

CUT THE HYPE. ACCURACY AND STANDARDS COME FIRST

Gholson Lyon, Assistant Professor, Cold Spring Harbor Laboratory

WITH INCREASED MEDIA AND POLITICAL ATTENTION, THE GENOMICS HYPE MACHINE IS IN FULL SWING. TO REALISE THE PROMISE OF PRECISION MEDICINE, THERE IS STILL A LONG WAY TO GO. GHOLSON LYON TALKS US THROUGH SOME OF THE KEY AREAS THAT NEED TO BE ADDRESSED, AND THE BENEFITS OF BROADENING YOUR EXPERIENCE.

Genomics is growing in just about every way imaginable. Advances in sequencing technology and cloud computing are making genomic analysis more accessible to researchers; more and more genomic data is being produced; and the political backing is also bringing more funding. It is a very exciting time to be involved in genomics at any level, but we can't get too excited just yet.

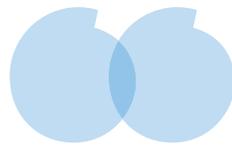
Managing stakeholder expectations is a constant battle. Especially when it comes to ensuring that robust and accurate science remain a priority over short term goals. Hype is great, in that it shows people are excited, but there's still a long way to go.

FLG: The political arena on both sides of the Atlantic has been very vocal about their support and backing of Genomics and Precision Medicine in recent weeks. Is there a risk in overhyping the public?

GL: Yes, of course. There seems to be an enormous amount of hype in the field of human genomics, and we must constantly remind people that the path to new drugs and/or the prevention of disease can take on the order of decades.

FLG: There's a sense of a growing disconnect between the growing political element and those doing the actual research. Is that funding being focused in the right areas of genomic research at the moment?

GL: More funding should be directed toward developing more accurate and faster sequencing methods, along with engaging much more with software engineers and cloud-based computing. We have to collectively develop ways to store and share millions of genomes going forward, and this is what various scientists and



“THIS \$1000 DOLLAR GENOME IS STILL A MYTH, AS YOU ARE QUITE CORRECT THAT SUCH A THING IS NOT ACCURATE”

members of the Global Alliance for Genomics and Health are working on. From my perspective, the emphasis should be on further technology development and the implementation of highly accurate genome sequencing. This is part of what we have been working on in collaboration with Michael Schatz and others at Cold Spring Harbor Laboratory and elsewhere. Using whole genome datasets from 10 members of one family, my graduate students, Jason O'Rawe, Han Fang, and Yiyang Wu, showed that one can increase the reliability of the biological inferences with an integrative bioinformatics pipeline, including a new algorithm, Scalpel, developed by Giuseppe Narzisi and Michael Schatz, for more accurate identification of indels. We find 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, and we derive greater accuracy when more pipelines are used in conjunction with data encompassing a larger portion of the family. We also collaborated with Min (Max) He and Kai Wang on the development of SeqHBase, a big data-

based toolset for analysing family-based sequencing data, and we demonstrated SeqHBase's high efficiency and scalability on several disorders, including a new syndrome, which we are currently calling RykDax Syndrome, where we identified a maternally inherited missense variant in an X-chromosomal gene, TAF1. A “genotype-first” approach led us to other families with variants in TAF1 and containing individuals having a remarkably similar clinical presentation.

FLG: Genomic technology has, and continues to develop at a rapid pace. At the center of this was the race to the ‘\$1000 dollar genome’. In pursuing faster and cheaper sequencing options, accuracy suffered. It has resulted in a drastic growth in the rate at which sequencing data is being produced. Was the compromise towards speed and affordability worth it? →

GL: This \$1000 dollar genome is still a myth, as you are quite correct that such a thing is not accurate. We published papers in 2013 and 2014 showing that the accuracy of “whole genome sequencing” is far from ideal, and we have recommended that 60x coverage is needed to detect >95% of indels, if one uses 100 base pair reads from Illumina. This mythical \$1000 dollar genome is only at 30x coverage and this is the amortized price on the current Illumina X10 instrument, which assumes running the instrument constantly to churn out 18,000 such genomes per year. I have heard of some genome centers charging \$1500 for this 30x coverage, although the real price is closer to \$2,000, I am told, and this does not include any of the costs for analysis. Therefore, for a 60x coverage Illumina genome, the current cost is \$3,000, and even this sort of genome still lacks many regions that can only be filled in with much longer reads. There is certainly ongoing improvement and innovation, including from other sequencing companies, so that hopefully in a few years’ time, the accuracy will be much improved. However, there is still plenty of work to be done.

FLG: Are people sequencing too much at the moment? It seems that every week we see a newly completed genome announced. Is this an indication of the potential power of genomics or are we just in a period in which sequencing technology is being used simply because it has become more accessible?

GL: About four years ago, I started broadly calling for better standards in terms of exome sequencing. However, the response basically was that implementing better standards makes the cost of research too much. So, we find ourselves in the situation of having had tens of thousands of exomes sequenced in research environments, where the only variants that can be returned to the research participants are ones that get Sanger validation (or some other validation) in a clinical environment. This was mentioned

and discussed yet again at the Precision Medicine Workshop held recently by the NIH, with various people lamenting the fact that there is such a huge divide between the “research” and “clinical” worlds, but the response once again concerned mostly cost. So, from this perspective, the main work right now should focus on getting the cost of sequencing down much further, along with getting higher accuracy with longer reads, so that eventually people can get their whole genomes sequenced in clinical environments, where the chances for sample-swaps and other inaccuracies will be less. I do applaud that the FDA has finally approved a direct to consumer genetic test for Bloom Syndrome by 23andMe, but I certainly hope that the pace and scope of such approvals will dramatically increase. We need to get to a world of highly accurate and relatively cheap (~\$100 genomes), so that it is then cost-effective to sequence millions of people, and then collect and analyse these data in aggregate to begin to understand how any particular genotypes express themselves among many different genetic backgrounds. Such things will only be possible with broad data sharing, including on the level of phenotype data. One can see that this very broad sharing of data, including pictures, is possible, as demonstrated by innovative companies like Facebook, 23andMe, PatientsLikeMe, and Ancestry.com, although the privacy concerns and issues with genetic data are a big issue that must be carefully considered. There have also been some recent innovations in face recognition and image processing, where one can begin to classify genetic syndromes based on photographs.

FLG: In terms of technology coming through today, what do you feel is going to make the most useful impact?

GL: The developments going on right now at Pacific Biosciences and Oxford Nanopore are very promising in terms of longer reads, for sure. Researchers at Cold Spring Harbor Laboratory, including



COLD SPRING HARBOR LABORATORY

Cold Spring Harbor Laboratory is a world-renowned, private research and education institution with research programs in Cancer, Neuroscience, Plant Biology, Quantitative Biology, and Bioinformatics & Genomics. The ultimate goal is to apply this research on basic biological mechanisms to improve the diagnosis and treatment of cancer, neurological disorders and other diseases.

The Genomics Program at CSHL is comprised of faculty working across disciplines and research areas. Their main research interests are genomic organisation, structural variation of the human genome as related to disease, computational genomics and transcriptional modelling, and sequencing technology. Faculty in the Genomics program conduct research in the areas of human genetics, functional genomics and small RNA biology.

Dick McCombie and Michael Schatz, have been working hard on assessing and pushing forward these technologies from Pacific Biosciences and Oxford Nanopore, and it seems that we might only be a few years away from highly accurate and relatively inexpensive human whole genomes. It is also useful that people at the FDA seem to be beginning to engage more with how to regulate direct to consumer genetic testing in broader fashion. One can also see that companies like Google are getting interested in this sector, based on their cloud computing capabilities, so one could imagine that this could further enable broader data sharing.

FLG: The road to the promised 'Genomic Revolution' seems to keep stretching out. One of the first major milestones was the completion of the Human Genome Project. Back then, President Clinton suggested that "our children's children will know the term 'cancer' as a constellation of stars." The miracle cures that the public were hoping for, never arrived. It was, however, the starting point for a lot of great work. Now with large scale sequencing projects taking place, it feels like the expectation is that a whole bunch of variants are going to be found that we can drug and eradicate disease. The reality is that it is very unlikely that a single gene holds the answer. So the next step will be to take all of that data and turn it into useful information. Is there enough focus on funding the kinds of rigorous functional studies that will help deliver something tangible to patients?

GL: Although it seems that most people focus on the development of drugs to treat illness after it has started, I personally feel that large-scale sequencing might actually be more effective in terms of early detection and prevention of disease. This remains to be proven, of course, but there are early indications that screening of family members (known as cascade carrier screening) can help to identify other members of the family carrying particular mutations and thus at increased risk. Such people can at least know about their elevated risk, and there are some instances in which people can take action. Perhaps the most famous example of this is in women with prominent family histories of breast and ovarian cancer, who also carry mutations in the BRCA1 gene. Such women can undergo more intensive screenings, or in some cases, made famous by Angelina Jolie, elect to undergo interventions to reduce their risk, such as with mastectomies. Of course, the decision to undergo such a radical intervention particularly when the expression of this phenotype is quite variable is something that needs to undergo much more study, but on an individual level, some people are acting on their genetic information already.

FLG: You've had some experience on the front lines as a clinician yourself. Here in the UK, one of the biggest challenges the 100,000 Genomes Project is trying to address is how to integrate genomics into the NHS. This will have to look at work force planning, in particular around how to use and support Genetic Counsellors. How difficult can it be to introduce something as big as genomics into regular clinical practice?

GL: This is indeed an enormous challenge, particularly in the United States, where we do not have a national health care system. In addition, the number of genetic counsellors in America is on the order of only a few thousand, for a population numbering well over 300 million people. There is also a pressing need to educate health care professionals much more about genetics. I am constantly reminding people in the genomics world that we live in a tiny bubble in comparison to the vast landscape of healthcare in America, and



Gholson Lyon
Assistant Professor
Cold Spring Harbor Laboratory

Gholson Lyon is on faculty at Cold Spring Harbor Laboratory and is a research scientist at the Utah Foundation for Biomedical Research. He is also a board-certified child, adolescent and adult psychiatrist. He earned an M.Phil. in Genetics at the University of Cambridge, then received a Ph.D. and M.D. through the combined Cornell/Sloan-Kettering/Rockefeller University training program. He started his independent research career in 2009, after finishing clinical residencies in child, adolescent and adult psychiatry. In addition to his research on the genetics of neuropsychiatric illnesses, Gholson is focusing on the genetic basis of rare Mendelian diseases and the development of clinical-grade exome and whole genome sequencing.

"ABOUT FOUR YEARS AGO, I STARTED BROADLY CALLING FOR BETTER STANDARDS IN TERMS OF EXOME SEQUENCING. HOWEVER, THE RESPONSE BASICALLY WAS THAT IMPLEMENTING BETTER STANDARDS MAKES THE COST OF RESEARCH TOO MUCH"

we have to prove clinical validity and utility in order to get any sort of wider adoption of exome or whole genome sequencing. Such things will be enabled by better technology, higher standards, lower costs, and broader data sharing.

FLG: Your research is focused on the genetics of neuropsychiatric illnesses at Cold Spring Harbor. What drew you to that particular field?

GL: The short answer is that these illnesses are among the most fascinating in all of medicine. The long answer concerns the broad training that I have had, which requires me to explain my background a bit. I conducted some of my training in the Cornell/Rockefeller/Sloan-Kettering M.D./Ph.D. program. Just prior to that, I had spent one year as a Rotary Scholar at the University of Cambridge, England, working toward a Master's degree in Genetics at the Wellcome CRC Institute, with Martin Evans as my research mentor. By the time I finished my Ph.D. at Rockefeller with Tom Muir and Richard Novick and returned to medical school, I had been exposed to a variety of research experiences. I had conducted research in various laboratories at Dartmouth College, the NIH, the University of Cambridge and the Cornell/Rockefeller/Sloan-Kettering M.D./Ph.D. program. Some of my research interests had included by then: thyroid hormone and its effects on the brain during development; cancer research with a focus on chemoprevention; the creation and phenotypic characterization of mouse models of human disease; the structure and function of proteins, including the use of chemistry to synthesize unusual protein variants and to analyze complex mixtures; and the development of novel anti-infectives for *Staphylococcus aureus* and other bacterial infections. It is safe to say that I had, and continue to have, wide-ranging interests in biological chemistry, which broadly defined encompasses the targeted use of chemistry to elucidate biological processes and vice-versa. →

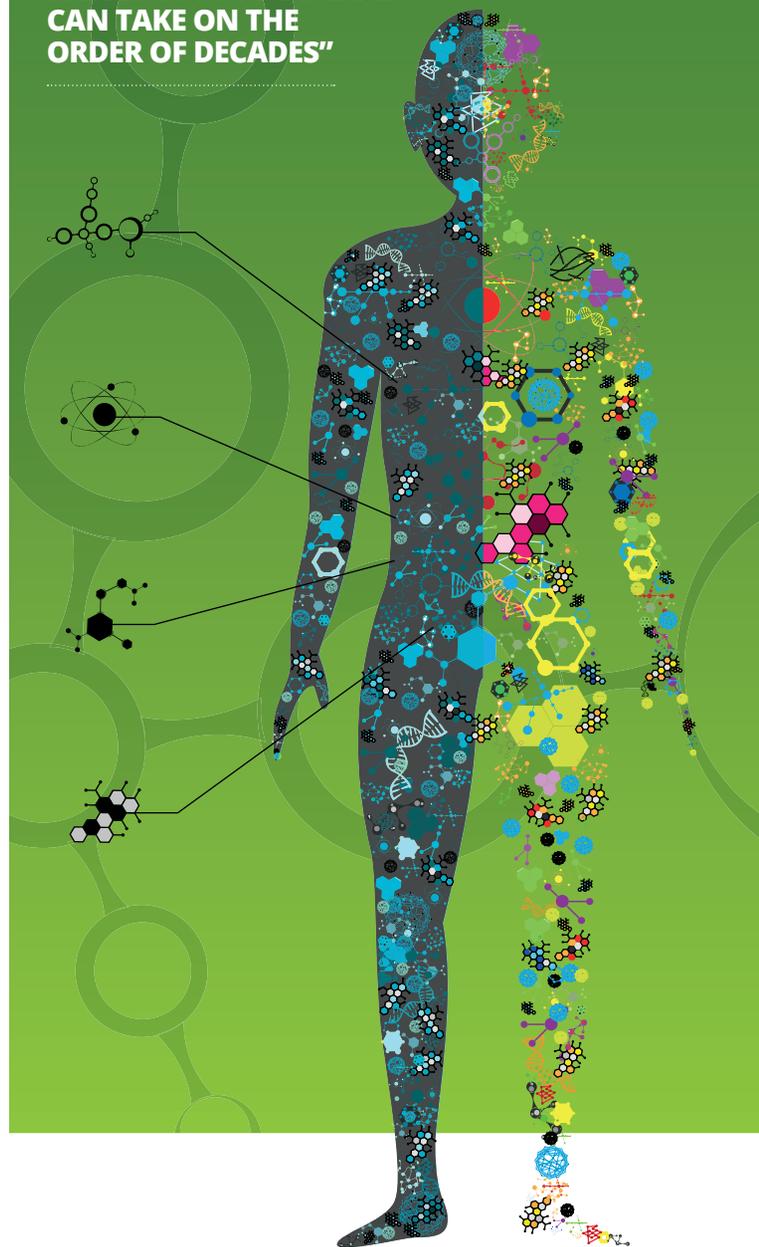
Upon returning to medical school and rotating through the many different clinical rotations, I was struck by the fact that the most fascinating and challenging clinical cases for me involved illnesses affecting the brain and mind, due mostly to our relative lack of knowledge of the complexities of psychiatric and neurologic illnesses. I realized that much remains to be discovered and that one could fill an entire career focusing on the clinical and basic science aspects of diseases like intellectual disability, autism, schizophrenia, obsessive-compulsive disorder, Tourette syndrome, or other brain-based illnesses. In order to make substantive contributions in this area, I felt that I needed extensive clinical exposure to these illnesses, so I decided in the year 2003 to pursue clinical residency training, first as a psychiatry resident at Columbia and the New York State Psychiatric Institute, followed by additional training in child and adolescent psychiatry at New York University, Bellevue Hospital and Rockland Children's State Psychiatric Hospital. Upon completion of 5 years of clinical training, I started my independent career in 2009 in the state of Utah, and I was recruited in 2012 to establish an active laboratory at Cold Spring Harbor. We focus on the discovery of families with rare diseases and/or increased prevalence for syndromes such as intellectual disability, autism, Tourette syndrome and schizophrenia. Once we identify mutations that likely contribute to a disease, we undertake detailed functional studies of these mutations and the biological processes affected. Proving the biological relevance for newly discovered mutations is the major problem, so having access to research participants and derived tissues is critically important, hence the need to engage directly with families.

FLG: You split your time as a researcher at the Utah Foundation for Biomedical Research (UFBR) as well. What kind of projects are you working on at the moment?

GL: We study the breadth and depth of genetic variants in Utah, where there is a large founding population, large family structures and good genealogical records, which enables well powered family-based genetic studies for rare diseases. We use exome and whole genome sequencing (WGS) to identify mutations that segregate with various idiopathic syndromes, and we undertake comprehensive functional studies of many of the newly identified mutations. This has led to the discovery of many new genetic syndromes, including Ogden Syndrome, RBCK1 Syndrome, and most recently RykDax Syndrome. This latter syndrome presents with severe intellectual disability (ID), a characteristic intergluteal crease, and very distinctive facial features. We continue to advocate for more comprehensive and accurate whole genome analyses in large pedigrees, and we have collected ~2000 DNA samples to date from >100 families in Utah, including detailed phenotyping information. Some of these samples have undergone exome or whole genome sequencing, and we are currently analyzing these data. This includes the ongoing analysis of whole genomes from 3 families with singleton cases of autism, and an analysis of nine whole genomes from a pedigree with Prader-Willi Syndrome (PWS), Hereditary Hemochromatosis, Familial Dysautonomia (FD), and Tourette Syndrome.

FLG: We interviewed Michael Vellard, from Ultragenyx last December. He had family reasons for wanting to develop treatments for rare diseases, and was very passionate about the potential relevance of his research to other, more common, indications. What attracted you to start researching into rare diseases?

"THERE SEEMS TO BE AN ENORMOUS AMOUNT OF HYPE IN THE FIELD OF HUMAN GENOMICS, AND WE MUST CONSTANTLY REMIND PEOPLE THAT THE PATH TO NEW DRUGS AND/OR THE PREVENTION OF DISEASE CAN TAKE ON THE ORDER OF DECADES"



GL: When I finished my M.D.-Ph.D. training in 2004, I was full of ideas about what research I might do in my future academic career. However, I decided to broaden my training by undertaking the clinical residency. Now, many years later, I am so glad that I undertook clinical training, as my eyes are now fully open to the complexity and nuance of the human condition, which cannot possibly be understood fully by studying the outcome of mutations only in mice or other lower animal models. Due mostly to my clinical training, my outlook has broadened to include a focus on rare diseases with very strong phenotypes, as these provide a window into very interesting and important biology. I believe that my background in genetics, chemistry, pharmacology, and medicine allows me to interface with basic

scientists and clinicians in the discovery and characterization of new genetic syndromes. It is critical for suitably trained physician-scientists or other broadly trained individuals to be involved with careful phenotyping and collection of human pedigrees with particular disorders, followed by a well-thought out experimental design in terms of whole genome sequencing and follow-up experiments.

FLG: Going back to the NIH Precision Medicine Workshop, do you think more people should try get a broad experience across research and clinical practice to try eliminate the present divide between the two?

GL: Yes, definitely. There is a major dearth of physician-scientists and other broadly trained scientists right now in America, and there is also a hyper-specialization that has occurred partly due to the way that NIH funding is determined, including evaluation of research grants by hyper-specialized study sections. I have been constantly amazed that there is very little reward in the current system for people with broad, interdisciplinary training, and in fact, I have received grant evaluations somehow lamenting the fact that I am not “focused enough” on one particular topic. I have heard that the NIH is trying to figure out ways to support broadly trained individuals, and I would certainly support such efforts. We definitely need to figure out ways to bridge this substantial divide between research and clinical practice, and this can only be done with the aid of people trained in both areas. I have been incredibly lucky at Cold Spring Harbor Laboratory that the president, Bruce Stillman, and the chancellor emeritus, Jim Watson, have been so supportive of my work.

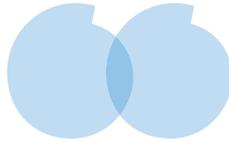
FLG: As genomics develops as a field over the next 15-20 years, what would you like to have achieved through your work?

GL: I have been advocating for more rigorous standards regarding the collection of human genetic data, including with the accuracy of variant calls. Instead of compartmentalizing research and medicine, the two should be integrated so that physicians who are most familiar with human “phenotypes,” can inform the other arms of science. This is certainly beginning to happen, and I am very much helping to push this forward.

FLG: For students at the start of a potential career in genomic research, would you have any advice or recommendations about how they should look to develop their areas of focus?

GL: Get a broad training! For me, science is all about trying to understand biology using whatever tools I can bring to the table. Therefore, genomics is just one of many tools out there. Over the course of my career, I have used many other tools, including cell culture, genome-edited mice, peptide chemistry, protein expression, mass spectrometry, and yeast genetics. So, I tend to pick biological questions and then figure out the best tools and techniques that I might need to answer the questions.

FLG: Since you joined Twitter in 2011, you have been steadily developing an impressive following. It seems that twitter is the



“THE MAIN WORK RIGHT NOW SHOULD FOCUS ON GETTING THE COST OF SEQUENCING DOWN MUCH FURTHER, ALONG WITH GETTING HIGHER ACCURACY WITH LONGER READS, SO THAT EVENTUALLY PEOPLE CAN GET THEIR WHOLE GENOMES SEQUENCED IN CLINICAL ENVIRONMENTS”

social media platform of choice for a lot of researchers at the moment. As a frequent tweeter, you give a lot to the community. What’s your main motivation for maintaining your online presence? Have you had anything out of the ordinary happen to you online yet?

GL: I use Twitter as a way to communicate with other scientists and the general public. It is a great way to keep up to date on what is happening in science in general, along with also seeing how the blogosphere reacts to overly hyped papers. Most people do not want to take the time to criticize papers, either at all or in various snail-mail venues such as “letters to the editor”, whereas a few scientists are at least willing to send out a tweet or post something on a blog to call out various papers. I do think that such analyses help to alert people regarding recidivist behaviours on the part of some scientists who tend to overly hype their results. This is particularly prominent for some people who tend to issue overly dramatic

(and sometimes misleading) press releases about their work. These things take time, but it is my hope that the younger generation of scientists will learn from such things on Twitter that it is actually damaging to your career and reputation to engage in so much hype and spin. People may or may not directly call out such behaviour publicly, but they certainly talk about these things at meetings and in other venues, and a poor reputation ultimately leads to less funding and support for your work. In regards to your other question, luckily, I have had mostly positive interactions on Twitter to date.

FLG: Is there anything else you would like to mention to our readers?

GL: The incentive structure in academic science is really skewed in favour of publications. This results in the churning out of many substandard papers, all due to the fact that each person has traditionally been held to the standard that they must be first or last author on some decent number of publications in order to be considered for certain grant monies. This dis-incentivizes collaboration, and I have personally witnessed behaviours involving withholding genetic data and pedigrees due to the fact that some group demands that they absolutely must be the sole first and/or last author on some paper. I have tried to counteract such behaviour by contributing to and helping to promote the BioRxiv preprint server, which was started by Cold Spring Harbor Laboratory Press as a way to encourage the open sharing of results. I am also on the editorial board of a new journal from CSHL Press, called Molecular Case Studies, which aims to present genomic and molecular analyses of individuals or cohorts alongside their clinical presentations and phenotypic information. The plan is to have a rapid peer-review process that is based on technical evaluation of the analyses performed, not the novelty of findings, and offers a swift, clear path to publication.

FLG: Thank you very much for your time, and good luck with your research!

GL: Thank you! ■