American Society of Human Genetics 62nd Annual Meeting November 6–10, 2012 San Francisco, California

NHGRI participation is designated by red print.

ANCILLARY EVENTS

Workshop

Social Media + Scientists = Success: Strategies for Using Social Media to Benefit Your Research, Your Career, and Your Connections

Separate advance registration required. Please check the appropriate box on the meeting registration form.
7 – 8:30 p.m., Wed., Nov. 7
Moscone Center, Room 310, Esplanade Level (South)

This workshop, which is moderated by Jonathan Gitlin, science policy analyst at NHGRI, will involve short presentations from panelists who have demonstrated a high degree of effectiveness using social media, followed by a highly interactive discussion with the audience. The panelists, including Jeannine Mjoseth, deputy chief of communications at NHGRI, are drawn from a cross-section of the genomics community and include an early-stage researcher (Daniel MacArthur, Harvard University), a senior faculty investigator and former editor for Nature (Chris Gunter, HudsonAlpha), and an industry representative in the commercial sector (Shirley Wu, 23andMe).

This workshop will open with a brief introduction to the different social media platforms that will be discussed and an overview of their uses and benefits for scientists at any stage in their career. The session will also feature a brief "Social Media 101" tutorial that shows attendees how to actually use the particular social media tools that the speakers will be discussing in this workshop (either live online or as PowerPoint slides with screenshots). Attendees will leave the workshop armed with knowledge about effective ways to use social media in the context of their genetics/genomics work.

Seating is limited, so please register early. Refreshments are included in the **$25 USD** registration fee. Seats may be reserved only by registering in advance. Tickets will be issued along with your meeting badge. **Tickets are non-refundable.**

Play

An Interactive Play: Drama, Discourse and Genomics: From IRBs to IFs

Separate advance registration required.

7 pm-9 p.m., Fri., Nov. 9
Moscone Center, Room 300, Esplanade Level

This interactive session weaves audience participation with the premiere of an original new **vignette-play** that illuminates ethical, psychological, social, legal, and policy concerns surrounding the sharing of information generated by next-generation sequencing. Between each act of the fictionalized play, **audience and actors are engaged in discussion**, as the vignettes evolve -- from an Institutional Review Board (IRB) meeting through the informed consent process to the disclosure of incidental findings (IFs). The presenters, Lynn Bush and Karen Rothenberg, delve into controversial ethical issues including:

How much and what should be reported?

- To whom?
- Under what circumstances?
- How sharing medical information differs in varying cultures.

Volunteer actors include:

- James Evans, Univ. of North Carolina
- Carla Easter, NHGRI
- Virginia Sybert, Univ. of Washington
- Kristy Lee, Univ. of North Carolina
- Rosario Isasi, McGill University
- Amy McGuire, Baylor University
- Carlos Bustamante, Stanford University
- Barbara Biesecker, NHGRI
- Barbara Koenig, UCSF
- Ruth Fischbach , Columbia University
- Kwame Anyaney-Yeboa, Columbia University
- Kandamurugu (Murugu) Manickam, Nationwide Childrens

This dramatization is intended to foster a deeper appreciation for conflicts when potentially disclosing massive amounts of genomic information. The dynamics of this play revolve around a family (including a child symptomatic with an autosomal recessive disorder and an "unaffected" sibling), their geneticist, genetic counselor and a discussion among IRB members, as they all experience the challenges of a genomic research study. Interspersed throughout the play, we explore together the complex implications of genomic information, particularly dilemmas raised by the return of results. Contextual subtleties will be brought to life by the volunteer actors who are ASHG attendees.

Variations in scenarios are shared to highlight the role that context plays when disclosing incidental findings. The goal of this interactive play is to provide a creative approach to enhance understanding of the inherent tensions and implications to both patient-participants and professionals in deciding whether to share genomic information with family members of varying ages, each with different values and vulnerabilities.

Seating is limited, so please register early. The **$10 USD** fee will offset the cost of evening refreshments. A cash bar will be available. Seats may be reserved only by registering in advance. Tickets will be issued along with your meeting badge. **Tickets are non-refundable.**

Award

Victor A. McKusick Leadership Award

8:20-8:40 a.m., Sat., Nov. 10
Moscone Center, Hall D, Lower Level North

The American Society of Human Genetics (ASHG) has named Francis Collins, M.D., Ph.D., Director of the National Institutes of Health, as the recipient of the 2012 Victor A. McKusick Leadership Award. Dr. Collins, the sixth recipient of the award, will be honored at the ASHG 62nd annual meeting on Saturday, Nov. 10, at the Moscone Convention Center in San Francisco.

This award, named in honor of the late Dr. Victor A. McKusick, recognizes individuals whose professional achievements have fostered and enriched the development of human genetics. "Recipients exemplify the enduring leadership and vision required to ensure that the field of human genetics will flourish and successfully assimilate into the broader context of science, medicine, and health, while also making major contributions to awareness or understanding of human genetics by policy makers or by the general public," said Joann Boughman, Ph.D., Executive Vice President of ASHG.

Dr. Collins was selected for the 2012 award because of his extensive achievements in genetics research, his efforts to advance health science and technology through policy and education, and his stellar leadership of the genetics community in mapping the human genome. "The revolution that was dreamed of at the start of the Human Genome project is currently being realized," Boughman said. "Today's medical geneticists, genetic counselors, and other health professionals are increasingly able to identify genes associated with both single-gene and complex diseases."

INVITED SESSIONS AND SESSIONS MODERATED BY NHGRI INVESTIGATORS

8 –10 a.m.
Wed., Nov. 7

Concurrent Invited Session I (3-10)
SESSION 10 – Metabolism, Metals, and Neurodegeneration: Toward Enhanced Understanding of Disease Mechanisms and Rational Therapeutics
Room 130, Lower Level North, Moscone Center
Moderators: Stephen G. Kaler, NICHD/NIH
Susan J. Hayflick, Oregon Hlth. & Sci. Univ.

This session will focus on the expanding knowledge concerning the influence of inborn errors of metabolism and disorders of trace metal homeostasis on neurodegeneration. Increasing numbers of clinical phenotypes and molecular defects are now associated with disordered metabolism in the central and peripheral nervous systems. These include Alzheimer and Parkinson diseases, Menkes and Wilson diseases, ATP7A-related distal motor neuropathy, acetyl CoA transporter 1-related hypocupremia, pantothenate kinase-associated neurodegeneration, infantile neuroaxonal dystrophy, dystonia-parkinsonism, Friedreich ataxia, hemochromatosis, and iron sulfur cluster scaffold myopathy. Expert speakers will discuss and review translational research advances relevant to these conditions, as well as emerging data on disease mechanisms, pathophysiology, and potential novel remedies.

9:40 AM **Exploring the link between glucocerebrosidase mutations and Parkinson disease.** E. Sidransky. NHGRI/NIH.

4:30–6:45 p.m.
Thurs., Nov. 8
Concurrent Platform (abstract-driven) Session C (38-46)
SESSION 43 – Genetics of Craniofacial and Musculoskeletal Disorders
Room 124, Lower Level North, Moscone Center
Moderators: Irini Manoli, NHGRI/NIH
Siddharth Prakash, Univ. of Texas Hlth. Sci. Ctr. at Houston

4:30–6:45 p.m.
Thurs., Nov. 8
Concurrent Platform (abstract-driven) Session C (38-46)
SESSION 46 – Pharmacogenetics: From Discovery to Implementation
Room 123, Lower Level North, Moscone Center
Moderators: Toni Pollin, Univ. of Maryland Sch. of Med.
Edward Ramos, NHGRI/NIH

8–10:15 a.m.
Fri., Nov. 9
Concurrent Platform (abstract-driven) Session D (47-55)
SESSION 49 – Common Variants, Rare Variants, and Everything in-Between
Room 135, Lower Level North, Moscone Center
Moderators: Joan E. Bailey-Wilson, NIH/NHGRI
Rasika Mathias, Johns Hopkins Univ. Sch. of Med.

9:40–11:40 a.m.
Sat., Nov. 10
Concurrent Invited Session III (73-80)
SESSION 73 – Returning Results from Large-Scale Sequencing: Where the Rubber Meets the Road
Gateway Ballroom 103, Lower Level South, Moscone Center
Moderators: Leslie G. Biesecker, NHGRI/NIH
Robert C. Green, Brigham and Women's Hosp.

is on the brink of a revolution, as large-scale medical sequencing (LSMS) is now available for patient care, marking the dawn of genomic medicine. LSMS may be used in patients with a family history or symptoms of a disease for diagnosis or to predict future health risks for prevention and surveillance. Developing standards and procedures for the use of LSMS in clinical medicine is critical and there is a need for empiric data to determine how to interpret, analyze, and return results. A critical question is how to handle secondary findings from LSMS. Speakers in this session will share their pioneering research initiatives that explore the translation of LSMS into meaningful clinical information and the delivery to patients. The speakers will report on a range of studies, including patients ascertained with diseases, healthy volunteers, adults, and children, research versus clinical, and rare versus common disease. All of the speakers will focus on results, and not just opinions, to inform the practice of genomic medicine and future research studies on the return of results. Dr. Biesecker will introduce the session and present ClinSeq, a large cohort LSMS study with return of results to subjects. Dr. Green will present the results of two efforts to formulate consensus on return of incidental findings in LSMS. Dr. Kingsmore will present experience with neonatal and pediatric diagnosis by LSMS. Finally, Dr. Veltman will present the results of a pilot using LSMS to diagnose genetically heterogeneous diseases with required return of all medically relevant results.

9:40 AM **ClinSeq: A pilot study of large-scale medical sequencing in research and implications for clinical genomic medicine.** L. G. Biesecker. NHGRI/NIH.

PLENARY SESSIONS

		Abstract Numbers
Tuesday, November 6		
2. **4:30–6:30pm:** Plenary Abstract Presentations	Hall D	**#1–#6**
Wednesday, November 7		
10:30am–12:45 pm: Concurrent Platform Session A (11–19):		
11. Genetics of Autism Spectrum Disorders	Hall D	**#7–#15**
12. New Methods for Big Data	Ballroom 103	**#16–#24**
13. Cancer Genetics I: Rare Variants	Room 135	**#25–#33**
14. Quantitation and Measurement of Regulatory Oversight by the Cell	Room 134	**#34–#42**
15. New Loci for Obesity, Diabetes, and Related Traits	Ballroom 1	**#52–#60**
17. Chromosomes and Disease	Room 130	**#70–#78**
18. Prenatal and Perinatal Genetics	Room 130	**#70–#78**
19. Vascular and Congenital Heart Disease	Room 123	**#79–#87**
Thursday, November 8		
10:30am–12:45pm: Concurrent Platform Session B (29–37):		
29. Next-Generation Sequencing: Methods and Applications	Hall D	**#88–#96**
30. Genetics and Intellectual Disability	Ballroom 103	**#97–#105**
31. GWAS from Head to Toe	Room 135	**#106–#114**
32. Cardiovascular Genetics: GWAS and Beyond	Room 134	**#115–#123**
33. Clinical Genetics: Mutations, Mutations and Syndromes	Ballroom 104	**#124–#132**
34. Cancer Genetics II: Clinical Translation	Room 124	**#133–#141**
35. Ethical, Legal, Social and Policy Issues	Room 132	**#142–#150**
36. Chipping Away at Autoimmune Disease	Room 130	**#151–#159**
37. Metabolic Disease Discoveries	Room 123	**#160–#168**
Thursday, November 8		
4:30pm–6:45pm: Concurrent Platform Session C (38–46):		
38. A Sequencing Jamboree: Exomes to Genomes	Hall D	**#169–#177**
39. Admixture and Demography	Ballroom 103	**#178–#186**
40. Analysis of Multilocus Systems	Room 135	**#187–#195**
41. Genes Underlying Neurological Disease	Room 134	**#196–#204**
42. Cancer Genetics III: Common Variants	Ballroom 104	**#205–#213**
43. Genetics of Craniofacial and Musculoskeletal Disorders	Room 124	**#214–#222**
44. Tools for Phenotype Analysis	Room 132	**#223–#231**
45. Therapy of Genetic Disorders	Room 130	**#232–#240**
46. Pharmacogenetics: From Discovery to Implementation	Room 123	**#241–#249**
Friday, November 9		
8:00am–10:15am: Concurrent Platform Session D (47–55):		
47. Structural and Regulatory Genomic Variation	Hall D	**#250–#258**
48. Neuropsychiatric Disorders	Ballroom 103	**#259–#267**
49. Common Variants, Rare Variants, and Everything in-Between	Room 135	**#268–#276**
50. Population Genetics Genome-Wide	Room 134	**#277–#285**
51. Endless Forms Most Beautiful: Variant Discovery in Genomic Data	Ballroom 104	**#286–#294**
52. Clinical Genetics: Complex Mechanisms and Exome-Discovery	Room 124	**#295–#303**
53. From SNP to Function in Complex Traits	Room 132	**#304–#312**

54. Genetic Counseling and Clinical Testing	Room 130	**#313–#321**
55. Mitochondrial Disorders and Ciliopathies	Room 123	**#322–#330**

Friday, November 9
4:30pm–6:45pm: Concurrent Platform Session E (59–67):

59. Genome Structure and Variation	Hall D	**#331–#339**
60. Advances in Neurodegenerative Disease	Ballroom 103	**#340–#348**
61. Missing Heritability, Interactions and Sequencing	Room 135	**#349–#357**
62. Exome Sequencing Uncovers Etiology of Mendelian Disease	Room 134	**#358–#366**
63. Transcriptional Regulation, Variation and Complexity	Ballroom 104	**#367–#375**
65. Advances in Ocular Genetics	Room 132	**#385–#393**
66. Cancer Genetics: Somatic Variants	Room 130	**#394–#402**
67. Developmental Insights into Human Malformations	Room 123	**#403–#411**

NHGRI PLENARY ABSTRACTS

13. Cancer Genetics I: Rare Variants

33

Somatic activating mutations in PIK3CA cause progressive segmental overgrowth. *M.J. Lindhurst1, V.E.R. Parker2, F. Payne3, J.C. Sapp1, S. Rudge4, J. Harris2, A.M. Witkowski1, Q. Zhang4, M.P. Groeneveld2, C.E. Scott3, A. Daly3, S.M. Huson5, L.L. Tosi6, M.L. Cunningham7, T.N. Darling8, J. Geer9, Z. Gucev10, P.A. Kreiger11, V.R. Sutton12, M.M. Thacker13, C. Tziotzios14, A.K. Dixon15, T. Helliwell16, S. O'Rahilly2,17, D.B. Savage2,17, M.J.O. Wakelam4, R.K. Semple2,17, I. Barroso2,3, L.G. Biesecker1*. 1) The National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 2) The University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Cambridge, UK; 3) The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK; 4) The Babraham Institute, Babraham Research Campus, Cambridge UK; 5) Genetics Unit, Manchester Academic Health Science Centre, Manchester, UK; 6) Division of Orthopaedics, Children's National Medical Center, Washington DC; 7) Division of Craniofacial Medicine, University of Washington School of Medicine, Seattle, WA; 8) Department of Dermatology, Uniformed Services University of the Health Sciences, Bethesda, MD; 9) Greenwood Genetics Center, Greenwood, SC; 10) Department of Endocrinology and Genetics, Skopje Medical Faculty, Skopje, Macedonia; 11) Department of Pathology, A.I. duPont Hospital for Children, Wilmington, DE; 12) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 13) Department of Orthopaedic Surgery, A.I. duPont Hospital for Children, Wilmington, DE; 14) Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK; 15) School of Clinical Medicine, University of Cambridge, Cambridge, UK; 16) Liverpool Cancer Research UK Center, University of Liverpool, Liverpool, UK; 17) The National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge, UK.

The phosphatidylinositol-3-kinase (PI3K)/AKT signaling pathway is critical for cellular growth and metabolism. Mutations in key genes have been identified in numerous tumor samples, while loss of PTEN function or activation of AKT1, AKT2 or AKT3 have been implicated in disorders that feature overgrowth and/or hypoglycemia. We performed exome sequencing of DNA from affected and unaffected skin fibroblasts from a patient (C1) with unclassified severe overgrowth of the lower extremities and identified a cancerassociated variant in *PIK3CA* in DNA from the affected sample that was not present in the unaffected sample. This variant, c.3140A>T which predicts p.His1047Leu, was also found in DNA isolated from other affected tissues from patient C1 including muscle, bone, fibrous and adipose tissue. We then screened affected cells and tissues from 15 patients with clinical features that overlapped with patient C1 and found the p.His1047Leu variant in two patients and a second variant, p.His1047Arg, in eight patients. The mutation burden amongst the samples harboring these variants ranged from <1% to 50% but neither variant was found in any blood samples from these patients. The predominant finding in this cohort was segmental progressive overgrowth of subcutaneous, muscular and visceral fibroadipose tissue with skeletal overgrowth that was sometimes, but not always, distorting. The severity varied remarkably and ranged from massive overgrowth of both legs to overgrowth limited to two rays of one foot. Despite having mosaic overgrowth that is both progressive and sporadic, these patients did not meet the clinical criteria for Proteus syndrome. Their features more closely match those of CLOVES syndrome, however, these patients lacked the complex truncal vascular malformations that are commonly found in patients with CLOVES. We tested affected tissue from two patients that met the criteria for CLOVES syndrome and found a *PIK3CA* p.Glu542Lys variant in one patient and *PIK3CA* p.Glu545Lys in the other. These variants are also activating mutations commonly found in tumor samples. Finally, in a patient with an isolated congenital linear verrucous epidermal nevus, we found *PIK3CA* p.Glu545Lys in keratinocytes but not fibroblasts isolated from the lesion. These findings expand the spectrum of phenotypes associated with somatic activation of PI3K signaling and suggest multiple therapeutic targets for patients with progressive segmental overgrowth.

15. **New Loci for Obesity, Diabetes, and Related Traits**

46

Exome analysis in 8,232 Finnish men identifies novel loci and lowfrequency variants for insulin processing and secretion. *J.R. Huyghe1, A.U. Jackson1, M.P. Fogarty2, A. Stancˇáková3, H.M. Stringham1, M.L. Buchkovich2, C. Fuchsberger1, J. Paananen3, P.S. Chines4, T.M. Teslovich1, J.M. Romm5, H. Ling5, I. McMullen5, R. Ingersoll5, E.W. Pugh5, K.F. Doheny5, J. Kuusisto4, L.J. Scott1, F.S. Collins4, G.R. Abecasis1, R.M. Watanabe6, M. Boehnke1, M. Laakso3, K.L Mohlke2*. 1) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA; 2) Department of Genetics, University of North Carolina, Chapel Hill, NC, USA; 3) Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; 4) Genome Technology Branch, National Human Genome Research Institute, Bethesda, MD, USA; 5) The Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD, USA; 6) Department of Preventive Medicine, Keck School of Medicine of USC, CA, USA.

Insulin secretion plays a critical role in maintenance of blood glucose levels, and failure to secrete sufficient insulin is a hallmark of type 2 diabetes. Genome-wide association studies (GWAS) have identified loci involved in these processes; however, many genetic contributions remain undefined. Until now it has not been possible to study the role of low-frequency (minor allele frequency (MAF) < 5%) nonsynonymous variants in complex traits exome-wide and systematically in large samples. To determine the role of low-frequency nonsynonymous variants in insulin processing and secretion, we designed an exome array based on variants discovered by sequencing > 10,000 subjects and genotyped ⍰242,000 SNPs on the Illumina Infinium HumanExome Beadchip in 8,232 non-diabetic Finnish men from the population- based Metabolic Syndrome in Men (METSIM) study. We identified two novel genes harboring low-frequency variants associated with insulin processing and secretion: TBC1D30 (2.0% MAF) and KANK1 (3.0% MAF), both of which function in G protein signaling pathways. We also identified coding low-frequency variants at two known loci. A nonsynonymous variant in SGSM2 (1.4% MAF) was independent of the GWAS-identified common signal. A nonsense variant in MADD (3.7% MAF) occurred on a haplotype containing the most associated SNP of two independent common GWAS signals at this locus. The nonsense allele, associated with decreased insulin secretion, always occurred with the GWAS allele associated with decreased secretion, and adjusting for one variant in a conditional analysis decreased but ccccccdid not eliminate association for the other variant. Finally, we observed that linkage disequilibrium with nonsynonymous variants in this region can extend up to 1 Mb. The interpretation of both single-variant and gene-based tests therefore needs to consider the effects of distant common SNPs, an especially important consideration when exome sequence data are analyzed in the absence of data on the surrounding noncoding regions. In conclusion, although sequencing will still be needed for a fully comprehensive analysis of variants, this study provides proof of principle that exome array genotyping is a valuable approach for identifying low-frequency functional variants, and for fine-mapping of GWAS-identified loci, in complex traits.

41. **Genes Underlying Neurological Disease**

199

Vps37A causes a novel form of complex Hereditary Spastic Paraparesis. *T. Falik-Zaccai1,2,5, Y. Zivony-Elboum1,5, W. Westbroek3, D. Savitzki4, Y. Shoval1, Y. Anikster6, A. Waters7, R. Kleta7,8*. 1) Institute of Human Genetics, Western Galilee Hosp, Nahariya, Israel; 2) The Galilee Faculty of Medicine - Bar Ilan, Tzfat, Israel; 3) Section on Human Biochemical Genetics, Medical Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda MD, USA; 4) Department of Child Development, Western Galilee Hospital-Nahariya, Israel; 5) Rappaport Faculty of Medicine, Technion, Haifa, Israel; 6) Metabolic Unit, Edmond and Lily Safra Children's Hospital, Sheba Medical Center, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel; 7) Nephro-Urology Unit, Great Ormond Street Hospital, London WC1N 3JH, UK; 8) Division of Medicine, University College London, London NW3 2PF, UK.

Hereditary Spastic Paraplegia (HSP) comprises a heterogeneous group of neurodegenerative disorders characterized by progressive lower limb spasticity, retrograde degeneration of the crossed cortico-spinal tracts, and thinning of the posterior columns in the spinal cord. Complicated forms (CHSP) are characterized by the addition of such neurological features as spastic quadriparesis, seizures, dementia, amyotrophy, extrapyramidal disturbance, cerebral or cerebellar atrophy, optic atrophy, and peripheral neuropathy, as well as by extra neurologic manifestations such as dysmorphism, albinism, retinitis pigmentosa. CHSP forms are generally inherited as autosomal recessive (AR) traits. Currently, more than 40 HSP loci and 21 causative genes for pure and complicated HSP forms have been identified. We report members of two unrelated kindred of Arab Moslem origin who present with infantile spastic paraparesis of upper and lower limbs, mild intellectual disability, kyphosis, pectus carinatum, and hypertrichosis. We performed neurological and developmental examinations on the affected individuals. We conducted whole genome linkage and haplotype analyses, followed by sequencing of candidate genes; RNA and protein expression studies, and investigations on knockdown morpholino oligonucleotide injected zebrafish. We characterize a novel form of AR CHSP. MRI studies of brain and spinal cord were normal. Within a single significantly linked locus we identified a homozygous missense mutation c.1146A>T (p.K382N) in the Vacuolar Protein Sorting 37A (Vps37A) gene, fully penetrant and segregating with the disease in both families. Mobility was significantly reduced in Vps37A knockdown morpholino oligonucleotide injected zebrafish, supporting the causal relationship between mutations in this gene and the phenotype described in our patients. We provide evidence for involvement of Vps37A, a member of the endosomal sorting complex required for transport (ESCRT) system, in upper motor neuron disease. The ESCRT system has been shown to play a central role in intracellular trafficking, in the maturation of multivesicular bodies and the sorting of ubiquitinated membrane proteins into internal luminal vesicles. Further investigation of mechanisms by which dysfunction of this gene cause CHSP will contribute to the understanding of intracellular trafficking of vesicles by the ESCRT machinery and its relevance to CHSP.

42. **Cancer Genetics III: Common Variants**

211

Statistical fine mapping of regions containing melanoma susceptibility genes identified through genome-wide association studies. *J.H. Barrett1, J.C. Taylor1, M. Brossard2, A.M. Goldstein3, P.A. Kanetsky4, E.M. Gillanders5, J.A. Newton Bishop1, D.T. Bishop1, F. Demenais2, M.M. Iles1, GenoMEL consortium.* 1) Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, St James's University Hospital, University of Leeds, Leeds, United Kingdom; 2) INSERM, U946, Fondation Jean- Dausset-CEPH, Paris, France; 3) Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute. National Institutes of Health, Bethesda, Maryland, USA; 4) Centre for Clinical Epidemiology & Biostatistics, University of Pennsylvania, Philadelphia, Pennyslvania, USA; 5) Inherited Disease Research Branch, National Human Genome Research Institute, US National Institutes of Health, Baltimore, Maryland, USA.

Genome-wide association (GWA) studies of melanoma have resulted in the identification and confirmation of approximately 15 loci associated with disease risk. In most of these regions the causal variant(s), and sometimes even which genes are implicated, are still not known. We have applied statistical methods to data from the GenoMEL GWA study (Barrett et al, Nature Genetics, 2011; 43:1108-13) to refine each locus of interest and identify the most parsimonious model(s) explaining the association. Genotypes were imputed in at least 2Mb around each locus using IMPUTEv2 (Howie et al, PLoS Genetics, 2009) with 1000 Genomes (March 2012 release) as reference panel after stringent quality control. All genotyped single nucleotide polymorphisms (SNPs) and imputed SNPs with an INFO score >0.8 were analysed using a gene dosage approach in standard single locus logistic regression analysis adjusting for population structure. This analysis was repeated conditioning on the top genotyped and then the top imputed SNP. Hyperlasso (Hoggart et al, PLoS Genetics, 2008), which implements a form of penalized logistic regression, was applied to all SNPs in the region to select a set of associated SNPs, and the model including these selected SNPs was further characterised using multiple logistic regression. In the 16q24.3 region close to *CDK10* and *MC1R* this approach allowed us to show (and confirm by sequencing) that the signal is explained by the known functional variants in *MC1R*, despite the fact that the original signal is closer to other genes. The results obtained varied markedly across the other loci. At one extreme, for the region on 11q14-q21 around the tyrosinase gene, the association was completely explained by one SNP (the most strongly associated imputed SNP, which is in strong linkage disequilibrium (r2=0.91) with the missense variant rs1126809). For over half the loci a single SNP explaining the association could not be identified. For example, for the 5p15.33 region around *TERT* and *CLPTM1L*, despite the fact that the initial signal was confined to a narrow region, the simplest model included a two-SNP haplotype and possibly one other SNP, suggesting either a more complex mechanism or the involvement of SNP(s) neither genotyped nor imputed. Statistical fine mapping is a useful first step in the attempt to identify the causal mechanisms giving rise to association signals. Next steps will include sequencing of more narrowly defined regions and functional experiments.

45. **Therapy of Genetic Disorders**

239

Positive effects of short course androgen therapy on the neurodevelopmental outcome in boys with 47, XXY syndrome at 36 and 72 months of age. *C. Samango-Sprouse1,3, E. Stapleton3, T. Sadeghin3, F. Mitchell3, T. Dixon3, M. Kingery3, A. Gropman1,2*. 1) George Washington University of the Health Sciences, Washington, DC; 2) Department of Neurology, Children's National Medical Center, Washington, DC; 3) NDC for Young Children, Davidsonville, MD.

Background: The effects of early androgen treatment on neurodevelopmental performance in prepubertal males with 47, XXY have not been well investigated. Androgens have a profound effect on modulating neurodevelopment, brain function and behavioral outcomes from as early as 16 weeks gestation throughout adulthood. Androgen insufficiency has been described during puberty in several studies in XXY and suggests that hormone replacement therapy may have a positive outcome on brain function. Males with XXY have CNS abnormalities including endocrine and neurocognitive deficits with language based learning disabilities, dyspraxia, and musculoskeletal anomalies. **Objective:** To determine if an early course of androgen treatment (3 injections of testosterone enanthate, 25mg, each) could have a positive impact on neurodevelopmental function in XXY boys immediately and later. **Methods:** 101 prenatally diagnosed males with 47, XXY participated in comprehensive neurodevelopmental assessments. One group (n=34) received androgen treatment in infancy and the second was untreated (n= 67). Statistical analysis was completed to determine if there was a treatment effect at 36 and 72 months on multiple domains of development. **Results:** At 36 months, there was a significant positive treatment effect in multiple neurodevelopmental domains on the WISC-IV of Vocabulary p= .0007, Comprehension p=.0099, VIQ p=.0225 and FSIQ p=.0203. A positive treatment effect was observed in Vocabulary p=.0052, Comprehension p= .0165, intellectual abilities of VIQ p=.0081 and FSIQ p=.0270 and neuromotor skills p=.0197 at 72 months. **Conclusion:** Improved function was observed in neurodevelopmental performance in XXY males at 36 and 72 months when treated with a short course of androgen in infancy and revealed significant improvement in areas of the brain with known androgen receptors, which have been described as deficient in XXY males. Demonstrating that an early course of hormonal replacement may have an extended positive effect and giving support to the link between neurobiological treatment and neurodevelopmental outcome and the possibility for newborn screening resulting in early treatment for XXY associated developmental disabilities.

46. **Pharmacogenetics: From Discovery to Implementation**

241

Pharmacogenomics, ancestry and clinical decision making for global populations. *E. Ramos1, A. Doumatey1, H. Huang1, D. Shriner1, G. Chen1, S. Callier2, J. Zhou1, A. Adeyemo1, H. Mcleod3, C. Rotimi1*. 1) Center for Research on Genomics and Global Health, NIH/NHGRI, Bethesda, MD; 2) Department of Clinical Research and Leadership, School of Medicine and Health Sciences, George Washington University, Washington, DC; 3) Institute for Pharmacogenomics and Individualized Therapies, University of North Carolina, Chapel Hill, NC.

A significant component of individualizing patient care will be largely attributable to our understanding of the influence the human genome has on adverse drug reactions, dosing and other treatment modalities (i.e., pharmacogenomics). However, best practices for laying the groundwork towards treating a single individual include the consideration of genetic variation from multiple populations. Therefore, we examined 19 global populations sampled from 5 continents allowing for analysis at various levels, including inter- and intracontinental comparisons of variants relevant to drug metabolism. Specifically, we genotyped several African-ancestry populations using a platform that contains nearly 2000 variants selected from roughly 230 genes known to be involved in the absorption, distribution, metabolism, or excretion of drugs. For the remaining populations, we extracted these markers from the publically available genome sequences of the 1000 Genomes project. Minor allele frequencies (MAFs) were calculated and compared across all populations in addition to measurements of population differentiation. The set of markers were analyzed as a whole; however, a subset of markers (42 total) identified to be clinically useful was highlighted. We identified several clinically actionable single nucleotide polymorphisms (SNPs) that vary among populations in this study. For example, rs9923231, associated with warfarin metabolism (an anti-platelet drug) showed MAFs of nearly the entire range (from 0.02 in Yorubans to 0.95 in Han Chinese from Beijing). However, variation was not limited to just global comparisons. We also observed markers that varied in frequency within a given country illustrated by rs776746 (cyclosporine metabolism) in the Kenyan samples (MAF of 0.88 and 0.54 for Luhya and Masaai, respectively). In addition, the MAF for rs1801280 that predicted the acetylator phenotype of NAT2 was 0.71 in Spanish Iberian samples compared to other western European counterparts or Latin American samples, which ranged from 0.34 to 0.46. We highlighted in this study clinically actionable pharmacogenomic markers where group labels such as “black” or “Hispanic” could be a barrier to safe and effective drug selection. The data highlights the importance of casting a wide net when trying to assess the profile of clinically relevant genetic variation. Our data also speak to other ethical and social issues such as access as well as relevant public policy implications.

47. **Structural and Regulatory Genomic Variation**

258

A SNP associated with skin cancer and pigmentation disrupts a melanocyte enhancer in an intron of *IRF4*. *D.U. Gorkin1, S.K. Loftus2, D. Lee3, M.A. Beer1,3, W.J. Pavan2, A.S. McCallion1*. 1) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA; 2) Genetic Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA; 3) Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21201, USA.

Genome-Wide Association Studies (GWAS) have identified Single Nucleotide Polymorphisms (SNPs) associated with hundreds of human phenotypes. However, efforts to identify the causative variants underlying these associations have been hindered by limited knowledge of the location and sequence composition of functional non-coding sequences. Focusing on phenotypes that involve pigment cells (melanocytes), we recently developed two tools that facilitate the identification of variants impacting functional non-coding sequences: 1) a genome-wide catalog of 2,489 melanocyte enhancers identified by ChIP-seq (EP300/H3K4me1) with a high validation rate in functional assays of 86% *in vitro* (43/50) and 70% *in vivo* (7/10), and 2) a vocabulary of sequence motifs derived by

machine learning (Lee D. et al. 2011) that are predictive of melanocyte enhancer function with power to indentify additional enhancers genome-wide in both the mouse and human genomes. We used these tools to identify a melanocyte enhancer in an intron on *IRF4* that contains a SNP (rs12203592) associated with skin cancer (melanoma, basal cell carcinoma, and squamous cell carcinoma), nevus count, freckling, hair color, and eye color (Han J. et al. 2011; Duffy D.L. et al. 2010; Han J. et al. 2008; Gathany A.H. et al. 2009; Eriksson N. et al. 2010). The sequence containing rs12203592 drives strong reporter gene expression in cultured melanocytes, and shows additional hallmarks of melanocyte enhancer activity including EP300 binding and H3K4me1 enrichment in mouse melanocytes, and DNase I hypersensitivity in human primary melanocytes and melanoma cell lines. Based on our enhancer sequence vocabulary, the risk allele of rs12203592 (T) is predicted to have a strongly negative impact on enhancer function. We demonstrate that this allele significantly diminishes the ability of the enhancer to drive reporter expression in melanocytes (P= 2.7e-5). We will discuss our progress in uncovering the molecular basis of the association between rs12203592 and melanocyte-related phenotypes, as well as how knowledge of functional non-coding sequences can be systematically applied to identify variants that disrupt the function of these sequences.

53. **From SNP to Function in Complex Traits**

304

The type 2 diabetes (T2D) risk allele of rs11603334 increases *ARAP1* promoter activity and is associated with increased *ARAP1* mRNA in pancreatic islets. *J.R. Kulzer1, M.L. Stitzel2, M.A. Morken2, F.S. Collins2, K.L. Mohlke1*. 1) Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC; 2) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD.

Genome-wide association studies have identified many loci associated with T2D and pancreatic islet dysfunction, but for most, the functional variants and target genes have not been determined. We are investigating molecular and biological mechanism(s) underlying association of the *ARAP1* locus with T2D and fasting proinsulin. Index SNPs rs11603334 and rs1552224 are in perfect linkage disequilibrium (LD) (r2 = 1.0) and located 113 bp apart within the 5'UTR of the short isoform of *ARAP1*. All SNPs in high LD (r2 > 0.8) with the index SNPs are non-coding. A third SNP, rs7109575 (r2 = .86 with the index SNPs), is located at the transcription start site of the long isoform of *ARAP1*. We hypothesized that one of these three SNPs influences the transcriptional regulation of *ARAP1*. We measured allele-specific mRNA levels from 87 human islet samples and determined that the index SNP T2D risk alleles are associated with increased *ARAP1* mRNA (*P* ⍰⍰.01 for each of two assays), suggesting that one or more risk variants increase *ARAP1* transcriptional activity or message stability. We then evaluated the allele-specific effects of rs11603334, rs1552224, and rs7109575 on transcriptional activity at the *ARAP1* promoters by performing dual luciferase reporter assays in the INS-1-derived rat beta cell line 832/13. The two-SNP haplotype containing the T2D risk alleles of rs11603334 (G) and rs1552224 (T) reproducibly increased promoter activity two-fold compared to the nonrisk haplotype (*P* ⍰⍰.001). When the effects of rs11603334 and rs1552224 were separated by site-directed mutagenesis, the G allele of rs11603334 exhibited two-fold increased transcriptional activity (*P* < .001), while rs1552224 showed no effect. The DNA region surrounding rs7109575 demonstrated strong promoter activity, but that activity did not differ between SNP alleles. Taken together, these results suggest that the rs11603334 T2D risk allele increases *ARAP1* mRNA expression by upregulating transcriptional activity at the promoter of the short isoform. Ongoing studies focus on identifying transcription factors differentially bound to rs11603334 and examine the consequences of increased ARAP1 expression on insulin processing and secretion. Investigating the molecular mechanisms underlying T2D-associated loci is an important step toward identifying genes that contribute to T2D susceptibility.

54. **Genetic Counseling and Clinical Testing**

316

Intentions to receive individual results from whole-genome sequencing among participants in the ClinSeqTM study. *B.B. Biesecker 1, F.M. Facio2, H. Eidem2, T. Fisher1, S. Brooks2, A. Linn2, K.A. Kaphingst3, L.G. Biesecker2*. 1) Social and Behavioral Research Branch, National Human Genome Research Institute, Bethesda, MD; 2) Genetic Diseases Research Branch, National Human Genome Research Institute, Bethesda, MD; 3) Department of Surgery, Washington University School of Medicine, St. Louis, MO.

Objectives: Genome sequencing has been rapidly integrated into clinical research, and is currently marketed to health care practitioners and consumers alike. The volume of sequencing data generated for a single individual and the wide range of findings from whole genome sequencing raise critical questions about the return of results and their potential value for endusers. Methods: We conducted a mixed-methods study of 311 sequential participants in the NIH ClinSeqTM study to assess general preferences and specific attitudes toward learning results and perceived opinions of valued others. We tested how these variables predicted intentions to receive results within four categories of findings ranging from medically actionable to variants of unknown significance. Results: Two hundred ninety-four participants indicated a preference to learn their genome sequencing results and six were unsure. Most often participants cited disease prevention as their reason, including intention to change their lifestyle behaviors. A third expressed a general desire to know, reflecting those who generally valued information and others who sought to understand the personal implications of findings. Participants had positive attitudes, strong perceived social norms and strong intentions to learn results overall, although there were significant mean differences among four categories of findings (p<0.01). Attitudes and social norms for medically actionable and carrier results were most similar and rated the highest. Attitudes and norms significantly contributed to the variance in intentions. Among these early adopters there was overwhelming enthusiasm to learn results. Conclusions: Participants distinguished among the types and quality of information they may receive despite strong intentions to learn all results presented. These intentions were motivated by confidence in their ability to use the information to prevent future disease and a belief in the value of even uninterpretable information. It behooves investigators to facilitate participants' desire to learn a range of information from genomic sequencing, while promoting realistic expectations for its clinical and personal utility.

62. **Exome Sequencing Uncovers Etiology of Mendelian Disease**

358

Loss of function mutations in known human disease genes in 572 exomes. *J. Johnston1, K. Lewis1, D. Ng1, S. Gonsalves1, J. Mullikin2,3, L. Biesecker1,2*. 1) Genetic Disease research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 2) NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 3) Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD.

Genome and exome sequencing costs continue to fall and many individuals are undergoing these assessments as research participants and patients. The issue of secondary findings in exome analysis is controversial and data are needed on their meaning in otherwise healthy individuals. The genetics literature includes a substantial fraction of papers describing causative variants based on minimal data. The default is to assume such a variant, especially if it falls in a gene with a function related to the patient's phenotype, is causative. To better understand the frequency of potentially causative mutations in apparently healthy persons, we have analyzed potential loss of function mutations including stop, frameshift and splice site alterations in 572 subjects of the ClinSeq™ project. A total of 9,421 potential loss of function variants were identified. As we were interested in clinical significance, further analyses were restricted to variants predicted to alter the protein-coding region of the transcripts annotated in the Human Gene Mutation Database (HGMD). This reduced our variant list to 835. Forty-five variants predicted to cause cancer susceptibility, lipid disorders, or cardiomyopathy/ channelopathy phenotypes have been analyzed separately and previously reported and were removed from our dataset. Final filters included frequency, quality and predicted inheritance. As our goal was to understand the impact of these variants on health, we filtered for

dominant or x-linked conditions. This filtering reduced our variant list to 54 variants. Eight of these variants were present in HGMD and the remaining 46 were novel. Five variants were present in multiple individuals. Phenotypes predicted to result from the identified variants included intellectual disability, developmental disorders including OFD1 and CHARGE syndrome, neuromuscular disorders including Becker muscular dystrophy, polycystic kidney disease, deafness and cataract. Proband and family history suggested a small minority of these are truly pathogenic. We will summarize the further proband and family investigations undertaken to evaluate the associated phenotypes. We recognize that an approximately 10% incidence of such predicted mutations in an otherwise healthy adult cohort is unreasonable. These data highlight challenges associated with interpreting apparently pathogenic null variants in exome and genome sequencing.

359

The Problem of Multiple Plausible Molecular Diagnoses in Next Generation Sequencing Data: The NIH Undiagnosed Diseases Program Experience. *D. Adams1,2, C. Boerkoel2, K. Fuentes Fajardo2, P. Cherukuri2, M. Sincan2, C. Toro2, C. Tifft1,2, W. Gahl1,2, T. Markello1,2*. 1) NHGRI, NIH, Bethesda, MD; 2) Undiagnosed Diseases Program, NIH, Bethesda, MD.

Introduction: Exome sequencing (ES) is a powerful diagnostic tool that is now available for clinical use. Methods and standards for interpreting the resulting data have yet to be established. Substantial practical challenges arise from clinical application of ES including genetic counseling, proof of disease association and secondary variants. The NIH Undiagnosed Diseases Program evaluates patients with complex medical syndromes. The use of ES for selected participants has revealed an additional characteristic of ES data—the presence of multiple likely-pathogenic variants in single individuals. **Methods:** 380 exome sequences have been obtained in 80 families. The resulting variants are subjected to an extensive set of filters including population frequency, segregation consistency, consistent high-quality genotypability, alignment correctness, and predicted pathogenicity. Promising variants are Sanger validated then assessed utilizing data from extensive clinical phenotyping. The resulting "high quality" variants are referred for further experimental validation. **Results:** To date, 54 families have generated high quality variants for further study. Of those, 39 have multiple high-quality variants. In an increasing number of individuals, we are finding multiple DNA mutations that are either known to be pathogenic or demonstrate severe pathogenic potential at the in vitro, cell biological and/or model organism level. In one example, two sibs were affected by an autosomal recessive disorder. In addition, each sib had one new-dominant mutation in a different gene; each sib had one shared, and one unshared, inherited condition. **Discussion:** Our patient cohort comprises undiagnosed patients who have undergone extensive prior evaluation. Our early experience with genome and exome sequencing suggests that some patients will be affected by multiple interacting disorders, rather than by a single condition. An alternative hypothesis is that one of the detected variants explains the entire syndrome, and other environmental or inherited disease modifiers have complicated prior efforts at diagnosis. In either case, our data highlights the fact that the ES may reveal increasingly complicated relationships between DNA variation and medical disease.

POSTER SESSIONS

The program and abstract/poster board number next to each listing is followed by a **W** (Wednesday), or **T** (Thursday), or **F** (Friday) to indicate the day on which authors must be present at their poster boards. Posters will remain on the boards for all three days (Wednesday through Friday).

Session Title /Topic Area	***Start #***	***End #***
Genome Structure, Variation and Function	#412	#600
Pharmacogenetics	#601	#655
Metabolic Disorders	#656	#753
Therapy for Genetic Disorders	#754	#804
Development	#805	#854
Cytogenetics	#855	#980
Cancer Genetics	#981	#1279
Statistical Genetics and Genetic Epidemiology	#1280	#1583
Cardiovascular Genetics	#1584	#1715
Genetic Counseling and Clinical Testing	#1716	#1796
Ethical, Legal, Social and Policy Issues in Genetics	#1797	#1850
Genetics Education	#1851	#1863
Health Services Research	#1864	#1877
Complex Traits and Polygenic Disorders	#1878	#2369
Neurodegeneration	#2370	#2661
Molecular Basis of Mendelian Disorders	#2662	#2966
Prenatal, Perinatal and Reproductive Genetics	#2967	#3034
Clinical Genetics and Dysmorphology	#3035	#3250
Evolutionary and Population Genetics	#3251	#3414
Epigenetics	#3415	#3525
Bioinformatics and Genomic Technology	#3526	#3735

POSTER SCHEDULE

Wednesday, November 7
10:00 am– 4:30 pm Posters open for viewing
2:15 pm– 4:15 pm Poster Session I (W)
2:15 pm– 3:15 pm *(odd poster board numbers; author must be present)*
3:15 pm– 4:15 pm *(even poster board numbers; author must be present)*

Thursday, November 8
7:00 am– 4:30 pm Posters open for viewing
2:15 pm– 4:15 pm Poster Session II (T)
2:15 pm– 3:15 pm *(odd poster board numbers; author must be present)*
3:15 pm– 4:15 pm *(even poster board numbers; author must be present)*

Friday, November 9
7:00 am– 4:30 pm Posters open for viewing
2:15 pm– 4:15 pm Poster Session III (F)
2:15 pm– 3:15 pm *(odd poster board numbers; author must be present)*
3:15 pm– 4:15 pm *(even poster board numbers; author must be present)*
4:15 pm– 4:30 pm Authors must remove posters
4:30 pm Exhibit Hall and Posters close

NHGRI POSTER ABSTRACTS

Metabolic Disorders

664T

Exploring the phenotype of MMACHC deficiency (*cblC*) in zebrafish with morpholinos and zinc finger nuclease targeted mutagenesis. *J.L. Sloan1, K. Bishop2, T.L. Blake2, R.J. Chandler1, B.P. Brooks3, R. Sood2, C.P. Venditti1*. 1) Organic Acid Research Section, Genetics and Molecular Biology Branch, , Bethesda,
MD; 2) Zebrafish Core Facility, Genetics and Molecular Biology Branch, National Human Genome Research Institute, Bethesda, MD; 3) Ophthalmic Genetics and Visual Function Branch, National Eye Institute, Bethesda, MD.
Cobalamin C deficiency (*cblC*) features a combined impairment of the cobalamin dependent enzymes, methionine synthase and methylmalonyl-CoA mutase. It is caused by mutations in the *MMACHC*gene, which encodes an enzyme that is suspected to participate in cobalamin decyanation and intracellular trafficking. The clinical spectrum of *cblC* is wide and patients can have developmental abnormalities, such as congenital microcephaly, heart defects and intrauterine growth retardation (IUGR). To examine the underlying pathophysiology of this disorder, we have studied the phenotype in zebrafish using two approaches. FITC-tagged morpholinos targeting the cognate ATG and an exonic junction were designed and used to knockdown zebrafish *MMACHC*. Morphants began to display defects at 24 hours, including brain necrosis and diminished movement when compared to controls. The morphants had delayed hatching and were significantly smaller in size, had less blood, smaller heads and eyes, and were neurologically impaired compared to controls. By 96 hours, injected fish displayed pericardial edema. Both morpholinos produced a similar phenotype, which was dose-dependent. The same phenotype was observed in the p53-/- fish, suggesting it is not due to off-target effects. Further investigation using transgenic fish expressing GFP in the CNS revealed underdeveloped brains and absent Rohon-Beard neurons. Additionally, histologic examination revealed a fatty liver and abnormalities in the retina. Metabolic analysis of extracts derived from the morphants showed increased methylmalonic acid (MMA) and cystathionine, and MMA was increased in fish incubated with the precursor, propionic acid (PA). Furthermore, the mutant fish were sensitive to PA, such that they died earlier than the mutant fish not incubated with PA. To extend our morpholino studies, a zinc finger nuclease targeted to *MMACHC* was engineered and an allelic series was created: p.V31fsX78, p.H43PfsX21, p.L44HinsM. Weare in the process of breeding and phenotyping these mutant lines. The model presented here is the first animal model of *cblC* and faithfully replicates some of the more severe findings observed in humans. Furthermore, it demonstrates the utility of zebrafish to easily examine aspects of metabolic diseases that will be difficult to study in other organisms, such as embryonic manifestations, and should facilitate the exploration of the underlying pathophysiological mechanisms and testing of new therapies for *cblC*.

674T

The clinical spectrum of combined malonic and methylmalonic aciduria,

a defect in the intramitochondrial fatty-acid-biosynthesis pathway.
I. Manoli1, J.L. Sloan1, J.J. Johnston2, L. Peller2, J.C. Sapp2, L.G. Biesecker2, C.P. Venditti1. 1) Genetics and Molecular Biology Branch, NHGRI, NIH, Bethesda, MD; 2) Genetic Disease Research Branch, NHGRI, NIH, Bethesda, MD.

We have identified acyl-CoA synthetase family member 3 (*ACSF3*) as the gene encoding a malonyl- and methylmalonyl-CoA synthetase residing in the mitochondrial matrix. ACSF3 deficiency causes a form of combined malonic and methylmalonic aciduria/acidemia (CMAMMA). In these patients, methylmalonic acid predominates and malonyl-CoA decarboxylase activity is normal. *ACSF3* mutations occur with a minor allele frequency of 0.005 in 5,950 control individuals, predicting that CMAMMA has a population incidence of 1:40,000. To further delineate the phenotypes associated with this enzymopathy, we evaluated fifteen subjects, 1.5–64 years of age, with CMAMMA. Seven subjects were diagnosed in adulthood with neurological or psychiatric manifestations (seizures, memory problems, psychiatric disease and/or cognitive decline, peripheral sensory neuropathy and optic nerve atrophy), after exclusion of vitamin B12 deficiency and other more common diagnoses. Three subjects presented during childhood with symptoms suggestive of an intermediary metabolic disorder (coma, ketoacidosis, hypoglycemia, failure to thrive, elevated transaminases), while the five others had more general findings including microcephaly, dystonia, axial hypotonia, muscle weakness, seizures, mild dysmorphic features, autism and global developmental delay. Three out of ten apparently unaffected siblings were diagnosed with CMAMMA after the index case was ascertained in each family. Missense mutations and one in-frame deletion were located in highly conserved acyl-CoA synthetase motifs, while none of the patients carried two nonsense mutations. One mutation, p.R558W, was present in 13/30 alleles. In one subject only one mutation was identified. In another, no damaging mutations were detected; however fibroblasts from this subject showed increased methylmalonic acid secretion following *in vitro* stimulation. Given the predicted population incidence, it is likely that many patients with ACSF3 deficiency remain undiagnosed. Furthermore, the variable clinical phenotypes observed in the patient cohort mandates the consideration of alternative diagnoses in addition to CMAMMA. The generation of an *ACSF3* mouse model will be required to understand the pathogenic role of mutations and to define disease pathophysiology.

709F
SATB2 Acts as an Activator of the UPF3B Gene. *P. Leoyklang1,2,3, K. Suphapeetiporn2,3, C. Srichomthong2,3, S. Tongkobpetch2,3, H. Dorward4, A.R. Cullinane4, M. Huizing4, W.A. Gahl4, V. Shotelersuk2,3*. 1) 1Biomedical Science Program, Faculty of Graduate School, Chulalongkorn University, Bangkok, Thailand; 2) 2The Center of Excellence for Medical Genetics, Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand; 3) 3Excellence Center for Medical Genetics, King Chulalongkorn Memorial Hospital, Thai Red Cross, Bangkok, Thailand; 4) Medical Genetics Branch, NHGRI (NIH), Bethesda, MD.

Two syndromic cognitive impairment disorders have very similar craniofacial dysmorphisms. One is caused by a heterozygous nonsense mutation of SATB2, a transcription regulator, and the other by heterozygous mutations leading to premature stop codons in UPF3B, encoding a member of the nonsense-mediated mRNA decay complex. We hypothesized that the products of these two causative genes functioned in the same pathway. Here, we demonstrated that the SATB2 nonsense mutation identified in our patient led to a truncated protein that localized to the nucleus, formed a dimer with the wild-type SATB2 and interfered with its normal activity. This suggested that the SATB2 nonsense mutation had a dominant negative effect leading to more severe clinical manifestations, compared to patients with SATB2 haploinsufficiency. We also showed that the patient's leukocytes had significantly decreased UPF3BmRNAcompared to controls. A chromatin immunoprecipitation assay demonstrated that the SATB2 protein bound to the promoter of the UPF3B gene. A luciferase reporter assay revealed that recombinant SATB2 protein significantly activated gene transcription using the UPF3B promoter. These findings indicate that SATB2 acts as an activator of the UPF3B gene through binding to its promoter. This study emphasizes the benefit of linking disorders with similar clinical phenotypes to elucidate

711F

Natural History Study of Patients with Hereditary Inclusion Body Myopathy (HIBM). *J.K. De Dios1, J. Shrader2, G. Joe2, C. Ciccone1, A. Mankodi3, J. Dastgir3, C. Bonnemann3, M. Bevans2, D. Draper1, J. McKew4, M. Huizing1, W.A. Gahl1, N. Carrillo-Carrasco4*. 1) Medical Genetics Branch (MGB), National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH), Bethesda, MD; 2) Clinical Center, NIH, Bethesda, MD; 3) National Institute of Neurological Disorders and Stroke (NINDS), NIH, Bethesda, MD; 4) Therapeutics for Rare and Neglected Diseases (TRND), National Center for Advancing Translational Sciences (NCATS), NIH, Bethesda, MD.

Background: HIBM, a rare disorder characterized by progressive muscle weakness, is caused by *GNE* mutations, impairing sialic acid biosynthesis. Weakness, apparent in early adulthood, progresses slowly from distal to proximal, first in lower and subsequently in upper extremities, with relative sparing of the quadriceps. There is no therapy available and patients eventually require a wheelchair. **Methods**: We have evaluated 17 patients (age range, 28 to 58 years) in a prospective, longitudinal, single-center study of HIBM (CT.gov: NCT01417533) to define the natural history, genotypephenotype correlations, prognostic factors, associated manifestations and potential biomarkers and outcome measures. **Results**: The mean age of presentation was 27 years with a mean diagnostic delay of 10 years (range: 2–19 years). Initial manifestations included falls, gait disturbance, foot drop, and overuse-related hand weakness (n=1). Various genotypes were seen, the most common being homozygous p.M712T (n=3). Decreased mean predicted ankle dorsiflexion (9 %), grip (34%) and knee extension (69%) strength were seen. Grip strength deficits emerged around 7.5 years after onset of first symptoms. The 6-minute walk was 68% of mean predicted distance and directly correlated with Activities-specific Balance Confidence Scale Scores. Less strength for selected muscle groups was also related to less balance confidence and longer duration of disease. MRI of lower extremities showed atrophy of muscles with relative sparing of the quadriceps. Prolonged QTc (n=1) and asymptomatic cardiomyopathy (n=1) were seen, but their association to HIBM remains unclear. CPK levels ranged from 161–1152 and decreased with progression of the disease. Mild elevation of ALT (n=2) and hypoalbuminemia (n=12) were seen; all patients had normal renal function tests. **Conclusions**: The characteristic presentation may be modified by muscle overuse. More data are needed to establish genotype/

phenotype correlations, but variability among siblings suggests involvement of other factors. Muscle strength and volume decreased with progression and correlated with functional measures and balance confidence. The heart was the only other organ involved as previously reported (Chai et al 2011). Planning clinical trials for potential therapies (ManNAc and sialic acid) is complicated by diagnostic delays and lack of natural history data. Selection of appropriate biomarkers and clinical outcome measures is under evaluation.

721F

Renal Growth in Isolated Methylmalonic Acidemia (MMA). *P. Kruszka, I. Manoli, J. Sloan, C.P. Venditti.* Genetics and Molecular Biology Branch, NHGRI, NIH, Bethesda, MD.

OBJECTIVES: A major complication of methylmalonic acidemia (MMA) is the development of tubulointerstitial nephritis and renal failure. The effect of MMA on kidney size has not been studied. This study examines kidney size variation with age and other measureable predictors and develops a kidney length nomogram. DESIGN: Renal ultrasounds, anthropometric measurements, and laboratory evaluations including serum creatinine (Cr), serum and urine methylmalonic acid, and cystatin C were conducted on 50 non-transplanted patients (ages 2.2 to 36.3 years) with enzymatic and mutation confirmed MMA. The patients were followed prospectively from 2004 to 2011, with some patients imaged sequentially, yielding 84 total ultrasounds. MMA patients were compared to a control group of 209 patients aged 0 to 19 years using the paired t-test. Linear and multiple regression analysis were used to study relationships between renal length and other clinical variables. RESULTS: Comparisons with age-matched controls showed a significant difference in renal length ($p < 0.05$) for most age groups. The regression equation for the control group was renal length (cm) = 6.79 + 0.22 * age (years) for children older than 1 year and for our study patients was 6.80 + 0.09 * age (years); $p < 0.001$ for both the constant and �age coefficient (R2 = 0.399). The most highly correlated single predictor model for MMA patients used height instead of age (R2 = 0.490). When serum Cr values were in the normal range (Cr < 1.3), the most predictive multiple regression model found height, age, and serum methylmalonic acid (MMAS) to be the optimal independent variables (R2 = 0.720). Cystatin C was comparable to MMAS in the multiple regression model and correlated closely with Cr and MMAS (Pearson correlation coefficients = 0.7901 and 0.7684, respectively). Urine methylmalonic acid concentration was not a significant predictor of renal length when controlling for age, height and Cr ($p = 0.849$). CONCLUSIONS: Renal length in MMA patients is significantly decreased compared to normal controls and predicted by a multiple regression model that uses the clinical variables of height, age, and MMAS when Cr is in normal range. In addition to generating a clinically useful nomogram, our results capture a marker of the natural history of MMA renal disease. The use of a simple index such as renal length will contribute to the study of therapeutic interventions designed to target the kidney in this population.

Therapy for Genetic Disorders

755W

Bone marrow derived cells as a stable source of sialic acid for mice with GNE myopathy. *MC. Malicdan[1,3], K. Momma[2], F. Funato[3], YK. Hayashi[3], I.*

Nonaka1, M. Huizing1, W. Gahl1,4, CF. Boerkoel1, I. Nishino3, S. Noguchi3.
1) MGB, NHGRI, National Institutes of Health, Bethesda, MD, USA; 2) Department of Neuromuscular Research, Natiional Institute of Neuroscience, NCNP, Tokyo, Japan; 3) Department of Neurology, National Defense Medical College, Saitama, Japan; 4) Office of the Clinical Director, NHGRI, NIH, Bethesda, MD, USA.

GNE myopathy, previously known as distal myopathy with rimmed vacuoles (DMRV), or hereditary inclusion body myopathy (hIBM), is an autosomal recessive myopathy characterized by progressive weakness and atrophy involving the distal muscles, and myofiber vacuolation and degeneration. GNE myopathy is secondary to mutations in the GNE gene, which encodes the bifunctional enzyme which catalyzes the first two crucial steps in sialic acid biosynthesis. We and others have shown that hyposialylation is one of the key players in disease pathomechanism. This notion is supported by our recent publications showing that the myopathic phenotype in the existing DMRV/hIBM mouse was prevented by giving exogenous oral sialic acid metabolites or modified ManNAc conjugates. Despite these promising results, designing an oral therapeutic regimen is often challenged by the the extremely rapid excretion of orally administered sialic acids that potentially requires frequent or almost continuous dosing. In this study, we established allogenic bone marrow transplantation (BMT) as a choice for therapy.Bone marrow cells were isolated from CAG-GFP expressing mice and injected intravenously to 30 weeks old DMRV mice after a single dose of sublethal irradiation. Chimerism and cell surface sialylation of peripheral leukocytes after RBC lysis was assessed four weeks after cell transplantation and monthly thereafter. Results show that there was a clear increment in cell surface sialylation and correlated with chimerism. GFP-positive donor cells were seen to engraft into skeletal muscles and other organs of DMRV/hIBM mice. DMRV/hIBM mice subjected to BMT had a marked improvement in lifespan and motor performance. Analysis of muscle contractile properties and pathology revealed an improvement in muscle phenotype. Evaluation of specific glycoproteins in the muscle demonstrated a recovery of cellular sialylation. GNE activity and sialic acid levels in several organs revealed a remarkable increase as compared to non-treated mutant mice. Our results provide a proof of concept for the utility of cell based therapies in DMRV/hIBM, and that hematopoetic cells can be a good source of sialic acid production. We envision that with this strategy, GNE myopathy patients can have a lasting supply of sialic acid that may benefit them towards disease recovery.

Psychiatric Genetics, Neurogenetics & Development

824F

Cc2d2a is required for cilia biogenesis. *S. Veleri1, TJ. Foskett1, SH. Manjunath1, A. Longo1, MA. English1,4, P. Liu2, J. Lei2, C. Gao3, RN. Fariss3, R. Sood4, RA. Rachel1, P. Liu4, L. Dong2, A. Swaroop1*. 1) NNRL, NEI, NIH, Bethesda, MD; 2) Genetic Engineering Core, NEI, NIH, Bethesda, MD; 3) Biological Imaging Core, NEI, NIH, Bethesda, MD; 4) NHGRI, NIH, Bethesda, MD.

Purpose: Mutations in CC2D2A (Coiled-Coil & C2 Domain containing 2A) are associated with Meckel-Gruber syndrome (MKS) and Joubert syndrome (JS), the two ciliopathies characterized by retinal dystrophy, mental retardation,

polydactyly, liver fibrosis, and polycystic kidney. CC2D2A protein is localized to basal body and transition zone of the cilia. The purpose of this study was to elucidate how CC2D2A mutations cause human disease by loss of function studies using model organisms. Methods: We used antisense splice blocking and translational blocking morpholinos (MOs) to knockdown zebrafish cc2d2a. The splice blocking MO mimicked a diseaseassociated splicing defect in human patients. The Cc2d2a-knockout (KO) mice were produced by targeted deletion of exons 6 to 8 using homologous recombination. Results: Knockdown of cc2d2a by both splice and translational blocking MOs in the zebrafish embryos demonstrated dose-dependent developmental defects, including brain atrophy and microphthalmia. The Cc2d2a-/- mice exhibit embryonic lethality, with extensive developmental defects that include situs inversus, heterotaxy, polydactyly, anophthalmia, hydrocephalus and liver fibrosis. Occasionally, Cc2d2a-/- animals survive for a month but display severe hydrocephalus and retinal dystrophy. The analysis of embryonic fibroblasts, embryonic node and kidney tubules showed that cilia biogenesis is disrupted by Cc2d2a-/- mutation. Conclusion: Our studies suggest that Cc2d2a function is critical for normal body plan and organ development that involves cilia-mediated signaling pathways.

Cancer Genetics

1001F

Meta-analysis of 25 prostate cancer associated SNPs in high-risk prostate cancer families: new evidence from the International Consortium for Prostate Cancer Genetics (ICPCG). *C. Teerlink1, S. Thibodeau2, D. Schaid3, K. Cooney4, E. Lange5, C. Maier6, J. Stanford7, E.A. Ostrander8, J. Schleutker9, G. Cancel-Tassin10, O. Cussenot10, R. Eeles11, D. Easton12, W. Isaacs13, J. Xu14, J. Carpten15, J. Bailey-Wilson16, F. Wiklund17, A. Whittemore18, W. Catalona19, W. Foulkes20, N. Camp1, L. Cannon-Albright1, 21, International Consortium for Prostate Cancer Genetics.* 1) Dept Internal Medicine, Univ Utah Sch Med, Salt Lake City, UT; 2) Department of Lab Medicine and Pathology, Mayo Clinic, Rochester, MN 55905; 3) Department of Health Sciences Research, Mayo Clinic, Rochester,MN55905; 4) Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109; 5) Departments of Genetics and Biostatistics, University of North Carolina, Chapel Hill, NC 27599; 6) Institute for Human Genetics, University of Ulm, Germany; 7) Fred Hutchinson Cancer Research Center (FHCRC), Division of Public Health Sciences, Seattle, WA 98195; 8) Cancer Genetics Branch, National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH), Bethesda, MD 20892; 9) Department of Medical Biochemistry and Genetics, University of Turku, Turku 20520, Finland; 10) CeRePP ICPCG Group, Hopital Tenon, Assistance Publique-Hopitaux de Paris, Paris75020, France; 11) The Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK; 12) Strangeways Laboratory, Worts Causeway, Cambridge CB1 8RN, UK; 13) Johns Hopkins University ICPCG Group, Department of Urology, Johns Hopkins Medical Institutions, Baltimore, MD 21287; 14) Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, NC 27157; 15) Translational Genomics Research Institute, Phoenix, AZ 85004; 16) Johns Hopkins University, Baltimore, MD 21224; 17) Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; 18) Department of Health Research and

Policy, Stanford School of Medicine, Stanford, CA 94305; 19) Northwestern University ICPCG Group, Northwestern University Feinberg School of Medicine, Chicago, IL 60611; 20) Program in Cancer Genetics, McGill University, Montreal, Quebec H3T 1E2, Canada; 21) George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, UT 84148.

Previous GWAS studies have reported associations between various SNPs and prostate cancer using cases unselected for family history. We report the results of a validation study of 25 SNPs reported to be associated with prostate cancer using a novel analysis of related familial prostate cancer cases. Fourteen study sites contributed a total of 11,783 genotyped samples, including 3,786 unaffected controls, 5,628 cases with non- or undeterminedaggressive disease, and 2,368 cases with aggressive disease. The majority of genotyped samples (n = 8,573) originated from 2,276 pedigrees, the remainder were individually sampled cases. Each site contributed its own controls. We used PedGenie software to account for known relationships. We conducted separate analyses for all prostate cancer and for aggressive prostate cancer only. For the non-specific prostate cancer phenotype, our analysis showed that 16 of the 25 SNPs were statistically significant ($p < 2E-3$ (= 0.05/25)), including SNPs on chromosomal bands 6q25, 7p15, 8q24, 10q11, 11q13, 17q12, 17q24, and Xp11. For the aggressive prostate cancer phenotype, our analysis showed that 8 of the 25 SNPs, were statistically significant ($p < 2E-3$), including most of the same chromosomal bands except 6q25 and 10q11 but also included a SNP at 2p15. The results of this analysis validate strong association to familial prostate cancer for the majority of SNPs considered and demonstrate the power of a GWAS approach that integrates data for related cases.

1081T

Histologic Types and Risk Factors in Familial Lung Cancer Cases from Southern Louisiana. *D. Mandal1, M. Haskins1, A. Bencaz1, J. Hutchinson1, J. Chambliss1, H. Rothschild1, J.E. Bailey-Wilson2.* 1) Dept Gen, LSU Hlth Sci Ctr, New Orleans, LA; 2) National Human Genome Research Institute, National Institutes of Health, Baltimore, MD.

The association between lung cancer (LC) and smoking is well known. However, only 15% of smokers are diagnosed with LC. In addition, about 10% of LC cases (22,000 cases per year in the U.S.) have at least one relative affected with LC. So, family history is a significant risk factor. In a case control study in Germany, adenocarcinoma was most prevalent in never smokers and in women; squamous cell carcinoma was most prevalent in male smokers. In the 1990's in the population in Louisiana, squamous cell carcinoma was also observed to be the most frequent type in LC cases in general(39.3%), with nearly equal numbers of adenocarcinoma (25.0%) and small cell subtypes (25.5%). While some studies have shown similar proportions of cases with a family history in all histologic subtypes of LC, others have suggested that a higher proportion of patients with squamous cell carcinoma have a family history of LC. However, less is known about the proportions of each histologic type in LC cases with a strong family history.The objective of the present study was to analyze histologic subtypes and their association with smoking behaviors and other risk factors of interest among familial LC cases from southern Louisiana. Eligible subjects (N=148) with 2 relatives affected with primary lung cancer were recruited from southern Louisiana. Diagnosis of primary LC was confirmed through medical records, and histologic subtype (N=114) was abstracted from pathology reports. About 81% of cases had non-small cell LC, with adenocarcinoma (40%) being the most common histologic subtype, followed by squamous cell at 28% and small cell at 18%. Histologic type of LC was most strongly related to pack-years of cigarette smoking and age at diagnosis in these familial cases, with squamous cell associated with higher mean pack-years of smoking and older age. Preliminary results indicated significant difference in the age of diagnosis between cases with non-small cell LC and those with small cell LC; mean number of pack years was about twice as high in cases with non-small cell LC. The frequency of adenocarcinoma in these familial cases was higher and of squamous cell carcinoma was lower than previously reported for the Louisiana LC population. This is consistent with a higher risk of adenocarcinoma in persons with a family history, which may be related to our previous observations that less smoking appears to be necessary for LC in persons with a familial risk of LC.

1261T

Exome sequencing approach for identification of causative mutations in neurofibromatosis type 1 (NF1)-associated plexiform neurofibromas.

A. Pemov1, H. Li2, M. Wallace2, D.R. Stewart1, NISC Comparative Sequencing Program, NIH Intramural Sequencing Center, NHGRI, NIH, Bethesda, MD. 1) NCI, NIH, Bethesda, MD; 2) University of Florida, Gainesville, FL.

Background. NF1 is an autosomal dominant tumor pre-disposition genetic disorder, caused by constitutive inactivation of one of two copies of the tumor suppressor *NF1*. Individuals with NF1 are prone to the development of both benign and malignant tumors including skin and plexiform neurofibromas (PNF). Although it is accepted that somatic inactivation of *NF1* is a necessary step in PNF development, little is known what other genes/pathways are disrupted/affected in the tumor. In this study, we performed wholeexome sequencing of 11 tumor samples matched with germline DNA obtained from 11 unrelated NF1 patients. **Methods.** Germline DNA was extracted from peripheral white blood cells and tumor DNA was obtained from primary Schwann cell (SC) cultures established from dissected plexiform neurofibromas. All cell cultures were of low passage and contained at least 70% SC. We used a capture kit from Illumina (“TruSeq”) that targets roughly 60 million bases consisting of the CCDS annotated gene set. We sequenced the captured DNA on the Illumina HiSeq platform until we had sufficient coverage to call genotypes with an arbitrary quality score (“MPG”) of 10 for at least 85% of targeted bases. We analyzed the data using an R-based statistical software designed to identify “driver” genes from “passengers.” **Results.** First, we analyzed the *NF1* locus in both germline and tumor DNA. We identified pathogenic mutations in the *NF1* gene in all but one germline sample. All germline mutations were also confirmed in the tumor DNA. In addition, we identified somatic inactivation of *NF1* in nine out of 11 tumors. Second, we evaluated concordance of the exome sequencing data with genotyping data obtained from Illumina 2.5M SNP-arrays. Out of ⍰45,000 SNPs genotyped in both platforms, 97% were concordant. Finally, we analyzed the data to identify genes that could play an important role in PNF tumorigenesis. After correcting P-values for multiple testing we identified eight frequently mutated statistically significant genes with false discovery rate (FDR) below 0.05. Evaluation of biological functions of the genes revealed that such processes as cell cycle, nonsense-mediated mRNA decay and Notch pathway signaling might play an important role in PNF tumorigenesis. **Conclusions and further steps.** To our knowledge, this is the first attempt to identify “driver” mutations in NF1-associated PNF via exome sequencing. We are validating select mutations with the classic Sanger approach.

Statistical Genetics and Genetic Epidemiology

1302T

Unraveling Phenotype Heterogeneity in Prostate Cancer Susceptibility in Finland Utilizing Covariate-Based Analysis. *C.D. Cropp1, C.L. Simpson1, T. Wahlfors3,4, A. George1,5, M.P.S. Jones2, U. Harper2, D. Ponciano-Jackson2, T. Tammela6, J. Schleutker3,4, J.E. Bailey-Wilson1.* 1) Inherited Disease Research Branch, National Human Genome Research Institute/ National Institutes of Health, Baltimore, MD; 2) Genomics Core/Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland; 3) Institute of Biomedical Technology/ BioMediTech, University of Tampere and Fimlab Laboratories, Tampere, Finland; 4) Department of Medical Biochemistry and Genetics, University of Turku, Turku, Finland; 5) Fox Chase Cancer Center, Philadelphia, Pennsylvania; 6) Department of Urology, Tampere University Hospital, University of Tampere, Tampere, Finland.

Prostate cancer is the most common male cancer in developed countries. Previously, we reported a genome-wide linkage scan in 69 Finnish Hereditary Prostate Cancer (HPC) families, which replicated the *HPC9* locus on 17q21-q22 and identified a locus on 2q37. We used ordered subset analysis (OSA) to detect other loci linked to HPC in subsets of families to identify and detect other loci linked to HPC incorporating age of onset as a traitrelated covariate to address genetic heterogeneity and strengthen the linkage findings previously reported. The overall mean age of onset across the families was 66.2±8.8 years while the range of individual onset ages ranged from 46 to 98 years. Although the highest OSA LOD score with a DLOD (p=0.02) was 2.876 on 15q26.2-q26.3 in a subset of 40 families ascending by age at onset, no other DLOD scores were significant after permutation testing. Since OSA uses a single covariate value per family, we used mean age of onset for each pedigree. To better capture the effect of age on the linkage signal, we used LODPAL to perform a linkage analysis in affected relative pairs, while adjusting for the age of each individual family member as a single covariate. Preliminary results revealed strong evidence of linkage to HPC on chromosome 15q was (LOD=4.9, 132cM) and 8q (LOD=3.1, 157cM). Permutations are ongoing to determine empirical p-values for these LOD scores.

1347T

Selection of sequence variants for quantitative traits using penalized regression: using LASSO, LARSand Elastic Net in the Tiled Regression framework. *Y. Kim1, B. Suktitipat1,2, A.J.M Sorant1, A.F. Wilson1*. 1) IDRB/ Genometrics Sec, NHGRI/NIH, Baltimore, MD; 2) Biochemistry Dept, Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand.

Selecting the set of independent genetic variants that contribute to the variation of quantitative traits can be challenging because of the multicollinearity caused by highly correlated sequence variants either in linkage or gametic disequilibrium; and the use of regression methods that ignore multicollinearity can be problematic. Penalized regression methods are known to be robust to the multicollinearity of variables — shrinking the coefficients of less informative variables close to or equal to zero until the regression model is optimized. In this study we investigate the performance of penalized regression methods and compare these results to those from stepwise regression as currently implemented in the Tiled Regression Analysis Program (TRAP v1.0).Wecompared the performance of three penalized regression methods: LASSO, Elastic Net (EN) and LARS, to results from stepwise regression (STEP). We simulated 4000 individuals with 300,000 variants based on HapMap data, and 100 replicates of quantitative traits from seven independent causal loci with the same effect size but with various locusheritability (h2) ranging from 0.00045 to 0.01 based on an additive genetic model. We examined the detection rates (% per 100 replicates) of each causal locus, precision (true positive/true positive + false positive) and sensitivity (true positive/ true positive + false negative) of all causal loci in the final subset under two scenarios: including and excluding causal variants. In the presence of causal variants, LASSO, EN, and STEP showed high detection rates (76–100%) for causal variants with locus-specific h2 > 0.005, but LARS and STEP were higher than LASSO and EN for detecting causal variants of h2 < 0.005. However, LARS showed the least precision (0.004) and sensitivity (0.42) for the final subsets. After excluding causal variants in causal loci, the penalized regression methods showed slightly higher detection rates (1–28%) than STEP (0–15%) — doing a somewhat better job at detecting variants that were correlated to the excluded causal variants. Overall sensitivity and precision were decreased compared to the results including causal variants. LARS showed the least precision (0.002), but observed similar sensitivity (0.18) with LASSO (0.16) and EN (0.16). In conclusion, LASSO and EN may be an effect tool in detecting variants that are correlated with causal variants when the causal variants are not among the variants considered.

1363F

A two-stage random forest approach to identify genetic variants using recombination hotspot information. *S. Szymczak1, Q. Li1, Y. Kim2, A. Dasgupta3, J.D. Malley4, J.E. Bailey-Wilson1.* 1) Statistical Genetics Section, Inherited Disease Research Branch, National Human Genome Research Institute, NIH, Baltimore, MD; 2) Genometrics Section, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD; 3) Clinical Sciences Section, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD; 4) Center for Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD.

In genome-wide association studies (GWAS) each single-nucleotide polymorphism (SNP) is usually analyzed separately for association with a disease. However, it is expected that multiple genetic loci jointly contribute to disease risk. Recently, several machine learning algorithms have been used to study many or even all genome-wide genetic variants simultaneously. One promising approach is random forest (RF), an ensemble method based on a large number of classification and regression trees trained on bootstrap samples. RF provides variable importance measures that can be used to select the most relevant SNPs. However, identification of true risk SNPs that are in strong linkage disequilibrium (LD) with non-risk SNPs is challenging.

Large regions of strong LD can lower the ranking of all SNPs in a causal region since they can serve as proxies for each other in different trees.
We propose a two-stage approach that uses information about recombination hotspots in the genome. In a first step, a random forest is trained for each region between two hotspots. SNPs within these regions are more likely to be in strong LD. Predicted case-control status or estimated probability of case/control status based on each region's SNPs then replaces the group of SNPs as a predictor variable. These new predictor variables for all hot spot blocks are used in a genome-wide random forest analysis in the second stage.
We compare our approach with a random forest analysis using all SNPs simultaneously based on simulated GWAS data mimicking local LD patterns observed in European samples of the 1,000 genomes project. Genotype data for 500K SNPs in 5,000 cases and 5,000 controls are generated using the software GWAsimulator. We compare the false-positive errors under a null model with no true genetic effects. For a power analysis several variants with different minor allele frequencies and in regions with different local LD structure are modeled as true risk SNPs with independent effects on the disease status.

1396F

Tiled Regression Improves the False Discovery Rate in Genome-Wide Association Studies. *B. Suktitipat1,2, Y. Kim1, A.J.M. Sorant1, H. Sung1, A.F. Wilson1.* 1) Genometrics Section, NHGRI/NIH, Baltimore, MD; 2) Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand.

Most genome-wide association studies (GWAS) aim to identify causal genes based on single variant effects ignoring the effects of all other variants in the genome. Ideally, multiple regression of all the variants in the genome would identify the set of independent causal variants; however, the requirement of complete data makes this approach problematic. Tiled Regression combines simple linear or logistic regression models, stepwise selection of variables and a staged approach. User-defined regions (“tiles”) of potentially correlated SNPs (e.g., based on hotspots or genes) are first considered separately and discarded if they show no evidence of association with a trait. Stepwise regression is then used to select independent significant SNPs from the remaining tiles, with higher order stepwise regression used at the chromosome and genome levels. We evaluated the performance of this approach with the Tiled Regression Analysis Program (TRAP v1.0) using simulated data. We simulated 4000 individuals with 306,097 markers spanning 53,060 tiles across the genome, based on phased data from the HapMap CEU population. To evaluate the false positive rate, we simulated 100 replicates of normally distributed random phenotypes. To evaluate the power, we simulated 100 replicates of quantitative traits from seven independent loci with the same effect size but with various locus specific heritability from 0.0005 to 0.01 assuming additive genetic model. A critical value of 0.0001 was used as the selection criteria for the stepwise regression for the tile, chromosome, and genome levels. We compared our results to conventional single marker analysis based on simple linear regression model testing one marker at a time. On average, the type I error for TRAP was 5E-05, about one-half of the nominal type I error rate. TRAP had a false discovery rate (FDR), calculated as a proportion of false positive signals divided by all positive signals, of 0.76, which was better than the FDR of

0.86 from single marker analysis using the same statistical significant level. Power estimated with TRAP was comparable to the power estimated from single marker analysis. In summary, in a GWAS setting, we found that TRAP has comparable power but with a lower FDR when compared to single marker analysis. Thus, the use of TRAP is beneficial in selecting independent markers contributing to quantitative trait and in reducing the false discovery rate in GWAS.

1447F

Merging Genomic Data for Research in the Electronic MEdical Records and GEnomics Network: Lessons learned in eMERGE.

M.D. Ritchie1, S. Setia1, G. Armstrong1, L. Armstrong2, Y. Bradford3, D.C. Crawford3, D.R. Crosslin4, M. de Andrade5, K. Doheny6, M.G. Hayes2, G. Jarvik4, I.J. Kullo5, R. Li7, T. Manolio7, M. Matsumoto5, C.A. McCarty8, D. Mirel9, S. Nelson4, L. Olson3, E. Pugh6, S. Purcell10, G. Tromp11, J.L. Haines3. 1) Biochemistry & Molecular Biology, Pennsylvania State University, University Park, PA; 2) Northwestern University, Chicago, IL; 3) Vanderbilt University, Nashville, TN; 4) University of Washington, Seattle WA; 5) Mayo Clinic, Rochester, MN; 6) CIDR, Johns Hopkins; 7) NHGRI, Bethesda, MD; 8) Essentia Rural Health, Duluth, MN; 9) Broad Genotyping Center, Boston, MA; 10) Mt Sinai, New York, NY; 11) Geisinger Medical Center, Danville, PA.

Biobanks linked to electronic health records (EHR) is an emerging area of research for dissecting the architecture of complex human traits. Electronic phenotyping algorithms are deployed in large EHR systems to “ascertain” samples for analysis. To achieve success, sample size is an important consideration to maximize statistical power. To this end, merging data generated from different genotyping platforms is often desired. The eMERGE network, an NHGRI funded initiative, has developed a pipeline for merging genomic data generated on a single platform as well as a new pipeline for merging data from different genotyping arrays based on imputation. eMERGE consists of seven sites, each with DNA databanks linked to EHRs. Over 42,000 samples have been genotyped using one of the available Affymetrix and Illumina genome-wide genotyping arrays. These data have been imputed using BEAGLE and the October, 2011 release of the 1000 Genomes cosmopolitan reference dataset. Because of the computational complexity of imputation and the large sample size of the merged eMERGE dataset (42,807 individuals), a distributed imputation pipeline was implemented. In this scheme, the genome was divided by SNPs into 30,000 marker “SNPlets” with 700 markers of overlap on each side, resulting in 510 SNPlets; the data were also divided into sample sets of n=2000 or less. This parallelized pipeline resulted in over 556 billion SNPs (more than 13 million per individual) based on hundreds of thousands of CPU hours. The dataset generated consists of genome-wide SNPs on thousands of individuals all linked to EHR systems where numerous phenotypes can be explored. The genotype data will be available for research in dbGAP. The lessons learned by this group of investigators will be valuable for the genomics community also dealing with the combining of large-scale genomic datasets. We will present the details of our imputation pipeline, including our quality control investigations in data of this scope and magnitude. This merged eMERGE dataset is an invaluable resource for the genomics community to discover genetic risk factors for common, complex diseases and pharmacogenomic traits.

1454W

Fine-Mapping in a Covariate-based Genomewide Linkage Scan of Lung Cancer Susceptibility. *C.L. Simpson1, T. Green1, B. Doan1,2, C.I. Amos3, S.M. Pinney4, E.Y. Kupert4, M. de Andrade5, P. Yang5, A.G. Schwartz6, P.R. Fain7, A. Gazdar8, J. Minna8, J.S. Wiest9, H. Rothschild10, D. Mandal10, M. You4, T.A. Coons11, C. Gaba12, M.W. Anderson4, J.E. Bailey-Wilson1.* 1) Inherited Disease Res Branch, NHGRI, NIH, Baltimore, MD; 2) Johns Hopkins School of Medicine, Baltimore, Maryland; 3) Department of Epidemiology, University of Texas, M.D. Anderson Cancer Center, Houston, Texas; 4) Medical College of Wisconsin, Milwaukee,Wisconsin; 5) Department of Health Sciences Research, Mayo Clinic Rochester, Minnesota; 6) Karmanos Cancer Institute, Wayne State University, Detroit, Michigan; 7) University of Colorado, Denver, Colorado; 8) University of Texas Southwestern Medical Center, Dallas, Texas; 9) National Cancer Institute, NIH, Bethesda, Maryland; 10) Louisiana State University Health Sciences Center, New Orleans, Louisiana; 11) Saccomanno Research Institute and John McConnell Math & Science Center of Western Colorado, Grand Junction, Colorado; 12) University of Toledo, Toledo, Ohio.

Lung cancer (LC) is a leading cause of death in the developed world, with over 160,000 deaths expected in the US in 2012. Environmental risk factors such as smoking and asbestos exposures are well known. However, only 15% of smokers develop LC, suggesting genetic effects or gene-environment (GxE) interactions. We previously mapped a major LC susceptibility locus to 6q23–q25, and discovered a rare risk haplotype in linked families that exhibits a GxE interaction between the 6q susceptibility locus and smoking. Genome-wide association studies have suggested other loci with common alleles of small effect on LC risk. However, these loci do not explain all familial risk of LC, suggesting that additional risk alleles exist. We have also found additional susceptibility loci using linkage analysis including environmental covariates on 6p (LOD=5.75, 74cM) and 6q (LOD=3.25, 173cM), with novel evidence of linkage on 12q24 (LOD=5.46, 150cM) and 22q11 (LOD=5.19,10cM). Linkage to lung and throat cancer was observed on 9p21 (LOD=4.97, 64cM). All analyses were on microsatellite data. Here we present the results of a fine-mapping linkage analysis, with data from the microsatellite study combined with a dense SNP map. The data were checked for Mendelian inconsistencies and low call rate and the marker allele frequencies were estimated from the data. Linkage analyses of LC (adjusting for personal smoking) to the combined microsatellite/SNP dataset using LODPAL will be presented.

1512T

Using whole exome sequencing to identify rare causal variants for oral clefts in multiplex families. *T.H. Beaty1, I. Ruczinski1, M.M. Parker1, J.B. Hetmanski1, P. Duggal1, M.A. Taub1, S. Szymczak2, Q. Li2, C. Cropp2, H. Ling3, E.W. Pugh3, Y.H. Wu-Chou4, J.E. Bailey-Wilson2, M.L. Marazita5, J.C. Murray6, E. Mangold7, M.M. Noethen7, K. Ludwig7, A.F. Scott8.* 1) School of Public Health, Johns Hopkins Univ., Baltimore, MD; 2) Inherited Disease Branch, NHGRI, NIH, Baltimore, MD; 3) Center for Inherited Disease Research, Baltimore, MD; 4) Chang Gung Memorial Hospital, Taipei, Taiwan; 5) School of Dental Medicine, Univ. of Pittsburgh, Pittsburgh, PA; 6) School of Medicine, Univ. of Iowa, Iowa City, IA; 7) Institute of Human Genetics, Univ. of Bonn, Bonn, Germany; 8) School of Medicine, Johns Hopkins Univ., Baltimore MD.

Non-syndromic oral clefts (cleft lip, cleft palate and cleft lip & palate) are common birth defects with a complex and heterogeneous etiology. Several genes and regions have been associated with risk in case-control and casefamily studies, and genome-wide linkage studies of multiplex families have identified additional regions likely to harbor causal genes. This whole exome sequencing (WES) study used 108 affected 2o and 3o relatives drawn from 52 multiplex families originally recruited for linkage (4 families with 3 affecteds and 48 families with 2 affecteds each). While the WES approach has been successful in identifying novel causal genes for Mendelian diseases, it has not previously been applied to oral clefts.
WES was done by the Center for Inherited Disease Research using the Agilent SureSelect v.4 capture reagents & Illumina HiSeq 2000 sequencers. Initially, we focused on truly novel single nucleotide variants (SNVs), i.e. not previously reported, shared by affected relatives within a family (exact genotype matches only) and predicted damaging by SIFT score (⍰0.05). A total of 516 novel SNVs were identified as shared between affected relatives in 52 families. Only one truly novel SNV (A⍰G at hg19 position 3056632) in ZNF764 was shared by affected relatives across 2 families, and these 2 families came from the same recruitment site. We focused on 334 candidate genes (plausible by molecular function, biological process or pathway) identified in Jugessur et al. (2009, PLoS ONE 4:e5385), and identified 5 genes: CDH1 (16q22), FGF8 (10q24), FGFR4 (5q35), GAD2 (10p11), and TRPS1 (8q24). Most of these genes are recognized candidates for oral clefts or cancers. Mutations in FGF8 underlie Kallman syndrome, which can include oral clefts; mutations in CDH1 have been reported in families with oral clefts and diffuse gastric cancer (MIM 192090); mutations in TRPS1 cause the abnormal craniofacial development of tricho-rhino-phalangeal syndrome. Observed SNVs included 1 Stop/Gain and 4 nonsynonymous SNVs. These affected individuals all have apparently inherited forms of oral clefts, and because they are heterozygous for damaging SNVs, they may result from haploinsufficiency of distinct gene products. Additional studies are needed to confirm causality.

1566T

The PhenX Toolkit: standard measures facilitate cross-study analyses.

C.M. Hamilton1, W. Huggins1, H. Pan1, D.B. Hancock1, J.G. Pratt1, J.A. Hammond1, T. Hendershot1, D.R. Maiese1, K.A. Tryka2, K. Sher3, K. Conway4, M. Scott5, W.R. Harlan6, J. Haines7, L.C. Strader1, H.A. Junkins8, E.M. Ramos8. 1) RTI International, Research Triangle Park, NC; 2) National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD; 3) University of Missouri, Columbia, MO; 4) National Institute on Drug Abuse, Rockville, MD; 5) National Institute on Alcohol Abuse and Alcoholism, Bethesda, MD; 6) Retired, Associate Director for Disease Prevention, Office of the Director, National Institutes of Health, Bethesda, MD; 7) Center for Human Genetics Research, Vanderbilt University, Nashville, TN; 8) National Human Genome Research Institute, Bethesda, MD.

The PhenX (consensus measures for Phenotypes and eXposures) Toolkit (https://www.phenxtoolkit.org/) is an online resource that catalogs broadly validated and well-established measures of phenotypes and exposures for use in genomic and other types of studies involving human subjects. The PhenX Toolkit currently includes 339 measures covering a broad scope of 21 research domains including Demographics, Anthropometrics, Cancer, Nutrition, Environmental Exposures, Neurology and Social Environments, and six specialty areas related to Substance Abuse and Addiction (SAA). Since its initial release in 2009, the Toolkit has logged 300,000 visits, and currently has 800 Registered Users. Toolkit visitors come from every state in the USA and 148 countries. Investigators can find measures of interest by browsing domains, collections or measures, or by searching using the Smart Query Tool. For each measure, the Toolkit provides a description of the measure, the rationale for its inclusion, detailed protocol(s) for collecting the data, and supporting documentation. The goal of the SAA project was to create six "Specialty" collections and one "Core" collection of SAA related measures. The SAA project used a streamlined version of the established PhenX consensus process to select 44 new measures which were added to the Toolkit in February, 2012. Another project, PhenX RISING (Real world, Implementation, SharING), brings together seven investigators who were awarded funds to incorporate PhenX measures into existing, population-based genomic studies. Over the course of the project, these early adopters are evaluating PhenX measures and recommending improvements to the Toolkit. The Toolkit also provides tools to help investigators integrate PhenX measures into their study design. That is, the Toolkit provides custom data collection worksheets to support data collection and custom data dictionaries to facilitate data submission to the database of Genotypes and Phenotypes (dbGaP). In an effort to link PhenX measures with data that is already in dbGaP, PhenX measures were mapped to 16 dbGaP studies. Of the 822 non-administrative variables included in these dbGaP studies, 79% could be mapped to PhenX measures. This mapping will help investigators identify studies in dbGaP that may be well-suited for various crossstudy analyses. Implications for integration of PhenX measures in dbGaP will be presented. Funding was provided by NHGRI 5U01HG004597 and 3U01HG004597-03S3.

Cardiovascular Genetics

1690T

Association of 87 traits related to coronary heart disease and rare sequence variants in the ClinSeq™ Study. *H. Sung1, B. Suktitipat1, K. Lewis2, D. Ng2, S. Gonsalves2, J.K. Teer2, N.F. Hansen3, J.C. Mullikin3,4, L.G. Biesecker2, A.F. Wilson1, NISC Comparative Sequencing Program.* 1) Genometrics Section, Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD; 2) Genetic Disease Research Branch, NHGRI, NIH, Bethesda, MD; 3) Comparative Genomics Unit, Genome Technology Branch, NHGRI, NIH, Bethesda, MD; 4) National Institutes of Health Intramural Sequencing Center (NISC), NHGRI, Bethesda, MD.

ClinSeq™ is a large scale medical sequencing study designed to investigate associations of rare sequence variants with traits related to coronary heart disease (CHD). The study currently includes 976 non-smoking patients, ages 45 to 65, with normal to severe coronary artery calcification with both traits and whole exome data. More than 200 CHD-related traits were measured at the NIH Clinical Research Center in Bethesda, MD; 87 quantitative traits were selected for the first pass analyses. Whole exome sequencing with two different capture regions was performed at the NIH Intramural Sequencing Center; 387 and 325 individuals were sequenced with the Agilent SureSelect 38 mb and 50 mb capture regions, respectively. Sequence variants (SVs) in common over both capture regions with calling rates > 98% and minor allele frequency (MAF) > 1% were used to check for misspecified population stratification by multidimensional scaling analysis - 635 Caucasians were included in these analyses. The two groups of capture regions were merged, yielding 645,364 SVs. Of these SVs, 68% and 31% had MAFs < 0.01 and < 0.001, respectively. The SVs with MAF < 0.01 were collapsed into a single derived variant for each genomic region defined by hotspot blocks. Collapsed variants were coded as the proportion of the minor allele within each region; common variants were coded as the number of minor alleles (scaled from 0 to 1). Tests of association of each SV with each trait were performed on age- and sex-adjusted traits on the untransformed, log-transformed and Box and Cox transformed data with simple linear regression. Associations between 23 traits and at least one SV were significant over both untransformed and transformed traits after Bonferroni correction (p value < 7.78e-08). Traits and the number of significant SVs (indicated in parentheses) included: coronary calcium score (208), echo right atrium pressure (117), corrected QT interval (38), direct bilirubin level (19), total bilirubin level (18), echo left atrial dimension (16), echo AV peak velocity (11), glucose (10), QT interval (8), echo E/A ratio (8), head circumference (7), platelet count (6) and progesterone (3). Tiled regression implemented in TRAP will be performed to identify the set of independently significant SVs that affect each trait. The genes for the associations will be described at the meeting.

1696T

Challenges in interpreting secondary variants from massively parallel sequencing, perspectives from the ClinSeq[TM study. *D. Ng1, J.J. Johnston1, K.L. Lewis1, S.G. Gonsalves1, L.N. Singh1, L.C. Peller1, J.K. Teer1,2, J.C. Mullikin2,3, L.G. Biesecker1,2*. 1) Genetic Disease Research Branch, NHGRI, NIH, Bethesda, MD; 2) NIH Intramural Sequencing Center, NHGRI, NIH, Bethesda, MD; 3) Genome Technology Branch, NHGRI, NIH, Bethesda, MD.

Massively parallel sequencing (MPS) has been used successfully to identify the genetic cause of rare disease (primary findings). MPS also generates incidental disease susceptibility variants (secondary findings). Interpreting secondary findings in asymptomatic individuals is a challenge as there is no consensus on the analysis and return of incidental genetic results. As part of our ongoing endeavor to study secondary variants, we selected 22 arrhythmia-associated and 41 cardiomyopathy-associated genes for analysis and present our findings. Methods: 572 whole exome sequences were annotated for variants in 63 genes with an algorithm that filtered results based on genotype quality, allele frequency, mutation type, and information in locus-specific databases. Variants were assigned pathologic scores ranging from 0–5 (0=poor genotype quality, 1=not pathogenic, 2=likely not pathogenic, 3=uncertain, 4=likely pathogenic, 5=pathogenic). Results: Seven arrhythmia-associated variants were designated likely pathogenic and three variants were pathogenic (KCNE1 p.Thr10Met, p.Arg98Trp; KCNH2 p.Arg3-12Cys). Two individuals had prolonged QTc interval (male with SCN4B p.Ser206Leu and female with SCN3B p.Leu10Pro). Three participants had relatives who died suddenly and one had a relative who died in infancy of unknown cause. There were 20 likely pathogenic and 3 definitely pathogenic (MYBPC3 IVS+1G>A; MYH7 p.Arg787Cys; PLN p.Leu39X) cardiomyopathy-associated variants. Among these 23 participants, 16 had abnormal ECG (sinus bradycardia n=13, premature ventricular contractions n=2, supraventricular contractions n=1). Echocardiogram showed one participant with a borderline thickened septum. Family history showed relatives with cardiomyopathy (n=1), congestive heart failure (n=5) and cardiac arrhythmia (n=11). Conclusion: We identified 6 asymptomatic individuals with pathogenic variants associated with Long QT syndrome (LQTS n=3), hypertrophic cardiomyopathy (HCM n=2) and dilated cardiomyopathy (DCM n=1). The incidence of these cardiac-associated variants in the ClinSeqTM cohort exceeds the estimated population frequency of LQTS (1/2500), HCM (1/500) and DCM (1/7500). The elevated incidence may be due to self-selection, false positives, or underestimation of the incidence of these disease phenotypes. Further family and functional studies are underway to address the pathogenicity, penetrance and medical management of secondary variants identified through MPS.

1702T

Cardiomyopathy as an emerging phenotype in Hereditary Inclusion Body Myopathy (HIBM). *S. Siebel1, S. McGehee2, A. Brofferio2, C. Ciccone1, M. Huizing1, L. Medne3, R. Finkel3, J. McKew4, W.A. Gahl1, N. Carrillo-Carrasco4.* 1) National Human Genome Research Institute, NIH, Bethesda, MD; 2) National Heart, Lung and Blood Institute, NIH, Bethesda, MD; 3) Children's Hospital of Philadelphia, Philadelphia, PA; 4) Therapeutics for Rare and Neglected Diseases (TRND), National Center for Advancing Translational Sciences (NCATS), NIH, Bethesda, MD.

Background: HIBM is an autosomal recessive neuromuscular disorder presenting during early adulthood with slowly progressive muscle weakness and atrophy. It is caused by GNE mutations in either the epimerase or kinase domains of this enzyme of sialic acid biosynthesis. More than 60, mainly missense, mutations have been described in different ethnic groups. The complete phenotypic characterization of HIBM is ongoing. **Case report**: Wedescribe a 37-year-old patient with HIBM and cardiomyopathy, evaluated under the Natural History Study of HIBM (CT.gov: NCT01417533). She is a compound heterozygote for missense mutations in exon 4 and exon 11 (p.V216A/p.A631V), affecting both the epimerase and kinase domains. Prior to her evaluation she denied any cardiac symptoms, but screening cardiac evaluation detected an arrhythmia with Holter monitoring; an echocardiogram showed global left ventricular hypokinesis with a reduced ejection fraction of 40%. A follow-up echocardiogram, while the patient was managed with a low-dose AT1 subtype angiotensin II receptor antagonist, showed an ejection fraction of 46%, global left ventricular hypokinesis and paradoxical septal motion. Cardiac stress MRI showed no evidence of ischemic disease. Her 29 year-old sister diagnosed with HIBM had a normal echocardiogram. **Conclusion:** We describe cardiomyopathy in a patient with HIBM and no other apparent predisposition to heart disease. The association of cardiomyopathy and HIBM remains to be established. The only other case report of dilated cardiomyopathy in HIBM was recently described in two siblings who were compound heterozygous for mutations in the kinase domain (p.F528C/A631V) (Chai, 2011). One of the mutations, p. A631V, was present in our current patient. In a recent study, cardiac myocytes derived from murine Gne-/- KO embryonic stem cells showed early degradation and a rapid decrease in their beating capacity, suggesting GNE may be an important factor in the development of cardiac tissue (Krentsis, 2011). Our case further supports the association between cardiomyopathy and HIBM, raises the possibility of a genotype-phenotype correlation and highlights the evolving and variable phenotype of HIBM. These data are encouraging to continue the systematic clinical evaluation and follow up of this patient group and the study of the role of GNE in cardiac development.

Ethical, Legal, Social and Policy Issues in Genetics

1845W

The NHGRI/NIH Clinical Sequencing Exploratory Research Program.

B.A. Ozenberger1, L.A. Hindorff1, S.D. Schully2, J. Boyer1, N. Lockhart1,2, L. Lund1, C. Mahomva1, A. Felsenfeld1, J.E. McEwen1. 1) National Human Genome Research Institute, NIH, Bethesda, MD; 2) National Cancer Institute, NIH, Bethesda, MD.

In 2010, the National Human Genome Research Institute (NHGRI) undertook an extensive expert consultation to formulate a strategic forecast for genomic research and the emerging application of genomic approaches in medicine, culminating in publication of a vision for genomics research (Charting a Course for Genomic Medicine, NATURE Feb 11, 2011). This strategic plan recognized the potential benefits to patients of comprehensive genomic data that soon will be available to clinicians with the rapid deployment of new DNA sequencing instruments and methods. NHGRI subsequently crafted the Clinical Sequencing Exploratory Research initiative to: 1) leverage the Institute's long-standing experience in genomic sequencing and analysis to ease the adoption of these methods into clinical care, 2) guide the development and dissemination of best practices for the integration of clinical sequencing into clinical care, and 3) research the ethical, legal, and psychosocial implications of bringing broad genomic data into clinical decision-making including, for example, evaluation of the risks and potential benefits associated with the return of incidental findings or information on variants of weak effect. Six grants were awarded in late 2011 [to Baylor College of Medicine, Brigham and Women's Hospital, Children's Hospital of Philadelphia, Dana Farber Cancer Institute, U of North Carolina, and U of Washington], with the National Cancer Institute cofunding two of the awards. The consortium of grantees will expand to additional sites with new awards expected to be issued in 2013. Over the next four years, these groups will generate and incorporate sequence data in the clinical care of patients and examine the relevant ethical, legal, and psychosocial issues. Several Working Groups are collaboratively defining the state of the art and exploring opportunities to coordinate efforts within and outside the consortium in the areas of: Sequencing Standards, Electronic Medical Records, Phenotype and Analysis Measures, Actionable Variants/Return of Results, Informed Consent, Outcomes and Measures for the psychosocial component of this research, and special issues relating to pediatric populations. The anticipated products and results from the Clinical Sequencing Exploratory Research consortium, and its role in NHGRI's broader efforts relating to genomic medicine, will be described.

Health Services Research

1886F

A missense mutation in Exon 2 of Adiponectin receptor 2 (ADIPOR2) is associated with serum insulin in overweight and obese African-American subjects. *A. Doumatey, G. Chen, J. Zhou, H. Huang, A.*

Adeyemo, C. Rotimi. NHGRI/CRGGH, National Institutes of Health, Bethesda, MD.

Introduction: Obesity (Ob) and overweight (OW) are associated with a number of metabolic disorders including insulin resistance, type 2 diabetes, hypertension, and dyslipidemia. Insulin is one of the most perturbed biochemical markers in OW and Ob, and it has been shown that adiponectin has an insulin-sensitizing action. Thus, we hypothesized that variants in genes encoding adiponectin and/or adiponectin receptors may influence insulin levels and action in Ob and OW persons. Methods: The subjects comprised 490 Ob and OW (BMI��25 kg/m2) African Americans from the Howard University Family study. Genotypes for SNPs in ADIPOQ, ADIPOR1 and ADIPOR2 were obtained from the Affymetrix genome- wide Human SNP array 6.0. Imputation was done using MACH with HapMap CEU and YRI as the reference population panels. A total of 169 genotyped and imputed SNPs ±50 kb around adiponectin (ADIPOQ) and adiponectin receptors (ADIPOR1, ADIPOR2) were tested for association with fasting serum insulin levels. Association models assumed an additive genetic model with adjustment for age, sex and the first two principal components of the genotypes. Results: Nine variants in AdipoR2 gene including a missense mutation (rs12298275) in exon 2 were associated with fasting insulin levels. In addition, one variant in CYB5R1 gene and four others in CACNA2D4 gene were associated with fasting insulin (p-values ranging from 4.1×10-2 to 5.4×10-3). CYB5R1 and CACNA2D4 are the closest genes to ADIPOR1 and ADIPOR2 respectively. Conclusion: These findings provide evidence for the role of adiponectin receptors in insulin metabolism in OW and Ob African Americans.

-

Complex Traits and Polygenic Disorders

1959W

Fine mapping of 6q27 for association with blood pressure. *B. Tayo1, B. Salako2, A. Luke1, X. Zhu3, A. Adeyemo4, C. Rotimi4, A. Ogunniyi2, R. Cooper1.* 1) Prev Med & Epidemiology, Loyola Univ Chicago, Maywood, IL; 2) University of Ibadan, Ibadan, Nigeria; 3) Case Western Reserve University, Cleveland, OH; 4) NIH Intramural Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, MD.

Hypertension is the most common cardiovascular condition in the world and accounts for a substantial proportion of adult mortality. Although elevated blood pressure (BP) has similar heritability to many other traits related to cardiovascular risk, genetic susceptibility loci have been difficult to localize. In a follow-up family-based association analysis of regions on chromosomes 6 and 7 linked to BP among Nigerians we observed strong preliminary evidence that the regions may influence susceptibility to elevations in BP. To refine identified association in 6q27, we conducted fine mapping analysis in a sample of 1614 unrelated adult Nigerians genotyped on Affymetrix 6.0 chip. The study sample comprised of 940 females and 674 males of which 797 were hypertensives. To increase single nucleotide polymorphisms (SNPs) density for the fine mapping, we performed genotype imputation using combined data from the HapMap and 1000 Genomes projects. A total of 3330 typed or imputed SNPs passed quality control in 6q27 and these were subsequently tested for association with BP and hypertension with covariates adjustment for sex, age and body mass index. After correction for multiple testing through Bonferroni adjustment, 7 intergenic SNPs within

70 kilobase pairs in 6q27 retained statistical significance (*P*<0.005) for association with hypertension. These data reinforce previous findings that variants in chromosome 6 influence susceptibility to elevated BP.

1983W

Preliminary data suggest an upstream IRX1 sequence variant identified in a family with kyphoscoliosis disrupts the wild-type expression pattern in zebrafish. *C.M. Justice1, K. Bishop2, B. Carrington2, P. Cruz3, K. Swindle4, R. Sood2,* N.H. Miller4, *A.F. Wilson1, NISC Comparative Sequencing Program.* 1) Genometrics Section, IDRB, NHGRI, NIH, Baltimore, MD; 2) Zebrafish Core, GMBB, NHGRI, NIH, Bethesda, MD; 3) NISC, NHGRI, NIH, Rockville, MD; 4) University of Colorado, The Children's Hospital, Denver, CO.

When idiopathic scoliosis, a structural lateral curvature of the spine ⍰⍰10;dg in individuals who are otherwise phenotypically normal, is associated with an increase in the normal thoracic kyphosis in the sagittal plane (⍰⍰40;dg), the condition is referred to as kyphoscoliosis. A previous model-independent linkage analysis of 7 families (53 individuals) with ⍰⍰2 family members with kyphoscoliosis identified candidate regions on 2q22, 5p15, 13q and 17q11, and analyses of single nucleotide polymorphism (SNP) markers narrowed the region on 5p15 to 3.5 Mb. This region contains only 3 genes, IRX1, IRX2 and IRX4, all members the Iroquois (IRX) gene family that codes for homeoproteins. The most significant linkage peak was in IRX1. The exons from IRX1, IRX2 and IRX4 were sequenced in all 53 individuals, but no functional sequence changes were identified. In this study, the highly conserved non-coding regions (HCNRs) 500 kb upstream and downstream from IRX1, IRX2 and IRX4 were sequenced in these 53 individuals in search of cis-regulatory elements. Quantitative association analysis revealed several SNPs with p-values < 0.01. The allele of 9 sequence variants (SVs; 2 of them novel) differed from the reference allele in a heterozygous state in 6 of 7 affected individuals in one family, and did not differ from the reference allele in all other individuals genotyped. One of these SVs is in a HCNR that functions as an enhancer in mice (enhancer.lbl.gov) and is 413 kb upstream from IRX1. We observed this SV in a heterozygous state (C/T) in 1 out of 90 controls. In order to determine if variation at this SV causes functional changes in vivo, an HCNR surrounding this SV (198 bp in size) was identified, amplified and cloned into a zebrafish enhancer detector (ZED) vector for Danio rerio transgenesis with the wild-type allele (C), the T allele, and the A allele reported by 1000 Genomes (MAF = 0.008). Preliminary data using transient transgenesis showed that the wild-type (C) allele drove strong expression in the midbrain and along the neural tube, while the fragments with the T and A alleles did not show a consistent definite expression pattern. We are in the process of generating stable transgenic lines for each allele. Our preliminary data suggest that sequence variation in a 198 bp HCNR disrupts the wild-type expression pattern in zebrafish and may be involved in the expression of kyphoscoliosis in humans.

2063F

Flipping sign test of GAWS summary statistics on multiple correlated traits. *Z. Zhang1, N. Franceschini2, T. Edwards3, B. Keating4, B. Tayo5, E. Fox6, A. Johnson7, Y. Sun8, Y. Sung9, M. Nalls10, J. Hunter11, A. Dresbach6, S. Musani6, G. Papanicolaou7, G. Lettre12, A. Adebowale13, R. Cooper5,*

A. Reiner14, D. Rao9, D. Levy7, X. Zhu1. 1) Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio 44106, USA; 2) Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, NC 27514, USA; 3) Center for Human Genetics Research, Vanderbilt Epidemiology Center, Department of Medicine, Vanderbilt University, Nashville, TN; 4) University of Pennsylvania School of Medicine, The Institute for Translational Medicine and Therapeutics, Philadelphia, PA, USA; 5) Department of Epidemiology and Preventive Medicine, Loyola University Stritch School of Medicine, Maywood, IL 60153, USA; 6) Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39126, USA; 7) Center for Population Studies, National Heart, Lung, and Blood Institute, Framingham, MA 01702, USA; 8) Department of Epidemiology, Emory University, Atlanta, GA 30322; 9) Division of Biostatistics, Washington University in St. Louis, MO; 10) Molecular Genetics Section, Laboratory of Neurogenetics, NIA, NIH, Bethesda, MD 20892, USA; 11) Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; 12) Montreal Heart Institute, Montréal, Canada; 13) Center for Research on Genomics and Global Health,National Human Genome Research Institute, Bethesda, MD 20892, USA; 14) Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, 98195, USA.

GWAS is a routine approach to detect the genetic determinants of complex traits, but requires a large sample size in order to overcome multiplicity and subtle effect sizes. Successful GWAS are often conducted through metaanalysis by large consortia, which usually operate by sharing summary statistics from cohorts with the same phenotypes. When multiple correlated phenotypes, such as hypertension status, systolic and diastolic blood pressure (BP) are available, multivariate analysis is often more powerful than single trait analysis. However, it is difficult to perform multivariate analysis with summary statistics. We propose a novel permutation procedure, META (Multiple-outcome Empirical Trait Analysis), for multiple correlated traits using GWAS summary statistics. META preserves the linkage disequilibrium patterns among SNPs as well as the correlation structure among traits, and adjusts significance for both multiple testing and trait correlation. META accomplishes this by randomly flipping the sign of regression effect estimates from each cohort to obtain the distribution of meta-analysis test statistics for each SNP under the null hypothesis and empirically estimating significance. Our simulations show that the proposed permutation method can improve power up to 40%; compared to single trait analysis. We applied our method to the GWAS summary statistics from the Continental Origins and Genetic Epidemiology Network (COGENT) consortium study of BP traits, which includes 19 African ancestry cohorts. The proposed method identified two additional novel BP/hypertension loci that were not detected by the single trait analysis. Our results suggest that a multivariate approach should be used when multiple correlated traits are studied, and furthermore that this enhanced analysis can also be done using summary statistics.

2108F

Ontology, visual, and informatics enhancements to the NHGRI Genome-wide Association Study (GWAS) catalog. *L.A. Hindorff1, J.A.L. MacArthur2, D. Welter2, T. Burdett2, P. Hall1, H.A. Junkins1, H. Parkinson2.* 1) Population Genomics, NHGRI, NIH, Bethesda, MD; 2) EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, UK.

The National Human Genome Research Institute's (NHGRI) GWAS Catalog (http://www.genome.gov/gwastudies) contains over 1,300 publications and 6,400 disease- or trait-associated genetic variants, which have been manually curated from published genome-wide association studies (GWAS). The availability of this curated, regularly updated, and downloadable resource contributes to the knowledge base of genetic variants associated with important diseases and conditions and enables cross-linking with other commonly used resources such as Ensembl, Database of Genotypes and Phenotypes (dbGaP), the Phenotype-Genotype Integrator (PheGenI) and UCSC Genome Browser. We have implemented several improvements to increase the usability of catalog data and ease of integration with other resources. First, catalog traits have been integrated into the Experimental Factor Ontology (Malone et al., 2010), an application ontology based on a paradigm of reusing elements from reference ontologies including PATO and ChEBI. This standardization of trait organization will facilitate increased integration of NHGRI GWAS catalog data with other sources and improved querying of the catalog. Building upon this ontology, generation of the iconic GWAS catalog diagram has been automated (http://wwwdev.ebi.ac.uk/fgpt/gwas) such that visual representations of GWAS associations across the human genome are now available filtered by different criteria, including disease category and p-value. These customized diagram images can be downloaded for use in publications and presentations. SNP-trait associations in the online version of the diagram are also interactive, with dynamic links to the GWAS catalog entry and to external resources such as Ensembl and dbSNP. Informatics improvements have also increased the efficiency and quality of the curation process. For example, the previous limit of 50 SNPdisease associations per paper is no longer in effect, with all qualifying associations with a p-value < 1x10−5 now included in the catalog. Ultimately, the benefits of standardized disease annotation, customized views of the GWAS catalog diagram, and efficient curation will yield an enhanced user experience and facilitate further integration of GWAS catalog data with other resources.
that this enhanced analysis can also be done using summary statistics.

2151W

A genome-wide association study identifies susceptibility loci for nonsyndromic sagittal craniosynostosis on chromosomes 20 and 7. *G. Yagnik1, 2, C.M. Justice3, 2, Y. Kim3, I. Peter4, E.W. Jabs4, X. Ye4, L. Shi4, M.L. Cunningham5, V. Kimonis6, T. Roscioli7, S.A. Wall8, A.O.M Wilkie8, 9, J. Stoler10, J.T. Richtsmeier11, Y. Heuzé11, P.A. Sanchez-Lara12, M.F. Buckley13, C.M. Druschel14, J.L. Mills15, M. Caggana16, P.A. Romitti17, D.M. Kay16, C. Senders18, P.J. Taub19, O.D. Klein20, J. Boggan21, C. Naydenov22, J. Kim1, A.F. Wilson3, S.A. Boyadjiev1.* 1) Department of Pediatrics, Section of Genetics, University of California, Davis, Sacramento, CA; 2) Authors with equal contribution; 3) Genometrics Section, IDRB, Division of Intramural Research, NHGRI, NIH, Baltimore, MD; 4) Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY; 5) Department of Pediatrics, Division of Craniofacial Medicine, University of Washington and Seattle Children's Research Institute, Seattle, WA; 6) Division of Genetics, Department of Pediatrics, University of California Irvine, Irvine, CA; 7) School of Women's and Children's Health, Sydney Children's Hospital, University of New South Wales, Sydney, Australia; 8) Craniofacial Unit, Oxford University Hospitals NHS Trust, John Radcliffe Hospital, Oxford

OX3 9DU, UK; 9) Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK; 10) Division of Genetics, Children's Hospital Boston, Harvard University, Boston, MA; 11) Department of Anthropology, Pennsylvania State University, University Park, PA; 12) Division of Genetics, Department of Pediatrics, University of South California, Los Angeles, CA; 13) Department of Haematology and Genetics, SEALS, Sydney, Australia; 14) Congenital Malformations Registry, New York State Department of Health, Albany, NY; 15) Division of Epidemiology, Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services, Bethesda, MD; 16) Division of Genetics, Wadsworth Center, New York State Department of Health, Albany, NY; 17) Department of Epidemiology, College of Public Health, The University of Iowa, Iowa City, IA; 18) Department of Otolaryngology, University of California Davis, Sacramento, CA; 19) Division of Plastic and Reconstructive Surgery, Kravis Children's Hospital, Mount Sinai Medical Center, New York, NY; 20) Departments of Orofacial Sciences and Pediatrics and Program in Craniofacial and Mesenchymal Biology, University of California San Francisco, San Francisco, CA; 21) Department of Neurological Surgery, University of California Davis, Sacramento, CA; 22) Department of Chemistry and Biochemistry, Medical University, Sofia, Bulgaria.

Sagittal craniosynostosis is a common congenital malformation, affecting approximately one out of 5,000 newborns. We conducted the first genomewide association study (GWAS) for non-syndromic sagittal craniosynostosis (sNSC) using 130 non-Hispanic white (NHW) case-parent trios. Robust associations were observed in a 120 kb region downstream of *BMP2* on chromosome 20p.12.3, flanked by rs1884302 (*P* ⍰⍰1.13 x 10−14; odds ratio [OR] = 4.58) and rs6140226 (*P* ⍰⍰3.4 x 10−11; OR = 0.24) and within a 167 kb region of *BBS9* on chromosome 7p14.3 between rs10262453 (*P* ⍰⍰1.61 x 10−10; OR=0.19) and rs17724206 (*P* ⍰⍰1.50 x 10−8; OR = 0.22). We replicated these associations for rs1884302 (*P* ⍰⍰4.39 x 10−31) and rs10262453 (*P* ⍰⍰3.50 x 10−14) in an independent NHW population of 172 unrelated sNSC probands and 548 unaffected controls. Both *BMP2* and *BBS9* implicated by these associations are biologically plausible genes with a role in skeletal development warranting functional studies to further understand the etiology of sNSC.

2155T

Ethnicity and ancestry information from genome-wide association studies: The NHGRI GWAS Catalog. *H.A. Junkins1, J.A.L. MacArthur2, P. Hall1, K.A. Harvey3, T.A. Manolio1, L.A. Hindorff1.* 1) Office of Population Genomics, NHGRI, National Institutes of Health, Bethesda, MD; 2) EMBLEBI, Wellcome Trust Genome Campus, Hinxton, UK; 3) Centers for Disease Control and Prevention, Atlanta, GA.

The impact of genome-wide association studies (GWAS) on the knowledge base of disease-associated variants has been vast, with over 1,300 publications and 6,400 disease- or trait-associated genetic variants identified to date (http://www.genome.gov/gwastudies). However, much of this knowledge is based upon studies of predominantly European-descent individuals, posing the question of whether these variants will also generalize to non-European populations and, importantly, to populations with disproportionate disease burdens. As part of the curation efforts for the National Human Genome Research InstituteGWAS Catalog, we developed a framework for systematically and consistently extracting ethnicity and ancestry information at multiple levels of granularity. Of 443 papers from January, 2011 through April, 2012 (35% of all papers), 56% exclusively report populations of European descent; 11% include both European and non-European populations, and 28% report only non-European populations. On a per-participant level, most participants were also classified as European descent. Of studies that included participants of non-European descent, the most frequently represented countries were China, Japan, Korea, and Singapore. Representation of non-European participants was most pronounced for the following traits: blood pressure, body mass index, HDL cholesterol and type 2 diabetes. Papers including non-European participants were more likely to be published in high impact journals (impact factor >7; OR 1.64, 95% CI 1.04–2.60), adjusted for study year, sample size, and previous GWAS publications. Preliminary data from a small study of 153 papers published from 2005–2008 suggest that a variety of statistical methods were used to account for population stratification in the analysis of GWAS data; an increased number of such methods was associated with use of more than one ethnic group in the discovery set, later publication date and publication in high-impact journals. These data need to be extended to more recent studies, which through meta-analyses of multiple cohorts have increased in complexity. Although these data are limited to GWAS published in 2011–2012, our work suggests that non-European participants are still greatly underrepresented in GWAS and that additional studies are needed to fully evaluate whether potentially functional variants identified in GWAS generalize to populations most impacted by chronic disease.

2219F

A copy number variant at the KIT ligand locus confers risk for canine squamous cell carcinoma of the digit. *E.A. Ostrander1, B. Decker1, E. Carlins1, B. VonHoldt2, G. Carpintero-Ramirez1, H.G. Parker1, R.K. Wayne2, D.M. Karyadi1*. 1) Cancer Genetics Branch, NHGRI/NIH, Bethesda, MD; 2) Dept of Ecology and Evolutionary Biology, UCLA, Los Angeles CA.

The domestic dog is a robust model for studying the genetics of disease susceptibility as many breeds harbor an elevated risk for certain cancers, despite the simple population structure common to all breeds. One example of breed-specific disease propensity is squamous cell carcinoma of the digit (SCCoD), a locally aggressive cancer that frequently causes lytic bone lesions and multiple toe recurrence. SCCoD is uncommon in most dog breeds, but is prevalent in Standard Poodles. Intriguingly, Poodles with dark coat color are at high risk for SCCoD, whereas Poodles of light coat colors are entirely unaffected, suggesting that interactions between multiple pathways are necessary for oncogenesis. We performed a genome-wide association study on Standard Poodle SCCoD cases compared to unrelated black Poodle controls. Allelic association was calculated with the single-locus chisquared significance test. The six most strongly associated SNPs (Praw= 5.62x10-5-1.20x10-7) were contiguous markers on canine chromosome 15 and all were statistically significant at the chromosome-wide level (Pempirical= 0.0299-7.00x10-5). Comparison of Poodle cases to other at-risk breeds initially refined the locus to 144.9Kb, a region that harbors 128 variants segregating with risk. The initial risk-associated haplotype is present in 92.9% cases, with 55% of cases are homozygous for the risk allele. While 52.8% of controls are heterozygotes, none are homozygous for the risk haplotype. The region lies upstream from the KIT Ligand (KITLG) gene. Fine mapping and sequencing reduced the locus to a minimal region of 29.7Kb. A dosage dependent effect of a CNV within the region is strongly associated with the disease phenotype in Standard Poodles. Allele-specific expression assays are ongoing to determine the relationship between the CNV and KITLG expression. Additional investigation highlighted a compensatory nonsense mutation in light colored standard poodles that explains their lack of disease, even if they carry the risk allele. The presence of disease is not due to coat color pigments, i.e. light colored Briards lack the compensatory mutation and get the disease if they carry the risk allele, similar to black Briards. This work highlights the utility of the canine system for disentangling the genetics of multigene cancer disorders. Our ongoing studies should highlight other genes contributing in the disease and, by extension, what role these genes play in human cancers.

Psychiatric Genetics, Neurogenetics & Neurodegeneration

2442W

Identification of a dosage-sensitive brain development gene within the critical region of 1q deletion syndrome. *E.A. Erickson1,2, W.A. Gahl2, C. Boerkoel1.* 1) NIH Undiagnosed Diseases Program Translational Laboratory, Bethesda, MD; 2) Section on Human Biochemical Genetics, Medical Genetics Branch, NHGRI, NIH, Bethesda, MD.

Submicroscopic terminal 1q deletion results in a syndrome consisting of severe psychomotor delays, aphasia, hypotonia, microcephaly, corpus callosum abnormalities, and facial dysmorphism. Though previous studies have narrowed the critical region for this phenotype to 1q43–1q44, the deleted genes responsible for the syndrome remain unknown. A potential contributor to this syndrome is the uncharacterized gene SCCPDH, which lies within the 1q44 critical region. We chose to investigate the effects of SCCPDH deletion on early brain development in zebrafish, since their development is easily observed and manipulated. RT-PCR analysis determined that the zebrafish SCCPDH homologue is normally expressed throughout development with its highest expression during the earliest stages. The SCCPDH homologue was knocked down in zebrafish by postfertilization embryo injection of a synthetic antisense oligomer, or morpholino (MO), designed to bind and block a splice site within SCCPDH pre-mRNA. At 24 and 48 hours post fertilization, the embryos had a dose-dependent defect in midbrain-hindbrain boundary formation and hydrocephalus, respectively. This same phenotype was observed with another MO that inhibits translation of SCCPDH transcripts. Demonstrating that the effects of the MO are specific to SCCPDH knockdown, co-injection of SCCPDH mRNA and the SCCPDH splice-site targeting MO rescued the phenotype; over 70% of the co-injected embryos were morphologically similar to the uninjected control embryos. Interestingly, the overexpression of SCCPDHmRNA alone caused cyclopia and poor brain development. These results indicate that the level of SCCPDH expression is highly regulated and that alterations beyond a threshold can result in abnormal brain development. This work suggests that SCCPDH plays an important role in early brain development and merits further investigation

2610W

Changes in the Human Transcriptome Caused by LMNB1 Duplication in a Case of Autosomal Dominant Leukodystrophy: An RNASeq study.

P. Cherukuri1, D. Simeonov1, K.V. Fuentes-Fajardo1, P. Zumbo2, C. Mason2, S. Lin3, Y.H. Fu3, C. Boerkoel1, T. Markello1, W. Gahl1, C. Toro1.
1) Undiagnosed Diseases Program, NHGRI / NIH, Bethesda, MD; 2) Department of Physiology and Biophysics, Weill Medical College, Cornell University, New York, NY; 3) Department of Neurology, University of California, San Francisco, San Francisco, California.

Autosomal dominant leukodystrophy (ADLD) is a rare, progressive, adultonset demyelinating disorder caused by duplications of the nuclear lamina gene, lamin B1 (LMNB1). LMNB1 is part of the stable fibrous meshwork of intermediate filaments (IFs) that underlie the nuclear membrane, forming the lamina. In addition, LMNB1 is tethered to chromatin at lamina-associated domains and has a role in the regulation of gene expression and gene silencing. In this study, we investigated the global impact of LMNB1 locus duplication on the expression profile of the poly-adenylated RNA fraction of the human transcriptome using massively parallel sequencing (RNASeq). 100ng of high-quality total RNA (RIN>0.8) for each experimental condition was isolated from primary fibroblast cell lines, and cDNA was prepared with the TruSeq mRNA prep kit (v2). Multiplexed cDNA libraries for each triplicate sample (2 samples (1 case & 1 normal); 6 datasets) were sequenced to 50×50 bp on the Illumina HiSeq2000, and post-processed with CASAVA (v1.8.2). Raw data were filtered for high-median quality (Q-value > 20) and then a total of 426 million paired-end (PE) reads (~71 million PE reads perreplicate) were processed on a High Performance Computing (HPC) cluster (SGE) through different RNASeq analysis pipelines (PPBS and ERANGE). Gene expression was quantified as reads-per-kilobase-per-million (RPKM). To test for dysregulation or disruption of gene-expression profiles of clusters of genes in genomic neighborhoods, we developed a computational approach to detect movement of genes' transcription outside of a stable boundary condition. Using this approach, we found that ~80 highly reproducible gene-clusters were significantly dysregulated. These clusters were significantly enriched (P-value=4.7e-29; hypergeometric distribution) for chromatin regions associated with interaction with the lamina. We are pursuing exploring the mechanism underlying these alterations in expression and evaluating genes within these boundaries for robustness of differential expression. This study presents a hypothesis for the mechanism underlying the neurodegeneration.

Molecular Basis of Mendelian Disorders

2838F

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. Cullinane1, J. Pan1, M.A. Merideth1, J.A. Curry1, J.G. White2, M. Huizing1, W.A. Gahl1.* 1) Medical Genetics Branch, NHGRI (NIH), Bethesda, MD; 2) Department of Laboratory Medicine, University of Minnesota, Minneapolis, MN.

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. Here we report a girl who had previously been screened for mutations in HPS-1 through HPS-6 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 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 directed to LROs. We demonstrated that TYRP1 interacts with the AP-1, AP-2 and AP-3 complexes, allowing this trafficking to occur. ATP7A, however, 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-3 complex) and directed to LROs.

2870T

The NIH Undiagnosed Diseases Program: Defining Pathogenicity for Personalized Medicine. *C.F. Boerkoel III, M.C. Malicdan, D. Adams, C. Toro, C. Tifft, W.A. Gahl, T. Markello.*NIH Undiagnosed Diseases Program, NIH/NHGRI, Bethesda, MD.

Human disease arises by maladaptation of humans to their ecological niche. Rare diseases, which affect ~8% of the population, frequently arise from strong genetic and epigenetic mutations causing maladaptation within a stable ecological niche. Within this context, a precise diagnosis is the first step to understanding illness and defining appropriate therapies. The NIH Undiagnosed Diseases Program uses genomics to approach diagnoses for ill individuals whose diseases have otherwise eluded identification. Even extending the genomic analysis to the community of the family, we confront the difficulty of defining what sequence variants are responsible for disease. In the absence of an epidemiological association, we consider four levels of evidence for pathogenicity: 1) statistical likelihood that the variant will occur in healthy individuals; 2) empiric evidence for pathological consequences of the variant on protein function; 3) associative evidence from systems biology considerations; and 4) recapitulation of disease in a model organism. Based on a study of a rare distal myopathy identified in a single family, we conclude that, in some instances, a systems biology approach garners sufficient evidence to prove disease causation and thereby abrogates the need for epidemiological association or attempt at full recapitulation of the disease in model organisms.

2938W

A single exome variant is the only expected variant by likelihood ratio for a rare heritable de novo dominant disorder in a three generation family with two affected. *S.M. Marchegiani*[1,5], *T.C. Markello*[1,2,3], *L.A. Wolfe*[1], *K. Fuentes-Fajardo*[1], *D.R. Adams*[1,2,3], *W.A. Gahl*[1,2,3], *J.C. Mullikin for NISC*[4], *T. Davis*[1], *J.P. Accardi*[1], *C.J. Tifft*[1,3], *C.F. Boerkoel*[1,2,3]. 1) NIH Undiagnosed Diseases Program, Bethesda, MD; 2) Medical Genetics Branch, National Human Genome Research Institute; 3) Office of the Clinical Director, National Human Genome Research Institute; 4) NIH Intramural Sequencing Center, National Human Genome Research Institute; 5) Department of Pediatrics, National Capital Consortium, Bethesda, MD.

High throughput sequencing (HTS) refined the predicted mutation rate to 1.3 × 10-8 per base pair per generation, which yields approximately 1.3 (n= 0-4, P>0.95) mutations per diploid exome per generation. Given the known mutation rate and size of the exome, there is a P<0.05 probability that exome sequencing (ES) will identify more than one variant that is both deleterious and consistent with inheritance in the case where a de novo mutation occurred in a parent and correctly segregates into two offspring in a third generation. We illustrate this concept by identifying one novel variant in a de novo affected father and his affected daughter within a three generation family. The father and daughter carry the diagnosis ablepharon macrostomia (AMS), a rare presumed dominant disorder characterized by absent eyelids, dysmorphic facies, lax redundant skin and ambiguous genitalia. Exome sequencing identified a T>A substitution in the 5'UTR of TRIM23, a well-conserved ADP-ribosylation factor (ARF) gene on Chromosome 5 that has a role in vesicular trafficking. The variant from exome sequencing is only present in the affected father and daughter, and is absent in all unaffected family members, including both paternal grandparents. This variant is not reported in current databases. Consanguinity was excluded by high density single nucleotide polymorphism (SNP) array analysis. Pseudodominant inheritance is unlikely for a rare disease in an outbred family.

Evidence of (non)penetrance was critically examined during clinical phenotyping, including comprehensive eye and skin exams with light and EM histopathology. We confirmed complete absence of features consistent with ablepharon macrostomia in the unaffected paternal grandparents, unaffected mother and unaffected brother, supporting a de novo dominant mutation in the father with subsequent inheritance by his daughter.Weadditionally established elastic fiber electron microscopy as another modality to clinically phenotype affected versus unaffected individuals in this family. The presence of a rare, heritable de novo dominant disorder in a three generation family provides an ideal model to propose that an identified and Sanger validated variant, like the one described for TRIM23 in this pedigree, is the only expected variant and thus likely causal for the condition in that family.

3070F

Clinical and molecular characterization of non-syndromic craniosynostosis: an International Consortium Approach.

M.L. Cunningham1, P.A. Romitti2, C.M. Justice3, A.F. Wilson3, T. Roscioli4, E. Oláh5, B. Bessenyei5, M.R. Passos-Bueno6, B. Wollnik7, A.O.M. Wilkie8, S.A. Boyadjiev Boyd9, International Craniosynostosis Consortium. 1) Dept of Pediatrics, Division of Craniofacial Medicine, University of Washington and Seattle Children's Research Institute, Seattle, WA, USA; 2) Department of Epidemiology, College of Public Health, The University of Iowa, Iowa City, IA; 3) Genometrics Section, IDRB, Division of Intramural Research, NHGRI, NIH, Baltimore, MD; 4) School of Women's and Children's Health, Sydney Children's Hospital, University of New South Wales, Sydney, Australia; 5) Department of Genetics, Debrecen, Hungary; 6) Department of Genetics, University of Sao Paolo, Brazil; 7) Institute of Human Genetics, University of Cologne, Germany; 8) Craniofacial Unit, Oxford University Hospitals NHS Trust, John Radcliffe Hospital, Oxford, United Kingdom; 9) Department of Pediatrics, Section of Genetics, University of California Davis, Sacramento, CA.

Non-syndromic craniosynostosis (NSC), the premature closure of one or more of the cranial vault sutures is a common congenital malformation, affecting approximately one out of 2,000 newborns. Rare mutations in the FGFR2, TWIST1, FREM1, LRIT3, EFNA4, ALX4, and RUNX2 duplications have been reported in a minor fraction of NSC cases; however, the etiology of this birth defect is not well understood. In an attempt to better characterize the clinical phenotype and to identify susceptibility loci and contributing environmental factors we have established an International Craniosynostosis Consortium (ICC) coordinating the effort of physicians and researchers at 10 sites in the United States, 4 craniofacial units in the United Kingdom, the Hungarian National Registry for Birth Defects and several major craniofacial centers in Australia, Brazil and Germany. A total of 741 families have been recruited and characterized and clinical databases with sample repository were created. A study website allows collaborators, referring physicians and members of the affected families to contribute to the goals of ICC (https://genetics.ucdmc.ucdavis.edu/icc.cfm). As a result, RUNX2, ALX4, FREM1, and LRIT3 variants predisposing to NSC were identified and reported by our group. The findings of the first genome-wide association study suggested a role for BMP2 and BBS9 in sagittal NSC. The goal of this paper is to present our current approach and to inform the genetics community of possible ways to contribute to the mission of the ICC (contact: simeon.boyd@ucdmc.ucdavis.edu).

3080F

Genotype/Phenotype correlation in Smith-Magenis Syndrome with abnormal 17p deletions.

T. Vilboux1, A.C.M Smith2, S. Chandrasekharappa3, C. Ciccone1, J. Blancato4, W.J. Introne2, W.A. Gahl1,2, M. Huizing1.
1) Medical Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD; 2) Office of the Clinical Director, National Human Genome Research Institute, NIH, Bethesda, MD; 3) Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD; 4) Department of Oncology, Georgetown University Medical Center, Washington, DC.

Smith-Magenis syndrome (SMS) is a complex developmental disorder characterized by an interstitial deletion of chromosome 17p11.2. This syndrome includes variable symptoms such as cognitive impairment, craniofacial dysmorphia, height-growth delay, infantile hypotonia, brachydactyly, attention deficit, decreased sensitivity to pain, self-injury, maladaptive behaviors, speech delay, hearing loss, and sleep disturbance. Most SMS patients have the same approximately 3.7Mb interstitial 17p11.2 genomic deletion. Haploinsufficiency of the *RAI1* gene is likely responsible for most of the SMS features, but haploinsufficiency of other genes within the deleted region may also contribute to the phenotype ofSMSpatients with 17p11.2 deletions. To further investigate the molecular basis of the SMS phenotype we studied phenotype/genotype correlation of 10 SMS patients with abnormal 17p deletions. We identified exact deletion sizes by multiple genomic techniques, including FISH (fluorescent in situ hybridization), MLPA (Multiplex Ligationdependent Probe Amplification) and quantitative PCR analyses. Furthermore, each patient's DNA was analyzed on a SNP-microarray (Illumina chip) to define the exact 17p breakpoints. With these data we refined the minimal Smith-Magenis syndrome critical region to an approximately 500-kb interval on 17p11.2 that includes RAI1 and 6 other genes. A deeper analysis of the genes in or outside this minimal critical region provided us with candidate genes for some of the unusual phenotypes of certain patients, including, obesity, severe speech delay, hearing-loss or immune defects. Analysis of genotype/phenotype correlations in patients with unusual deletions in complex syndromes like SMS may provide candidate genes for specific phenotypes within the syndrome. These findings may not only benefit the studied syndrome, but may also assist in molecular analysis of more common phenotypes (e.g. for SMS: autism, sleep abnormalities, obesity, speech delay, maladaptive behaviors).

3306T

Reproduction and immunity driven natural selection in the hominid *WFDC* locus. *Z. Ferreira1, 2, 3, S. Seixas2, A. Andres4, W. Kretzschmar5, J. Mullikin1, 6, W. Swanson7, M.K. Gonder8, S. Tishkoff9, A. Stone10, A.G. Clark11, E. Green1, 6, B. Hurle1, NIH Intramural Sequencing Center, Bethesda, MD.* 1) NHGRI, NIH, Bethesda, MD; 2) IPATIMUP, Porto, Portugal; 3) Department of Zoology and Anthropology, Faculty of Sciences, University of Porto, Porto, Portugal; 4) Genetic Diversity and Selection, Department of Evolutionary Genetics - Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; 5) Genomic Medicine and Statistics, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; 6) NIH Intramural Sequencing Center, Bethesda, MD; 7) University of Washington Seattle, WA; 8) Department of Biological Sciences University at Albany, State University of New York, Albany NY; 9) Departments of Genetics and Biology, University of Pennsylvania, Philadelphia, PA; 10) Department of Anthropology, Arizona State University, Tempe, AZ; 11) Department of Biology of Molecular Biology and Genetics, Cornell University, Ithaca, NY.

The *WFDC* and *SEMG* genes warrant a concerted evolutionary analysis because of their strikingly high *KA*/*KS*, indicative of response to adaptive pressures during vertebrate evolution. *WFDC* is the whey acidic protein (WAP) four-disulfide core domain locus located on human chr20q13. This locus spans 19 genes withWAP and/or Kunitz domains. These genes participate in antimicrobial, immune, fertility and tissue homoeostasis activities. WFDC-related genes include nearby genes encoding seminal proteins

Semenogelin 1 and 2 (SEMG1 and SEMG2). To understand better the selection pressures acting on *WFDC* genes in human populations, we sequenced 17 genes and 54 non-coding tags in 71 European (CEU), African (YRI) and Asian (CHB+JPT) individuals. We identified 484 Single Nucleotide Polymorphisms (SNPs), including 65 coding mutations, of which 49 were non-synonymous substitutions. Using classic neutrality tests we confirmed a signature of short-term balancing selection on *WFDC8* in Europeans; and a signature of positive selection spanning genes *PI3*, *SLPI*, *SEMG1* and *SEMG2*. Associated with the latter signal, we identified an unusually homogeneous derived haplotype with a frequency of 88% in Asians. A putative candidate variant targeted by selection is Thr56Ser in SEMG1, which may alter the proteolytic profile of SEMGI and antimicrobial activities of semen. Ancestral Thr56 was traced to the last common ancestor of SEMG1 in Old World Monkeys and hominoids 25 million years ago (MYA) - with derived allele Ser56 arising less than 0.25 MYA. These results are all consistent with patterns of variation seen in the 1000 genomes data as well. To examine if the adaptive pressures and patterns of genetic variation within the *WFDC* locus differ among hominoids, we sequenced 68 chimpanzees (*P. t. troglodytes*, *P. t. verus* and *P. t. ellioti*), the most closely related species with a different mating system and exposure to pathogens. For that, we generated 13 Mb of high-quality sequence data, identifying 847 SNPs, and we calculated summary statistics for each gene. Both the chimpanzee data and the human-chimp joint analysis indicate selective signals in the *EPPIN* and *WFDC6* genes, which have been shown to have active roles in primate fertility and immune response. This study provides further evidence that the *WFDC*s and *SEMG*s have been under strong adaptive pressures within hominid evolution, improving our knowledge of biological dynamics of rapidly evolving genomic regions in primates.

Bioinformatics and Genomic Technology

Bioinformatics and Genomic Technology
3533F

Simultaneous analysis of common and rare variants in complex traits.

G. Chen1, A. Yuan2, A. Bentley1, D. Shriner1, A. Adeyemo1, C. Rotimi1. 1) NHGRI/NIH, National Institutes of Health, Bethesda, MD; 2) National Human Genome Center, Howard University, Washington DC, USA.

Genome-wide association studies (GWAS) have facilitated the detection of common genetic variants underlying common traits and diseases. However, significant proportions of the underlying genetic variance for the traits investigated to date remain unexplained. Also, the predictive power of common variants identified by GWAS has not been encouraging. Given these observations along with the fact that the effects of rare variants are often unaccounted for by GWAS and the increasing availability of sequence data, we developed a new method that enables the simultaneous analysis of the association between individual common and rare variants while controlling for the potential confounding effects of covariates. We refer to this method as SCARVAsnp. SCARVAsnp is implemented in four stages: first, all common variants in a pre-specified region (e.g. gene) are evaluated individually; second, a union procedure is used to combine all rare variants (RVs) in the index region and the ratio of the log likelihood with one RV excluded to the log likelihood of a model with all the collapsed RVs is calculated. On the basis of simulation studies, a likelihood ratio □□1.3 is considered statistically significant; third, the direction of the association of the removed RV is determined by evaluating the change in □□values with the inclusion and exclusion of that RV. Lastly, significant common and rare SNPs along with covariates are included in a final regression model to evaluate the association between the index trait and variants in a pre-specified genomic region. We use simulated and real data sets to show that the method is simple to use, computationally efficient and that it can accurately identify both common and rare risk variants. This method overcomes several limitations of existing methods; for example, SCARVAsnp limits loss of statistical power by not including rare variants that are not associated with the trait of interest in the final model. The combined analysis of rare and common variants may be important in explaining the missing heritability from GWAS analyses. Also, SCARVAsnp takes into consideration the direction of association by effectively modelling positively and negatively rare associated variants.

3543F

Iterated Correction of a Diploid Parental Reference Sequence and its use during Exome Alignments for Disease Gene Discovery. *T.R. Gall1, D.R. Adams1,2,3, C.F. Boerkoel1, K. Fuentes-Fajardo1, W.A. Gahl1,2,3, P. Cherukuri1, M. Sincan1, C. Toro1, C.J. Tifft1,3, T.C. Markello1,2,3.* 1) NIH Undiagnosed Diseases Program, NIH, Bethesda, MD; 2) Medical Genetics Branch, NHGRI, NIH; 3) Office of the Clinical Director, NHGRI, NIH.

A major remaining problem in using next-generation sequencing is the substantial number of mismapped and misaligned sequences that lead to false-positive and false-negative errors when using alignment to a reference genome sequence as the method to reconstruct a patient's genome from short read sequences. This is due to the fact that the degree of difference between any one individual and the reference genome is no less than the difference between any one gene and an evolutionary homologue resulting from duplication and divergence that occurred within a family of genes. This is equivalent to low complexity since these closely related sequences require a longer read length to uniquely identify different locations, when compared to what is expected by a random sequence data. One way to address this is to use information, obtained independently and not from the short-read sequencing data, to improve the reference genome. The purpose of this modification is to more closely match the reference to the exome sequence being reconstructed. SNP chip data, haplotype data from the inheritance state of a nuclear family, and parental sequence data are all readily available sources of independent information that can be applied to resolve lowcomplexity region mapping and alignment ambiguities. We have established and tested a bioinformatic pipeline that takes SNP and parental sequence data from trios and finds the most likely alleles for the 4 independent haplotypes, generates a set of alignments, and then iterates these results for 3 cycles to improve the reference sequence used to generate the final alignment output for the patient. In addition simultaneous multiple sequence genotype calling, together with inheritance state-aware genotype calling, is used to produce the final variant list. This pipeline has produced measurable improvements from a simulated data set and for two trio data sets from a quartet sequenced in the Undiagnosed Diseases Program.

3616W

Comparing Protein Prediction Methods Using Disease-Causing Missense Variants. *P. Duggal1, Y. Kim2, M.K. Tilley2, M.M. Parker1, A. Maroo1, A.P. Klein1,3.* 1) Dept Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; 2) Inherited Disease Research Branch, NHGRI, NIH, Baltimore, MD; 3) Department of Oncology and Pathology, Johns Hopkins University, Baltimore, MD.

The success of next-generation sequencing studies to identify genetic variation associated with disease is dependent upon the ability distinguish between genetic variation that leads to a change in protein function and benign variation. Deciphering the role of these identified variants is often difficult because of low allele frequencies. Computational methods that classify the functionality of these variants are used to filter variants and to weight the relevance or importance of a given missense mutation in statistical analyses. We manually curated the Uniprot database, and the clinically associated variant list from the Human Genetic Mutation Database and identified 150 variants present in 1000 genomes data regardless of ethnicity. Using the OMIM database we identified 43 missense mutations from the 150 with published evidence of a functional role for the mutation, including differences in animal studies, clinical assays and binding affinities. We considered these 43 variants to be a “gold standard” since there were appreciable differences in function associated with each genetic change. Using 9 different prediction programs, we evaluated and compared the prediction performance. The programs used in this study were: SIFT, SNAP, PolyPhen-2, PMUT, VarioWatch, The Grantham Matrix Score, GERP++, SNPs&Go and MAPP. Depending on the algorithms and sources used in each protein prediction method, the prediction results were varied. Overall, the percent of variants predicted to be deleterious ranged from 51–77% using each program. Although for any given variant the percent of programs that predicted it to be deleterious ranged from 17–100%. These programs have been evaluated previously, but the use of a “gold standard” in which function was shown to be evident in the literature should provide the most informative comparison. The discrepancy in programs raises concerns about the use of an individual program to predict functionality in filtering or weighting analyses, since these may result in the exclusion of truly functional variants. This study suggests that multiple programs be considered when evaluating novel variants to limit the risk of false negatives, which will be especially important for complex diseases.