**Human Genetics and Genomic Medicine**

**Gholson Lyon**

with Max Doerfel, Jason O'Rawe, Yiyang Wu, Prashant Kota

The Lyon laboratory focuses on analyzing human genetic variation and its role in severe idiopathic neuropsychiatric disorders. We do this by studying large pedigrees living in the same geographic location, where one can study the expressivity and segregation of variants in a similar environmental background and with fewer population stratification concerns. Toward this end, we collect pedigrees in Utah and elsewhere, and then utilize exome and whole genome sequencing to find mutations that segregate with syndromes in the pedigrees. We focus on the discovery of families with rare diseases and/or increased prevalence for syndromes such as intellectual disability, autism and schizophrenia.

**X-linked Malformation and Infantile Lethality Syndrome (provisionally named Ogden Syndrome)**

with Max Doerfel, Yiyang Wu, Thomas Arnesen (Norway)

We have previously identified a lethal X-linked disorder of infancy comprising a distinct combination of distinctive craniofacial features producing an aged appearance, growth failure, hypotonia, global developmental delays, cryptorchidism, and acquired cardiac arrhythmias. The first family was identified in Ogden, Utah, with five affected boys in two generations of family members. A mutation was identified as a c.109T>C (p.Ser37Pro) variant in *NAA10*, a gene encoding the catalytic subunit of the major human N-terminal acetyltransferase (NatA). This same mutation was identified in a second unrelated family, with three affected boys in two generations. This X-linked Malformation and Infantile Lethality Syndrome has provisionally been named Ogden Syndrome, in honor of the hometown where the first family resides. This is the first human disease involving a defect in the N-terminal acetylation of proteins, a common (yet vastly understudied) modifications of eukaryotic proteins carried out by N-terminal

acetyltransferases (NATs). We are currently calling this new disease Ogden Syndrome, in honor of where the first family resides. There is significantly impaired biochemical activity of the mutant hNaa10p, suggesting that a reduction in acetylation of some unidentified proteins by hNaa10p might lead to this disease. There is currently very limited knowledge on the functional importance of Nt-acetylation at the protein level and at the organismal level. We have begun to study these processes in mammalian cell culture and yeast, along with setting up collaborations involving mice, zebrafish, and C. elegans.

**The Characterization and Analysis of an Idiopathic Intellectual Disability Syndrome via Whole Genome Sequencing Analysis**

with Jason O'Rawe, Yiyang Wu, Alan Rope and Jeffrey Swensen (University of Utah),

We have delineated a new idiopathic syndrome with intellectual disability and distinctive facial dysmorphology. The propositi are two affected male brothers, aged 10 and 12 respectively, with severe intellectual disability, autism-like behavior, attention deficit issues, and very distinctive facial features, including broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, relative hypertelorism, high-arched palate, and prominent ears. Their parents are nonconsanguineous and are both healthy, and the family history does not demonstrate any other members with anything resembling this current syndrome. X-chromosome inactivation assays reveal skewing in the mother, suggesting the possibility of an X-linked disorder. High-density genotyping arrays in the mother, father and two sons have not revealed previously known copy number variants (CNVs) that might contribute to the phenotype. Whole genome sequencing has led to the identification of several rare variants that are currently being characterized further, including in other unaffected members of the extended family.

**Expansion of collection efforts in Utah**

with Reid Robison and Clark Johnson (Utah Foundation for Biomedical Research), Kai Wang (University of Southern California, USC)

I have worked over the past year to collect dozens of additional families in Utah with severe neuropsychiatric disorders, as part of an Institutional Review Board (IRB)-approved protocol at the Utah Foundation for Biomedical Research (UFBR). We are collecting blood for DNA, RNA and peripheral blood mononuclear cells, with these cells to be used in future experiments to make and characterize induce pluripotent stem cells. For now, we are obtaining genotyping on the genomic DNA, and we have recently initiated whole genome sequencing of ~50 samples to a depth of >30x on the Illumina Hiseq platform, in an effort to understand the genetic basis of these disorders. The illnesses we are studying include intellectual disability, autism, schizophrenia, and other childhood-onset neuropsychiatric disorders.

**Toward more accurate variant calling for "personal genomes"**

with Jason O'Rawe, Yiyang Wu, Kai Wang (USC)

To facilitate the clinical implementation of genomic medicine by next-generation sequencing, it will be critically important to obtain accurate and consistent variant calls on personal genomes. Multiple software tools for variant calling are available, but it is unclear how comparable these tools are or what their relative merits in real-world scenarios might be. We sequenced 15 exomes from four families using the Illumina HiSeq 2000 platform and Agilent SureSelect v.2 capture kit, with ~120X mean coverage. We analyzed the raw data using near-default parameters with 5 different alignment and variant calling pipelines (SOAP, BWA-GATK, BWA-SNVer, GNUMAP, and BWA-SAMTools). We additionally sequenced a single whole genome using the Complete Genomics (CG) sequencing and analysis pipeline, with 95% of the exome region being covered by 20 or more reads per base. Finally, we attempted to validate 919 SNVs and 841 indels, including similar fractions of GATK-only, SOAP-only, and shared calls, on the MiSeq platform by amplicon sequencing with ~5000X average coverage. SNV

concordance between five Illumina pipelines across all 15 exomes is 57.4%, while 0.5-5.1% variants were called as unique to each pipeline. Indel concordance is only 26.8% between three indel calling pipelines, even after left-normalizing and intervalizing genomic coordinates by 20 base pairs. 11% of CG variants that fall within targeted regions in exome sequencing were not called by any of the Illumina-based exome analysis pipelines. Based on targeted amplicon sequencing on the MiSeq platform, 97.1%, 60.2% and 99.1% of the GATK-only, SOAP-only and shared SNVs can be validated, but only 54.0%, 44.6% and 78.1% of the GATK-only, SOAP-only and shared indels can be validated. Additionally, our analysis of two families, one containing four individuals and the other containing seven, demonstrates additional accuracy gained in variant discovery by having access to genetic data from a multi-generational family. Our results suggest that more caution should be exercised in genomic medicine settings when analyzing individual genomes, and to interpret positive and negative findings with scrutiny, especially for indels. Utilizing a combination of variants derived from multiple and orthogonal variant calling pipelines is a feasible first option for reducing false positives.

**Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress**

with Kai Wang (USC)

The pace of exome and genome sequencing is accelerating, with the identification of many new disease-causing mutations in research settings, and it is likely that whole exome or genome sequencing could have a major impact in the clinical arena in the relatively near future. However, the human genomics community is currently facing several challenges, including phenotyping, sample collection, sequencing strategies, bioinformatics analysis, biological validation of variant function, clinical interpretation and validity of variant data, and delivery of genomic information to various constituents. In light of this, I worked with my collaborator Kai Wang on a comprehensive review of these challenges and summarizing the bottlenecks for the

clinical application of exome and genome sequencing. We also suggested ways for moving the field forward, including the need for clinical-grade sample collection, high-quality sequencing data acquisition, digitalized phenotyping, rigorous generation of variant calls, and comprehensive functional annotation of variants. Additionally, we suggested that a "networking of science" model that encourages much more collaboration and online sharing of medical history, genomic data and biological knowledge, including among research participants and consumers/patients, will help establish how certain mutations may contribute toward the development of certain phenotypes.

**A proposal for an Analytic-Interpretive split (or a so-called "distributive model") across both clinical and research genomics.**

with Jeremy Segal (New York Genome Center)

I have been working on a policy piece related to the field of clinical genomics. In brief, the United States federal government mandates that any laboratory performing tests on human specimens "for the purpose of providing information for the diagnosis, prevention, or treatment of any disease" must satisfy the conditions set forth in the Clinical Laboratory Improvement Amendments (CLIA) of 1988. Most laboratories in academic research settings do not have sufficient standards in place to qualify them for CLIA approval. At the time CLIA was enacted, the separation of the clinical and research worlds seemed a fairly straightforward proposition. But today, the issues we face from a regulatory and ethical standpoint around genomics stem from the simple question: what do we do when it becomes difficult to draw a clear line of distinction between these two types of laboratory practices, particularly when researchers are working directly with families? Families afflicted with rare genetic disorders now have a reasonable expectation of definitive and potentially actionable results on the order of days to months, and all such families regardless of diagnosis are candidates for a relatively standardized genomic (rather than disease-specific mechanistic) analysis. The situation is

similar for with cancer, as standardized tumor-agnostic genomic analyses have a high likelihood of uncovering plausible drug targets during the lifetimes of some people, even those with late-stage disease. Some in the field might argue for the continued suitability of the "research first, clinical follow-up" model, with Sanger sequencing confirmation of "medically actionable" variants. But looking forward, as we learn more and more from the genome, and depending upon what types of data individuals want to receive from their genomes, the number of variants requiring confirmation will only increase, perhaps substantially. At $300 per variant (which is Sanger-based CLIA gene testing and unlikely to change anytime soon), the research model may rapidly become prohibitively expensive when used to manage care. Clearly history has shown that the quality of unregulated diagnostics is susceptible to perverse market forces, as many prior poor practices can be tied directly to economic conflicts of interest. At a minimum, CLIA erects standards to protect laboratories from these forces and provides an enforcement structure. The field would be prudent to embrace this protection, in light of the ongoing commoditization of sequencing and its associated potential for severe price competition.

As a solution, we are proposing an analytic-interpretive split (or a so-called "distributive model") across clinical genomics. This split model simply means that one laboratory performs sample processing/sequencing and then downstream laboratories can perform the interpretation and reporting. Thus, together, the laboratories perform all the functions that make up a laboratory test. The practical effect of this split would be to turn an exome or genome sequence into a discrete deliverable clinical unit that could be used for multiple purposes by downstream labs. The raw data, if individually approved, could be uploaded to electronic health records or others online sources, allowing its use in downstream investigative analysis. Another overlooked benefit of performing more standardized genome sequencing would be its effect on resultant datasets, and the community's resultant confidence therein. Enacting the type of split described herein allows organizations to leverage their strengths in sequencing or bioinformatics, thus lowering costs across the entire process. It will also help to provide guidance to the many

rapidly proliferating laboratories devoted to clinical genomic data mining. However, it is essential that we think of these companies as laboratories, even if they may be separate from the wet lab. Their post-analytic processing would simply be an extension of a prior wet-lab process, and thus included as a core component of a medical test.

**Publications**

**GJ Lyon*** and K. Wang*, Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress, Genome Med. 2012 Jul 26;4(7):58. *Co-Corresponding author.

**Lyon, G.J.*** Personalized medicine: Bring clinical standards to human-genetics research. Nature. 2012 Feb 15;482(7385):300-1, *Corresponding author.

Genome-wide association study of Tourette Syndrome, Scharf JM, Yu D, Mathews CA, Neale BM, Stewart SE, Fagerness JA, Evans P, Gamazon E, Edlund CK, Service SK, Tikhomirov A, 48 other authors, **Lyon GJ**, …18 other authors; for the North American Brain Expression Consortium, Hardy J; for the UK Human Brain Expression Database, 14 other authors, Oostra BA, McMahon WM, Freimer NB, Cox NJ, Pauls DL. Mol Psychiatry. 2012 Aug 14. doi: 10.1038/mp.2012.69.