the promoter region ($P=8.6\times10^{-14}$; OR 1.30, 95%CI: 1.21-1.39). Microarray / western blot studies found that the rs140490 risk allele increased UBE2L3 expression in B cells and monocytes from PBMC. Overexpression of UBE2L3 in combination with LUBAC in HEK293-NF-kB reporter cell line led to a marked upregulation in NF-kB activity, which was abolished by dominant-negative mutant UBE2L3[C86S]. RNAi blockade of UBE2L3 antagonised TNF signalling by inhibiting phosphorylation and degradation of the NF-kB sequestration protein IκBα. *Ex vivo* human B cells and monocytes were isolated from genotyped healthy twins stimulated with CD40L or TNF respectively and NF-kB translocation quantified by Imagestream analysis. rs140490 genotype was correlated with both basal NF-kB activation in healthy human individuals, as well as the sensitivity of NF-kB to CD40 stimulation in B cells and TNF stimulation in monocytes. Consistent with this functional effect of *UBE2L3* on CD40 signalling in B cells, rs140490 genotype correlated with increased plasmablast and plasma cell differentiation in SLE patients ($P<0.001$). **Conclusion:** This is the first study to show that the *UBE2L3* risk haplotype exerts a critical rate-limiting effect on TNF and CD40 signalling in primary human cells. Our functional data show the critical importance of UBE2L3 in regulation of LUBAC and NF-kB activation. By tracking NF-kB nuclear translocation in B cells and monocytes from genotyped individuals, this is the first demonstration that a complex trait variant at *UBE2L3* regulates both basal NF-kB activation and sensitivity of NF-kB to stimulation in *ex vivo* cells, resulting in accelerated B cell differentiation in SLE.

## 398

**A recombination allele of the lipase gene *CEL* and its pseudogene *CELP* encodes a hybrid protein and is a genetic risk factor for chronic pancreatitis.** *K. Fjeld[1,2], J. Rosendahl[3], J.M. Chen[4], D. Lasher[5], M. Cnop[6], B.B. Johansson[1], M. Ringdal[1], E. Masson[4], J. Mayerle[7], J. Mössner[3], C. Ruffert[3], S. Steine[8], E. Tjora[1], J. Torsvik[1], C. Ferec[4], F.U. Weiss[7], H. Witt[5], M.M. Lerch[7], P.R. Njølstad[1], S. Johansson[1,2], A. Molven[1,8,9].* 1) Department of Clinical Science, University of Bergen, Norway; 2) Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway; 3) Department for Internal Medicine, Neurology and Dermatology, Division of Gastroenterology, University of Leipzig, Germany; 4) Institut National de la Santé et de la Recherche Médicale (INSERM), U1078, Etablissement Français du Sang (EFS)-Bretagne, France; 5) Else Kröner-Fresenius-Zentrum für Ernährungsmedizin (EKFZ), Technische Universität München (TUM), Freising, Germany; 6) Laboratory of Experimental Medicine, Université Libre de Bruxelles, Belgium; 7) Department of Internal Medicine A, Ernst-Moritz-Arndt University, Germany; 8) Gade Laboratory for Pathology, Department of Clinical Medicine, University of Bergen, Norway; 9) Department of Pathology, Haukeland University Hospital, Bergen, Norway.

**Introduction:** Carboxyl-ester lipase (CEL) is a digestive enzyme that plays an important role in hydrolysis and absorption of cholesterol and lipid-soluble vitamin esters. Human *CEL* consists of 11 exons and the last exon is highly polymorphic, containing a variable number of tandem repeats (VNTR). We have previously described a monogenic syndrome of diabetes and exocrine pancreatic dysfunction caused by mutations in the *CEL* VNTR. Copy number variations (CNVs) of the *CEL* gene have also been reported although their detailed structures were not worked out. **Objectives:** The aim of this study was to characterize the structure of *CEL* CNVs and to explore their role in chronic pancreatitis (CP). **Methods:** We developed long-range PCR-based assays for screening of *CEL* CNVs. German and French materials of idiopathic CP (n=1136), alcoholic-related CP (n=853), and controls (n=4630) were investigated. For functional analysis, we transfected human embryonic kidney (HEK293) and mouse acinar (266-6) cells with CEL constructs, and performed Western blotting, immunostaining, confocal microscopy, deglycosylation treatment, enzyme activity measurements, and quantitative RT-PCR. **Results:** We identified two CNVs of *CEL*, one duplication and one deletion allele, probably resulting from non-allelic homologous recombination between *CEL* and its neighbouring pseudogene *CELP*. The deletion variant, a *CEL-CELP* hybrid allele (*CEL-HYB*), was strongly associated with idiopathic CP (meta-analysis: OR=6.4, $P$=1.1 x $10^{-16}$). The *CEL-HYB* variant was also found in 15/853 (3.9%) subjects with alcoholic CP compared with 26/3409 (0.8%) controls ($P$=0.016). *CEL-HYB* encodes a CEL protein where the coding VNTR sequence of CEL has been exchanged with that of CELP. This results in a truncated C-terminal of the protein and functional analyses of CEL-HYB showed impaired secretion, reduced enzyme activity, and a stimulation of autophagy. **Conclusion:** We have identified *CEL-HYB* as a new genetic risk factor in chronic pancreatitis. To our knowledge, this is the first lipase associated with the disease. Stimulation of autophagy by CEL-HYB suggests that the disease process may involve mechanisms different from those seen for other genetic risk factors for pancreatitis.

## 399

**Clozapine-induced agranulocytosis/granulocytopenia is associated with rare *HLA-DQB1* and *HLA-B* alleles.** *J.I. Goldstein[1,2], L.F. Jarskog[3], I. Cascorbi[4], M. Dettling[5], A.K. Malhotra[6,7,8], J. Nielsen[9,10], D. Rujescu[11,12], T. Werge[13,14,15], D.L. Levy[16,17], R.C. Josiassen[18], J.L. Kennedy[19], J.A. Lieberman[20], M.J. Daly[1,2], P.F. Sullivan[3,21,22], Clozapine Induced Agranulocytosis Consortium.* 1) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; 2) Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA; 3) Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA; 4) Institute of Experimental and Clinical Pharmacology, University Hospital Schleswig-Holstein, Kiel, Germany; 5) Department of Psychiatry and Psychotherapy, Charité-University Medicine, Berlin, Germany; 6) The Feinstein Institute for Medical Research, Manhasset, NY, USA; 7) The Hofstra NS-LIJ School of Medicine, Hempstead, NY, USA; 8) The Zucker Hillside Hospital, Glen Oaks, NY, USA; 9) Aalborg University Hospital, Psychiatry, Aalborg, Denmark; 10) Department of Clinical Medicine, Aalborg University, Denmark; 11) Department of Psychiatry, University of Halle, Halle, Germany; 12) Department of Psychiatry, University of Munich, Munich, Germany; 13) Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; 14) Institute of Biological Psychiatry, MHC Sct. Hans, Mental Health Services Copenhagen, Denmark; 15) The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark; 16) Department of Psychiatry, Harvard Medical School, Boston, MA, USA; 17) McLean Hospital, Belmont, MA, USA; 18) Department of Psychiatry, Drexel University, Philadelphia, PA, USA; 19) Center for Addiction and Mental Health, Toronto, Canada; 20) Department of Psychiatry, Columbia University and the New York State Psychiatric Institute, New York, NY, USA; 21) Department of Genetics, University of North Carolina, Chapel Hill, NC, USA; 22) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

Clozapine is a particularly effective antipsychotic medication but its use is curtailed by the risk of clozapine-induced agranulocytosis/granulocytopenia (CIAG), a severe adverse drug reaction occurring in up to 1% of treated individuals. CIAG can be fatal if not detected early. As a result, clozapine is underused despite its superior efficacy and particular utility in treatment-resistant schizophrenia. Identifying genetic risk factors for CIAG could enable safer and more widespread use of clozapine. Using multiple ascertainment schemes, we assembled the largest CIAG cohort to date (163 cases, 54.0% never previously reported). First, we performed a genome-wide association study (GWAS) in 161 CIAG cases and 1,196 controls of European ancestry. Second, we tested rare protein-coding variants for association using whole-exome sequencing data from 67 CIAG cases and 376 untreated population controls. For variants that were also genotyped on the exome array, we merged the sequencing data with exome array data from 81 CIAG cases and 3,294 controls of European ancestry. Finally, we imputed classical HLA alleles and amino acids. The SNP with the best evidence for association from the GWAS was in an intron in *HLA-B* (rs41549217, $P$=2.07e-7, OR=4.66, 95% CI 2.6-8.3). The top finding from the exome study was a low frequency missense variant in *BTNL2* (rs28362679, $P$=4.14e-7, OR=3.9, 95% CI 1.2-12.9). *BTNL2*, located in the major histocompatibility complex (MHC), was the only significant gene ($P$=7.0e-8) from a gene burden test for rare functional variants. From classical HLA allele imputation, we found two loci in the MHC to be independently associated with CIAG: a single amino acid in *HLA-DQB1* (126Q) ($P$=4.7e-14, OR=0.19, 95% CI 0.12-0.29) and an amino acid change in the extracellular binding pocket of *HLA-B* (158T) ($P$=6.4e-10, OR=3.3, 95% CI 2.3-4.9). Finally, we used a likelihood ratio test to determine that the two amino acid changes in *HLA-DQB1* and *HLA-B* were 23,000 times more likely to explain the association in the MHC region than the variants in *BTNL2*. Although our data clarify the contributions of HLA variation to CIAG, the odds ratios do not immediately suggest clinical application in screening. However, our genetic insights could further understanding of the biological process underlying CIAG and, as other non-genetic and genetic risk factors for CIAG are identified, constitute an important component of tests to screen patients for the safer use of clozapine.

## 400

**New Susceptibility Gene IKZF1 for Cold Medicine-Related Stevens-Johnson Syndrome/Toxic Epidermal Necrolysis with Severe Mucosal Involvement.** *M. UETA[1,2], H. Sawai[3], C. Sotozono[1], Y. Hitomi[3], N. Kaniwa[4], MK. Kim[5], KY. Seo[6], KC. Yoon[7], CK. Joo[8], C. Kannabiran[9], TH. Wakamatsu[10], V. Sangwan[11], V. Rathi[11], S. Basu[11], T. Ozeki[12], T. Mushiroda[12], E. Sugiyama[4], K. Maekawa[4], R. Nakamura[4], M. Aihara[13], K. Matsunaga[14], A. Sekine[15], JAP. Gomes[10], J. Hamuro[1], Y. Saito[4], M. Kubo[12], S. Kinoshita[1], K. Tokunaga[2].* 1) Kyoto Prefectural University of Medicine, Kyoto, Japan; 2) Research Center for Inflammation and Regenerative Medicine, Faculty of Life and Medical Sciences, Doshisha University, Kyoto, Japan; 3) Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; 4) Division of Medicinal Safety Science, National Institute of Health Sciences, Tokyo, Japan; 5) Department of Ophthalmology, Seoul National University College of Medicine, Seoul, Korea; 6) Department of Ophthalmology, Severance Hospital, Institute of Vision Research , Yonsei University College of Medicine, Seoul, Korea; 7) Department of Ophthalmology, Chonnam National University, Gwangju, Korea; 8) Department of Ophthalmology & Visual Science, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea; 9) Prof Brien Holden Eye Research Centre, L V Prasad Eye Institute, Hyderabad, India; 10) Department of Ophthalmology, Federal University of São Paulo, São Paulo, Brazil; 11) Cornea and Anterior Segment Services, L V Prasad Eye Institute, Hyderabad, India; 12) Research Group for Pharmacogenomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan; 13) Department of Environmental Immuno-Dermatology, Yokohama City University Graduate School of Medicine, Yokohama, Japan; 14) Department of Dermatology, Fujita Health University School of Medicine, Toyoake, Japan; 15) EBM Research Center, Kyoto University Graduate School of Medicine, Kyoto, Japan.

Stevens-Johnson syndrome (SJS) and its severe form, toxic epidermal necrolysis (TEN), are acute inflammatory vesiculobullous reactions of the skin and mucous membranes including the ocular surface, oral cavity, and genitals. These reactions are very rare but are often associated with inciting drugs and infectious agents. To identify susceptibility loci for cold medicine-related SJS/TEN (CM-SJS/TEN) with severe mucosal involvement (SMI), a genome-wide association study (GWAS) was performed in 808 Japanese individuals (117 cases and 691 controls). We found that the HLA-A region showed the strongest association with susceptibility to CM-SJS/TEN with SMI. Outside of the HLA region, there were 60 SNPs with $p < 10^{-3}$ in the GWAS. In the 11 SNPs that were $p < 10^{-5}$, loci IKZF1 and TSHZ2 showed especially low p-values. The 10 SNPs of the 11 SNPs with $p < 10^{-5}$, which functional TaqMan probes were available, were studied in a subsequent replication analysis by using an independent set of 208 Japanese samples (20 cases and 188 controls). In this first replication study, there were no SNPs with a significant association after applying the Bonferroni correction because of the relatively small sample size. However, the ORs of 9 SNPs of the 10 SNPs showed the same direction of association as those in the GWAS. Moreover, we genotyped these 10 SNPs in samples from the Korean population (34 cases and 90 controls), which is genetically close to the Japanese population. Although the number of Korean cases was small, we found a significant association between Korean CM-SJS/TEN with SMI and IKZF1. Furthermore, the meta-analysis with Japanese and Korean samples showed a genome-wide significant association with IKZF1. Moreover, in Indian samples (26 case and 56 controls), we also genotyped the SNP of IKZF1. Although the Indian population is genetically close to the European population and the sample size was small, we found significant associations between Indian CM-SJS/TEN with SMI and IKZF1. Furthermore, the meta-analysis with Japanese, Korean and Indian samples showed a genome-wide significant association between CM-SJS/TEN with SMI and IKZF1 ($p = 7.6\times10^{-11}$) . Furthermore, quantitative ratios of IKZF1 alternative splicing isoforms, Ik1 and Ik2, were significantly associated with the genotypes. These results indicate that the ratio of Ik2/Ik1 may be influenced by IKZF1 SNPs, which were significantly associated with susceptibility to CM-SJS/TEN with SMI.

## 401

**Prospective participant selection and ranking to maximize actionable PGx variants and discovery in the eMERGE Network.** *D. Crosslin[1,2], A. Gordon[1], P. Robertson[2], D. Hanna[2], D. Carrell[3], A. Scrol[3], I. Kullo[4], M. de Andrade[5], E. Baldwin[3], J. Grafton[3], K. Doheny[6], P. Crane[7], R. Li[8], S. Stallings[9], S. Verma[10], J. Wallace[10], M. Ritchie[10], M. Dorschner[2], E. Larson[3], D. Nickerson[2], G. Jarvik[1,2], The electronic Medical Records and Genomics (eMERGE) Network.* 1) Genome Sciences / Medical Genetics, University of Washington, Seattle, WA; 2) Department of Genome Sciences, University of Washington, Seattle, WA; 3) Group Health Research Institute, Center for Health Studies, Seattle, WA; 4) Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN; 5) Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN; 6) Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD; 7) Division of General Internal Medicine, University of Washington, Seattle, WA; 8) Office of Population Genomics, National Human Genome Research Institute, Bethesda, MD; 9) Department of Biomedical Informatics, Vanderbilt University, Nashville, TN; 10) Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA.

Some 9,000 participants in the eMERGE Network are being sequenced with the targeted Pharmacogenomics Research Network sequence platform (PGRNseq), thus linking electronic health records (EHR) to pharmacogenetic variant data to ultimately return actionable results. PGRNseq contains the coding regions, UTRs, and 2kb upstream for 84 pharmacogenes. To return CLIA results to participants at the Group Health Cooperative, we initially sequenced DNA from ≈900 participants (61% female) and selected 450 of these to re-consent, redraw, and ultimately validate variants. We designed an algorithm to harness data from ancestry, diagnosis codes, medication records, laboratory results, and variant-level bioinformatics to ensure selection of an informative sample for this project. The algorithm involved two steps. We enriched our sample for diversity by over-selecting for non-European ancestry participants, which included African (5%) and Asian (8%) ancestry. We enriched for participants with EHR evidence of actionable indications related to PGRNSeq genes, including malignant hyperthermia, long QT syndrome, hypertension, atrial fibrillation, congestive heart failure, and elevated creatine kinase values within six months of a statin medication. We annotated the ≈900 multi-sample VCF by a combination of SeattleSeq and SnpEff, with additional custom variables including evidence from ClinVar, OMIM, and HGMD with links to prior clinical associations. We focused our analyses on 28 actionable genes, largely driven by the Clinical Pharmacogenetics Implementation Consortium. We derived a ranking system based on the number of coding variants per participant (75.2±14.7), and the number of variants with high or moderate impact (11.5±3.9). Notably, we identified 11 stop-gained (1%) and 519 missense (20%) variants out of a total of 1,785 in these 28 genes. Finally, we prioritized variants to be returned to the EHR with prior clinical evidence of pathogenicity or annotated as stop-gain for the following genes: *CACNA1S* and *RYR1* (malignant hyperthermia); *SCN5A*, *KCNH2*, and *RYR2* (arrhythmia); and *LDLR* (high cholesterol). Our analytic pipeline, including participant-level variant indexing, custom annotation, and R and LaTeX scripts, will serve as a foundation for identification of potentially actionable variants and EHR integration. These data will inform pathogenicity of specific variants and practices for EHR integration of genomic data.

## 402

**Real-time Pharmacogenomics: Genetic Factors Impacting Phenylephrine Response During Surgery.** *J.M. Jeff[1], T. Joesph[2], K. Slivinski[1], M. Yee[3], A. Owusu Obeng[1,4,5], S.B. Ellis[1], E.P. Bottinger[1], O. Gottesman[1], M.A. Levin[2,6], E.E. Kenny[1,7,8,9].* 1) Charles F. Bronfman Institute for Personalized Medicine, Ichan School of Medicine at Mount Sinai, New York, NY; 2) Department of Anesthesiology, Icahn School of medicine at Mount Sinai Hospital, New York; 3) Carnegie Institution for Science, Dept. of Plant Biology, Stanford, CA; 4) Division of General Internal Medicine, Icahn School of medicine at Mount Sinai Hospital, New York; 5) Department of Pharmacy, Icahn School of Medicine at Mount Sinai Hospital, New York; 6) Division of Cardiothoracic Anesthesia, Icahn School of Medicine at Mount Sinai Hospital, New York; 7) Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY; 8) Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY; 9) Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, NY.

We present a novel pharmacogenomic approach in which we define real-time drug response to intravenous medication administered during surgery. We have built a database of surgical procedures that capture drug administration events and succeeding physiological responses, as well as type and length of surgery, anesthetic technique, quantity of fluids administered and other patient information. These data are linked to an internal Biobank of patients from New York, with genotype, sequence, and rich phenotypic data from the electronic medical record. Leveraging this resource, we investigated real-time response to phenylephrine, a selective $\alpha$1-adrenergic receptor agonist that is used to treat hypotension during surgery. After extensive analyses of possible confounders, we excluded patients whose mean arterial pressure (MAP) before or after drug bolus was outside the normal MAP range (30-130mmHG), cases with poorly recorded physiological data, and those who received >5L fluids or blood products during surgery, resulting in 866 patients for genetic analyses. We found that administration of a bolus of phenylephrine (50-200$\mu$g) increases MAP an average of 17.25 mmHg(SE±1.29), 13.3 mmHg(SE±1.01), and 15.21 mmHg(SE±0.98), in self-reported European-Americans (EA;n=168), African-Americans (AA;n=208) and Hispanic/Latino (HL;n=292), respectively. Overall EAs have a greater response to phenylephrine compared to HLs and AAs (F=6.24; P<0.008). We performed a GWAS with 1000 Genomes imputed genotypes to test whether genetic variation influences $\Delta$MAP after phenylephrine drug bolus. We confirmed associations in biologically relevant $\alpha$1-adrenergic receptors, *ADRA1A* and *ADRA1B* (p<10-3). Notably, we discovered a significant GWAS signal in EAs (rs2320937;chr2:134341287;P<4.7E-8;$\beta$=-0.71; MAF=0.26) that explains 18% of the variance in drug response and replicated in an independent cohort of EAs (n=50,rs62177704,chr2:134234306; P<1.7E-4, $\beta$=-1.49, MAF=0.07). Homozygous non-reference individuals have an attenuated drug response (2.43 mmHG,SE±6.33) compared to non-carriers (25.75 mmHG,SE±2.33). The signal resides upstream of NCK-associated protein 5 gene, *NCKAP5*, which is involved in the blood coagulation pathway via Factor Xa. This work demonstrates our ability to define real-time drug response, detect large effect alleles and novel genes affecting pharmaceutical response, and identify a subset of phenylephrine 'non-responders' with implications for personalized treatment during surgery.

## 403

**PGRN Network-wide Project: Transcriptome Analysis of Pharmacogenes in Human Tissues.** *C.E. French[1], A. Chhibber[2], E.R. Gamazon[3], S.W. Yee[2], X. Qin[4], E. Theusch[5], A. Webb[6], S.T. Weiss[7,8], M.W. Medina[5], E.G. Schuetz[9], A.L. George[10], R.M. Krauss[5], C.Q. Simmons[10], S.E. Scherer[4], N.J. Cox[3], K.M. Giacomini[2], S.E. Brenner[1].* 1) University of California, Berkeley, CA; 2) University of California, San Francisco, CA; 3) University of Chicago, Chicago, IL; 4) Baylor College of Medicine, Houston, TX; 5) Children's Hospital Oakland Research Institute, Oakland, CA; 6) Ohio State University, Columbus, OH; 7) Brigham and Women's Hospital, Boston, MA; 8) Harvard Medical School, Boston, MA; 9) St. Jude Children's Research Hospital, Memphis, TN; 10) Vanderbilt University, Nashville, TN.

Gene expression variation is crucial to the etiologies of common disorders and the molecular underpinnings of pharmacologic traits; however, the nature and extent of this variation remains poorly understood. The NIH Pharmacogenomics Research Network (PGRN) Network-wide RNA-seq project aims to create a community resource containing quantitative information on annotated and novel isoforms of genes involved in therapeutic and adverse drug response (pharmacogenes). Using 18 samples from each of 5 tissues of pharmacologic importance (liver, kidney, adipose, heart, and lymphoblastoid cell lines [LCLs]), we performed transcriptome profiling by RNA-Seq with the goal of determining differences in expression of pharmacogenes across tissues and between individuals. The data were analyzed for expression quantification, and we used the JuncBASE tool developed by members of our consortium to identify and quantify splicing events. In each of the tissues and LCLs, 11,223-15,416 genes were expressed at a substantial level. In pairwise comparisons of tissues, 105-211 pharmacogenes were differentially expressed (≥2-fold difference, FDR<0.1). For example, as expected, the CYP enzymes *CYP2C19* (MIM 124020) and *CYP2D6* (MIM 124030) were 10-fold and 100-fold more highly expressed in the liver than in other tissues. Other important drug metabolizing enzymes such as *DPYD* (MIM 612779) and *TPMT* (MIM 187680) showed more balanced gene expression patterns across the tissues. We observed that 72-93% of pharmacogenes are alternatively spliced within each tissue. There was substantial variation in both annotated and novel splicing events both between tissues and between individual samples of the same tissue. For example in *SLC22A7* (MIM 604995), a gene encoding a transporter for various drugs, we found evidence of a novel alternative last exon that is variably spliced between individuals. LCLs are important pre-clinical models for human genetic studies, but they highly express less than half of pharmacogenes as compared with the 66-83% expressed at a substantial level in each of the physiological tissues. However, a number of genes like *BRCA2* (MIM 600185) and *SLC6A4* (MIM 182138) are much higher in LCLs than the tissues, as are alternative splice events of many genes. In conclusion, these studies demonstrate that important pharmacogenes are variably expressed across tissues of pharmacologic relevance, and across different individuals, and that the vast majority are alternatively spliced.

## 404

**Transcriptome prediction in relevant tissues reveals mechanisms of drug-induced peripheral neuropathy.** *H.E. Wheeler[1], B.P. Schneider[2], D.L. Kroetz[3], K. Owzar[4], D.L. Hertz[5], H.L. McLeod[6], E.R. Gamazon[1], K.P. Shah[1], K.D. Miller[2], G.W. Sledge[2,7], N.J. Cox[1], M.E. Dolan[1], H.K. Im[8].* 1) Dept of Medicine, University of Chicago, Chicago, IL; 2) Dept of Medicine, Indiana University, Indianapolis, IN; 3) Dept of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA; 4) Dept of Biostatistics and Bioinformatics, Duke University, Durham, NC; 5) Dept of Clinical, Social, and Administrative Sciences, University of Michigan College of Pharmacy, Ann Arbor, MI; 6) Personalized Medicine Institute, Moffitt Cancer Center, Tampa, FL; 7) Dept of Medicine, Stanford University Medical Center, Stanford, CA; 8) Dept of Health Studies, University of Chicago, Chicago, IL.

The biological mechanisms underlying associations discovered in GWAS are often not well understood because few associated variants fall in the protein-coding regions of genes. For many traits, including chemotherapeutic toxicity, gene regulation is likely to play a crucial mechanistic role given the consistent enrichment of eQTLs among trait-associated variants. Our approach, called PrediXcan, harnesses the regulatory knowledge generated by eQTL studies to directly test for genes associated with complex traits. The Genotype-Tissue Expression (GTEx) Project has increased the number of relevant tissues for which genotype and expression data are available. Chemotherapy-induced peripheral neuropathy is the major dose-limiting toxicity for several anticancer drugs. Using genotype and tibial nerve expression data from the GTEx pilot phase, we computed predicted levels of gene expression in six genotyped clinical neuropathy cohorts and tested the predictions for association with the neuropathy phenotypes. An advantage of this gene-based approach is that the results are biologically interpretable, guiding follow-up experiments and future drug development. Positive correlation between predicted levels and phenotype indicates high expression of the gene is associated with neuropathy risk, and thus the gene represents a potential drug target, while negative correlation indicates low expression is associated with neuropathy risk. We performed a meta-analysis of the results from six clinical neuropathy cohorts and one gene, *TMED4*, reached genome-wide significance (Z-score = 4.8, unadjusted $P = 1.4 \times 10^{-6}$, Bonferroni $P = 0.02$). *TMED4* (transmembrane emp24 protein transport domain containing 4) is involved in vesicular protein trafficking and sensitizes cells to oxidative damage and cell death. Thus, peripheral nerves expressing high levels of *TMED4* may be more sensitive to drug-induced oxidative stress and apoptosis. The second most significant gene in the meta-analysis was *MARK3* (Z-score = -4.0, unadjusted $P = 6.7 \times 10^{-5}$, MAP/microtubule affinity-regulating kinase 3). These two genes were also the top two hits from a meta-analysis that only included the three (of six total) taxane-induced peripheral neuropathy cohorts (*TMED4* $P = 1.1 \times 10^{-5}$; *MARK3* $P = 2.1 \times 10^{-5}$). Given that taxanes target the microtubules, *MARK3* also represents a promising candidate for functional investigation. PrediXcan is applicable to GWAS of other complex traits and pharmacological phenotypes.

## 405

**Evidence for extensive pleiotropy among pharmacogenes.** *M.T. Oetjens[1], W.S. Bush[1,2], J.C. Denny[2,3], K.A. Birdwell[2], H.H. Dilks[1], S.A. Pendergrass[4,5], M.D. Ritchie[4,5], D.C. Crawford[1,6].* 1) Center for Human Genetics Research, Vanderbilt University, Nashville, TN; 2) Department of Biomedical Informatics, Vanderbilt University, Nashville, TN; 3) Department of Medicine, Vanderbilt University, Nashville, TN; 4) Center for Systems Genomics, The Pennsylvania State University, University Park, PA; 5) Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA; 6) Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN.

Genetic variants in drug-metabolizing enzymes and drug transporters are exemplars of pleiotropy as a result of their diverse roles in the metabolism of drugs, pollutants, and endogenous compounds. Variants in pharmacogenes that change systemic levels of these compounds can lead to an altered risk of disease in carriers. For instance, genome-wide association studies (GWAS) revealed the dualistic role of the *SLCO1B1* rs4149056 in serum bilirubin levels and statin induced myopathy. However, results from GWAS may underestimate the roles of pharmacogenes in disease risk. To systematically identify pleiotropic relationships among pharmacogenes, we performed phenome-wide association studies (PheWAS) using 6,092 European-descent subjects with DNA samples linked to de-identified electronic medical records as part of Vanderbilt University Medical Center's biorepository BioVU. All samples were genotyped on the Illumina ADME Core Panel, which was specialized for assaying 184 functional variants across 34 pharmacogenes. The ICD-9 codes present in these data were aggregated into 808 categories and matched control groups. We performed single SNP tests of association using logistic regression with each ICD-9 derived trait as an outcome adjusted for age and sex assuming an additive genetic model. We replicated five previously reported associations from the literature. For example, we robustly reproduced two associations between ADME genes and physiological traits: *ABCG2* rs2231142 and gout $p = 1.33 \times 10^{-7}$ (OR = 1.74, 95% CI = 1.46 - 2.07) and *SLCO1B1* rs4149056 and jaundice $p=2.44 \times 10^{-4}$ (OR = 1.67, 1.33 - 2.11). Epidemiological studies have reported an association between *CYP2C19* variants and development of gastrointestinal cancer, and we observed an association between *CYP2C19* rs4244285 and gastric cancer at $p < 0.05$ (OR = 1.69,1.09 - 2.63) and atrophic gastritis at $p= 2.44 \times 10^{-4}$ (OR= 1.99, 1.46 - 2.70). For novel associations, we set a Bonferroni corrected PheWAS significance threshold at $p < 5.76 \times 10^{-5}$. We detected one novel association between *SLC15A2* rs1143672 and renal osteodystrophy $p=2.29 \times 10^{-6}$ (OR = 0.60, 0.51 - 0.72). *SLC15A2* encodes PEPT2, a peptide transporter expressed in the proximal tubule of the kidney. To our knowledge, this is the first systematic screen for phenotypic associations using functional variants in pharmacogenes. Collectively, this ADME PheWAS suggests pharmacovariants may have systemic disease risk as well as altered drug response.

## 406

**Identifying risk variants for dihydropyrimidine dehydrogenase deficiency using an isogenic system of expression.** *S.M. Offer, S. Shrestha, C.R. Jerde, R.B. Diasio.* Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN.

Dihydropyrimidine dehydrogenase (DPD) deficiency [MIM 274270] manifests most frequently as severe clinical toxicity to the commonly prescribed anti-cancer drug 5-fluorouracil (5-FU), but has also been linked to various degrees of developmental delay in children, which is likely due to altered uracil catabolism. Genetic variations in the gene encoding DPD, *DPYD* [MIM 612779], have been suggested as the primary cause of DPD deficiency; however, given the rarity of individual candidate causal alleles within the gene, the functional consequences of many variants has not been evaluated. To address this important topic and identify which *DPYD* variants contribute to DPD deficiency, we have determined the effect of all 144 reported missense, nonsense, frameshift, and splice variants on DPD enzyme function, DPD protein stability, and expression of correctly-spliced *DPYD* mRNA. Each amino acid-changing variant was expressed in mammalian cells and the in vitro enzyme activity of expressed DPD measured. Approximately one-third of the reported variants directly impaired 5-FU catabolism to a degree that makes them candidate risk alleles for 5-FU toxicity. Based on publically-available allele frequency data, these variants are expected to be carried by 2-6% of the global population. To gain a broader appreciation for the spectrum of coding variations present within the gene, we performed targeted high throughput sequencing of the region encoding *DPYD* in a cohort of Somali-American individuals from Southeastern Minnesota, a population for which limited DNA sequence data are available. In addition to 8 previously reported variants, 13 novel missense, frameshift, and splice donor site variants were detected in this population. Collectively, these findings suggest that multiple rare variants contribute to DPD deficiency and that the spectrum of risk variants may vary greatly between racial groups. For that reason, targeted predictive tests developed in a single racial group are likely to incompletely genotype relevant alleles in other racial groups, necessitating a sequence-based approach for detecting risk alleles. Furthermore, the described system of phenotypically evaluating recombinantly-expressed variants should be applicable to additional drug-gene pathways and the study of other nucleotide metabolism disorders.