Human Genetics and Genomic Medicine

Gholson Lyon's lab focuses on analyzing human genetic variation and its role in severe neuropsychiatric disorders and rare diseases, including intellectual disability, autism, and schizophrenia. By recruiting large groups of related individuals living in the same geographic location (e.g., Utah), Lyon's lab can study the breadth and depth of genetic variants in a similar environmental background. Using the exome—the parts of the genome that code for protein—and whole-genome sequencing, the lab looks for mutations that segregate with syndromes in the various populations, and the lab undertakes comprehensive functional studies of many of the newly identified mutations.

Ogden Syndrome and the amino-terminal acetylation of proteins, with Max Doerfel, Yiyang Wu, Ronen Marmorstein (Philadelphia, PA), Thomas Arnesen (Norway), Nathalie Reuter (Norway), Petra van Damme (Belgium)

More than 85 % of human proteins are acetylated at their N-terminal amino group, hence, N-terminal acetylation (NTA) is one of the most abundant modifications of eukaryotic proteins. Despite its discovery more than 30 years ago, very little is known about the cellular effects/functions of this modification. In humans, 6 distinct N-terminal amino-acetyltransferases (NATs) catalyze the transfer of an acetyl group from acetyl-CoA to the Na-terminal amino group of their specific target proteins. The major human acetyltransferase, NatA, consists of an auxiliary subunit, Naa15, and a catalytic active subunit, Naa10. We have previously described two families with a lethal X-linked disorder of infancy called Ogden syndrome. This disorder comprises a distinct combination of an aged appearance, craniofacial anomalies, hypotonia, global developmental delays, cryptorchidism and cardiac arrhythmias. Using X chromosome exon sequencing, we identified a c.109T>C (p.Ser37Pro) variant in Naa10 as contributing to this disease. Biochemical analysis and immunoprecipitation assays in combination with LCMS demonstrated a reduced catalytic capacity and revealed an impaired binding of the S37P mutant towards specific interaction partners, including Naa15 and Naa50. Analysis of the N-terminal acetylome of patient cells revealed a decreased acetylation of a subset of NatA substrates, indicating that a reduced binding capability and an affected enzymatic activity of the Naa10 S37P mutation is a prominent feature in Ogden Syndrome. Characterization of *NAA10*/*NAA15* knockout yeast strains revealed various phenotypes, including growth defects at elevated temperatures and altered sensitivity towards cytotoxic stresses. These effects could be rescued by overexpressing human wild type Naa15/Naa10 from plasmids; however, overexpressing mutant Naa15/Naa10 S37P only partially rescue these effects. Interestingly, introduction of both human Naa15/Naa10 wt and S37P mutant into the endogenous locus of the corresponding yeast genes failed to reverse the effects. We also continued our efforts with establishing induced pluripotent stem cells (iPSCs) from skin fibroblasts from one of the boys with Ogden Syndrome, and we are also establishing knockin mice containing the mutation of interest in NAA10. Ongoing work will focus on characterizing the cells and mice.

RykDax Syndrome: Characterization and Analysis of an Idiopathic Intellectual Disability Syndrome, with Jason O'Rawe, Yiyang Wu, Han Fang, Laura Jimenez Barron, Ed Yang (Boston), Alan Rope (Oregon) and Jeffrey Swensen (Arizona), Reid Robison (Utah), Kai Wang (California)

We worked on the discovery of a new genetic syndrome, RykDax syndrome, driven by a whole genome sequencing study for one family from Utah with two affected male brothers, presenting with severe intellectual disability (ID), a characteristic intergluteal crease, and very distinctive facial features, including a broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, prominent periorbital ridges, deep-set eyes, relative hypertelorism, thin upper lip, a high-arched palate, prominent ears with thickened helices, and a pointed chin. This Caucasian family was recruited from Utah, USA, and Illumina-based WGS was performed on 10 members of this family, with additional Complete Genomics-based WGS performed on the nuclear portion of the family (mother, father and the two affected males). Using WGS datasets from 10 members of this family, we could increase the reliability of the biological inferences with an integrative bioinformatic pipeline. In combination with insights from clinical evaluations and medical diagnostic analyses, these DNA sequencing data were used in the study of three plausible genetic disease models that might uncover genetic contribution to the syndrome. We found a 2 to 5-fold difference in the number of variants detected as being relevant for various disease models when using different sets of sequencing data and analysis pipelines. We derived greater accuracy when more pipelines were used in conjunction with data encompassing a larger portion of the family, with the number of putative *de-novo* mutations being reduced by 80%, due to false negative calls in the parents. The boys carry a maternally inherited missense variant in a X-chromosomal gene *TAF1*, which we

consider as disease relevant. We took a "genotype-first" approach to find other families with variants in *TAF1* and containing individuals having a remarkably similar clinical presentation. TAF1 is the largest subunit of the general transcription factor IID (TFIID) multi-protein transcription complex, and our results have implicated mutations in TAF1 as playing a critical role in the development of this new intellectual disability syndrome. We published the mutation as part of a paper developing SeqHBase, described below, and we also posted a preprint with much more detailed clinical description to the BioRxiv preprint server, while we continue searching for mutations in TAF1 in additional families.

Development of comprehensive whole genome sequencing analysis pipelines, with Han Fang, Jason O'Rawe, Laura Jimenez Barron, Yiyang Wu, Michael Schatz, Giuseppe Narzisi, Kai Wang (California), Max He (Wisconsin).

We continued developing various bioinformatics approaches for the analysis of exome and whole genome sequencing data. For example, in one project, we showed that the accuracy of detection of small insertions and deletions (indels) is greater when using whole genome sequencing versus exon capture and sequencing. We also calculated that 60X WGS depth of coverage from the Illumina HiSeq platform is needed to recover 95% of indels detected by Scalpel. While this is higher than current sequencing practice, we proposed that the deeper coverage may save total project costs because of the greater accuracy and sensitivity. Finally, we investigated sources of INDEL errors (e.g., capture deficiency, PCR amplification, homopolymers). We reported over the past 12 months the results of several other ongoing bioinformatics projects as well, as shown in the below publications. For example, we developed SeqHBase, a big data-based toolset for analysing family-based sequencing data to detect *de novo*, inherited homozygous or compound heterozygous mutations that may be disease contributory. We demonstrated SeqHBase's high efficiency and scalability, which is necessary, as WGS and WES are rapidly becoming standard methods to study the genetics of familial disorders. We also recently published an opinion piece regarding the current state of uncertainty quantification in DNA sequencing applications, and we proposed methods that can be used for accounting and propagating these errors and their uncertainties through subsequent calculations.

Summarizing the state of human genetics, including the genetic architecture of human disease, with Jason O'Rawe.

We prepared a comprehensive book chapter summarizing the current state of human genetics, and we continue to expand upon this work. In brief, there are ~12 billion nucleotides in every cell of the human body, and there are ~25-100 trillion cells in each human body. Given somatic mosaicism, epigenetic changes and environmental differences, no two human beings are the same, particularly as there are only ~7 billion people on the planet. One of the next great challenges for studying human genetics will be to acknowledge and embrace complexity. Every human *is* unique, and the study of human disease phenotypes (and phenotypes in general) will be greatly enriched by moving from a deterministic to a more stochastic/probabilistic model. The dichotomous distinction between 'simple' and 'complex' diseases is completely artificial, and we argue instead for a model that considers a spectrum of diseases that are variably manifesting in each person. The rapid adoption of whole genome sequencing (WGS) and the Internet-mediated networking of people promise to yield more insight into this century-old debate. Comprehensive ancestry tracking and detailed family history data, when combined with WGS or at least cascade-carrier screening, might eventually facilitate a degree of genetic prediction for some diseases in the context of their familial and ancestral etiologies. However, it is important to remain humble, as our current state of knowledge is not yet sufficient, and in principle, any number of nucleotides in the genome, if mutated or modified in a certain way and at a certain time and place, might influence some phenotype during embryogenesis or postnatal life.

Expanding collection and sequencing of other rare genetic syndromes, with Jason O'Rawe, Yiyang Wu, Han Fang, Laura Jimenez Barron, Margaret Yoon, Ivan Iossifov, Reid Robison (Utah), Kai Wang (California) Alan Rope (Oregon), and others.

We continue to meet and collect many families in Utah and elsewhere with very rare, idiopathic genetic syndromes. The total number of DNA samples collected to date is approaching 2000, and this includes detailed phenotyping information. Some of these samples have undergone exome or whole genome sequencing, and we are currently analyzing 18 whole genomes and 35 exomes generated as part of this project. This includes the ongoing analysis of whole genomes from 3 families with singleton cases of autism, being performed in collaboration with Ivan Iossifov at CSHL, and an analysis of nine whole genomes from a pedigree with Prader–Willi Syndrome (PWS), Hereditary Hemochromatosis, Familial Dysautonomia (FD), and Tourette Syndrome. We did publish about one case in China where two siblings both began to develop

idiopathic progressive cognitive decline starting from age six, and were suspected to have an undiagnosed neurological disease. Initial clinical assessments included review of medical history, comprehensive physical examination, genetic testing for metabolic diseases, blood tests and brain imaging. We performed exome sequencing with Agilent SureSelect exon capture and Illumina HiSeq2000 platform, followed by variant annotation and selection of rare, shared mutations that fit a recessive model of inheritance. To assess functional impacts of candidate variants, we performed extensive biochemical tests in blood and urine, and examined their possible roles by protein structure modeling. Exome sequencing identified NAGLU as the most likely candidate gene with compound heterozygous mutations (chr17:40695717C>T and chr17:40693129A>G in hg19 coordinate). Sanger sequencing confirmed the recessive patterns of inheritance, leading to a genetic diagnosis of Sanfilippo syndrome (mucopolysaccharidosis IIIB). Biochemical tests confirmed the complete loss of activity of alpha-N-acetylglucosaminidase (encoded by NAGLU) in blood, as well as significantly elevated dermatan sulfate and heparan sulfate in urine. Structure modeling revealed the mechanism on how the two variants affect protein structural stability. This successful diagnosis of a rare genetic disorder with an atypical phenotypic presentation confirmed that such "genotype-first" approaches can particularly succeed in areas of the world with insufficient medical genetics expertise and with cost-prohibitive in-depth phenotyping

Writing policy and opinion pieces in genomic medicine, with Carol Barash and others.

The PI is an ongoing advocate for open access and data sharing, along with assisting with efforts at CSHL Press to roll out new initiatives, such as the BioRxiv preprint server and the new journal, Molecular Case Studies, which is designed to make it easier to publish and share the underlying data from single cases, families or small cohorts of human conditions.

Collaborating on genetics of Tourette Syndrome, with the Tourette Syndrome Association International Consortium for Genetics.

The PI continues to collaborate on this international effort to understand the genetics of Tourette Syndrome. Psychiatric comorbidity is common in Tourette syndrome (TS); when present, these conditions typically cause more distress and impairment than do tics. High rates of attention-deficit/hyperactivity disorder (ADHD) and obsessive-compulsive disorder (OCD) are well documented and thought to be core components of the TS phenotype; however, few studies have fully characterized other comorbidities. We therefore characterized the prevalence and impact of psychiatric comorbidity in a large sample of individuals with TS and their family members. The lifetime prevalence of comorbid psychiatric disorders, their heritabilities, and ages of risk were determined in participants with TS (N=1,374). The lifetime prevalence of any psychiatric comorbidity was 85.7%; 57.7% of participants had ≥2 psychiatric disorders. The mean number of lifetime comorbid diagnoses was 2.1. Prevalence of mood, anxiety, and disruptive behavior disorders was ~30%. The age of greatest risk for most psychiatric disorders was 4 to 10 years, except eating and substance use disorders, which begin in adolescence. TS was associated with increased risk of anxiety and decreased risk of substance use disorders; high rates of mood disorders among participants with TS may be accounted for by comorbid OCD. Parental history of ADHD was associated with a higher burden of non-OCD, non-ADHD psychiatric disorders. Genetic correlations between TS and mood, anxiety, and disruptive behavior disorders were better accounted for by ADHD and, for mood disorders, by OCD. This study confirmed that psychiatric comorbidities begin early in life and are extremely common among individuals with TS, and demonstrates that some may be mediated by the presence of comorbid OCD or ADHD.

References

1. Myklebust LM, Van Damme P, Støve SI, Dörfel M, Kalvik TV, Abboud A, Grauffel C, , Jonckheere V, Kaasa H, Liszczak G, Marmorstein R, Reuter N, **Lyon GJ**, Gevaert K, Arnesen T. (2014). Biochemical and cellular analysis of Ogden syndrome reveals downstream Nt-acetylation defects. Hum Mol Genet, pii: ddu611. PMID: 25489052.
2. **Lyon GJ**, Bird, L., Rope, A. "X-linked Malformations and Infantile Lethality". In: Zimmerman C, editor. Epstein's Inborn Errors of Development: Oxford University Press; 2015. (in press).
3. O'Rawe J, Rope A, Wu Y, Swensen J, Robison R, Wang K, **Lyon GJ**. (2015). A variant in TAF1 is associated with a new syndrome with severe intellectual disability and characteristic dysmorphic features. *BioRxiv preprint.* http://dx.doi.org/10.1101/014050

4. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, Wang Z, Wu Y, **Lyon GJ**, Wigler MH, Schatz MC. (2014). Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nature Methods, doi:10.1038/nmeth.3069. PMCID: PMC4180789

5. Fang H, Wu Y, Narzisi G, O'Rawe JA, Jimenez Barron LT, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC, **Lyon GJ**. (2014). Reducing INDEL calling errors in whole-genome and exome sequencing. Genome Medicine, 6:(89). doi:10.1186/s13073-014-0089-z. PMCID: PMC4240813.

6. Brownstein C, Beggs A, **Lyon GJ**, Margulies D, Kohane I, and 195 other co-authors. (2014). An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biology, 15(3): R53. PMCID: PMC4073084.

7. O'Rawe JA, Ferson S, **Lyon GJ**. (2015). Accounting for uncertainty in DNA sequencing data. Trends in Genetics. pii: S0168-9525 (14)00209-1. doi: 10.1016/j.tig.2014.12.002. PMID: 25579994.

8. He M, Person TN, Hebbring SJ, Heinzen E, Ye Z, Schrodi SJ, McPherson EW, Lin SM, Peissig PL, Brilliant MH, O'Rawe J, Robison RJ, **Lyon GJ**, Wang K. (2015). SeqHBase: a big data toolset for family-based sequencing data analysis. Journal of Medical Genetics. pii: jmedgenet-2014-102907. doi: 10.1136/jmedgenet-2014-102907. PMID: 25587064.

9. **Lyon GJ**, O'Rawe, J. "Clinical Genetics of Neurodevelopmental Disorders". In: Mitchell K, editor. The Genetics of Neurodevelopmental Disorders. (in press). Preprint also posted on BioRxiv preprint server: doi: 10.1101/000687: Wiley; 2015.

10. Shi L, Li B, Huang Y, Liu T, **Lyon GJ**, Xu A, Wang K. (2014). "Genotype-first" approaches on a curious case of idiopathic progressive cognitive decline. BMC Med Genomics, 7(1): 66. PMCID: PMC4267425.

11. Barash CI and **Lyon GJ**. (2014). Open access and data sharing: Easier said than done. Applied & Translational Genomics. doi: http://dx.doi.org/10.1016/j.atg.2014.09.008.

12. McGrath LM, Yu D, Marshall C, Davis LK, Thiruvahindrapuram B, Li B, Cappi C, Gerber G, Wolf A, Schroeder FA, Osiecki L, Lowe TL, **Lyon GJ**, Macciardi F, Maier W, … 57 other authors. (2014). Copy Number Variation in Obsessive-Compulsive Disorder and Tourette Syndrome: A Cross-Disorder Study. Journal of American Academy of Child & Adolescent Psychiatry, 53(8):910-9. PMCID: PMC4218748.

13. Hirschtritt ME, Lee PC, Pauls DL, Dion Y, Grados MA, Illmann C, King RA, Sandor P, McMahon WM, **Lyon GJ**, Cath DC, Kurlan R, Robertson MM, Osiecki L, Scharf JM, Mathews CA[#]. (2015). Lifetime Prevalence, Age of Risk, and Etiology of Comorbid Psychiatric Disorders in Tourette Syndrome. Tourette Syndrome Association International Consortium for Genetics. *(Manuscript accepted at JAMA Psychiatry).*