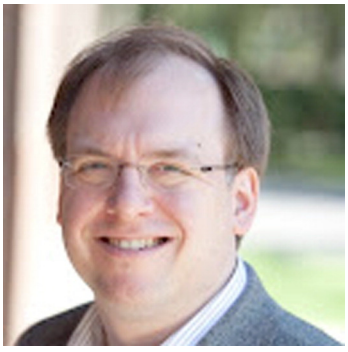
Applied & Translational Genomics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

Open access and data sharing: Easier said than done



This is an interview with Dr. Gholson Lyon, Assistant Professor and researcher at Cold Spring Harbor Laboratory. Dr. Lyon's work focuses on understanding the pathophysiological basis of neuropsychiatric conditions, with a long-term goal of expanding access to preventive services and treatment for these disorders. Dr. Lyon is also committed to the open discussion and management of the ethical implications of human genetics research, along with helping to move whole genome sequencing into the clinical world. He frequently speaks about the challenges of integrating genomics into clinical medicine and is vocal about the need for open access and data sharing and was among the 14,699 academicians who boycotted Elsevier for its restrictions on the dissemination of publications.

1. Open access initiatives have revolutionized knowledge generation and distribution. Is open access crucial to advancing genomics? Why or why not?

GL: Open access is a vehicle to democraticizing data because it ensures that anyone can get the data. But accessing the data is not useful unless one can also access reports based on the data.

2. If data are described in a publication, should researchers have access also to the actual data set to test the reproducibility of results, or conduct different analyses?

Yes, definitely. Currently, the NIH policy and practice are to place human genome sequencing data and genotyping data into controlled access databases, such as dbGaP. People say dbGaP is accessible but, in fact, access to it is impossibly hard. In order to gain access, one must apply through controlled access data committees. The process of applying involves formal applications entailing an enormous amount of paper work, including documentation of ethical assurances and IRB and HIPPA compliance. Last I checked, you had to be a Principal Investigator (P.I.) in order to even apply. Even applying as a P.I., I've been rejected because the committees wanted additional explanations regarding what I wanted to do with the data. So, in my view the NIH is extraordinarily conservative about whom they give the data to.

The access issue points to a problem with how bureaucrats and ethicists work together to develop policy guidance and practice standards. Both have the best intentions but they end up creating enormous bureaucratic hoops that make it very difficult for anyone to get access to the data, other than the very few established players. Obtaining access is easier if you're a researcher working at a major genome sequencing center with substantial NIH support because there are teams of people paid to help gain access to these data. But if you're working in a small genomics lab, it's basically a non-starter, because small labs can't afford to take the time and effort to fulfill these requirements, particularly if their chances of success are low. And the slim chance of success applies to genomic or genotyping data. Importantly, I think that there are a lot of younger people who could do breakthrough work if given the opportunity but as unestablished investigators, it's very hard if not impossible for them to gain access to it. Open access initiatives are intended to democratize data in just this way, namely permitting novices, as it were, the opportunity to gain access to data and to use it to advance knowledge. But in practice, access to this data is very limited and we may be paying a great societal price for these restrictions.

For these reasons, I'm pessimistic in 2014 that there will be near-term changes to enable truly open sharing of human genomic and phenotype data. There seems to be more of a way forward researching non-human organisms, such as plants and yeast, because these genomes are much more openly accessible. Perhaps it will take the next 5 years, or more, for people to become comfortable with truly open sharing of human genomic data.

3. The problem you describe is broader if you consider researchers in less resourced institutions or countries who would like access to publicly funded data. Might there be a solution that would enable you and others in less resourced institutions to continue to work on human data?

GL: I think there are established centers like the Broad, Baylor, WashU St. Louis, and some newly emerging players like the NY Genome Center, BGI, and MT. Sinai Genomics that have researchers with substantial resources, in the order of tens to hundreds of millions of dollars, and so are able to conduct very active projects where they're sequencing tens of thousands to hundreds of thousands of exomes and eventually whole genomes. And for the most part, I think that right now we're living in a very siloed world where certain genomic centers, such as those I just mentioned, will continue to dominate for the next 5 years. When the cost of a whole genome is about $100 using small, relatively cheap sequencing devices, then at that point there might be an inflection where people buy their own sequencing devices and generate and analyze their own data and the siloed nature of data generation and analysis ends. But until we can get a whole genome for around $100 I think the siloed status quo will continue to prevail. Well-funded research institutions will dominate the field, commercial companies will offer more targeted panels, increasing their revenue, and some academics, such

http://dx.doi.org/10.1016/j.atg.2014.09.008

Please cite this article as: Barash, C.I., Lyon, G., Open access and data sharing: Easier said than done, Appl. Transl. Genomic. (2014), http://dx.doi.org/10.1016/j.atg.2014.09.008

as myself, might just work with smaller genomes in non-human organisms and thus avoid dealing with the current bureaucracy governing access to human genomic data.

4. This 5-year period you describe would seem to have a significant negative impact on innovation. Are there dire consequences for researchers who can't get access to data or even society at large?

GL: I see very little leadership in America that is working on ways to engage people to collect and share a large amount of genomic data with detailed longitudinal phenotypic information. Without public engagement we will not have the data we need to implement genomics into clinical medicine. While, there's been talk about the Million Veteran Program, progress has been extremely slow on the uptake. The program is supposed to collect and sequence a million veterans. A contract was given to the company Personalis to sequence 1000 of those genomes but it is not clear whether sequencing is being done in any clinical grade manner, and there has been little comment by the government about the status of that project. Also, the NIH recently announced the delay of the National Children's Study that was supposed to start in 2015. After years of pilot studies, they were finally going to start collecting from birth the genomes of 100,000 children along with all kinds of clinical details. But now political gridlock, lack of funding and all sorts of other issues give the project an uncertain future. As far as I can tell, there's no leadership in the US government advocating for the need to be doing these kinds of big genome sequencing projects. To me, sequencing a million people in America in a clinical-grade manner and returning results directly to the donors would be very helpful to determine whether any mutation has any particular predictive value for certain phenotypes. There are other countries such as the UK that are embracing this. They've said that they're going to sequence 100,000 people. BGI in China has been talking about sequencing a million people. So other countries are very quickly moving ahead of America because of this lack of leadership and funding.

5. There are some sequencing projects underway; the privately funded PGP, and NIH funded BabySeq. Do you feel that these projects are unable to make even a small impact?

GL: I'm glad that the BabySeq is underway but it's small relative to a million people. These projects will make a small impact because of their small size, but I certainly hope that they scale up soon. Many of these projects could learn something from the consumer interfaces developed by companies such as 23andMe and Ancestry.com.

6. Your point about leadership is well taken. Two newly established global organizations would seem to be picking up some of this slack you refer to. The new Global Alliance for Genomic Health is establishing standards for enabling greater global data sharing and the Genomic Medicine Alliance aims to identify common strategies for implementing translational research and removing common delivery system and policy barriers to adopting genomics based care. Might these initiatives propel the need for data sharing?

GL: Sharing will happen, although perhaps not as quickly as I might like. I sometimes go back and read the literature on the Human Genome Project (HGP) and am reminded of what was happening with the HGP before Craig Venter announced he was going to undertake shotgun sequencing. Right now we have a governmental bureaucracy that's moving extraordinarily slowly such that maybe in 50 years time we'll have newborns sequenced on a large scale. What we need are visionaries like Elon Musk or Steve Jobs with deep pockets who announce that they're going to sequence whole genomes for a million people in the next five years and do it because they have the money and will to succeed. There are a few companies that have started talking about such an effort, like Human Longevity, which Craig Venter recently announced, but they haven't gained traction yet. But unless someone takes the ball and runs with it, I think we're trapped in the status quo.

7. Certainly there would be advantages to this kind of industry funded initiative but might any industry driven initiative, with proprietary interests, doom guarantees of open access and availability in the public arena?

GL: That's true, but what I'm saying is that we really need someone to shake it up and figure out a way to accelerate the governmental efforts. I think certain key things need to come together; namely public engagement and technological ease of uploading data. For example, Apple will be releasing I-health on their new operating system. They're trying to create a platform where people can upload and integrate all sorts of mobile health data. I could easily imagine that some people, including at Illumina, could easily build on that health platform; get your sequence, upload it to the platform and share it quickly with others. So, there's a lot of potential. I'm waiting to see if something develops in the next year or two that's relatively constructive.

8. Given that data requires human donors and to some extent they have to consent to third party access, do you think privacy concerns pose an obstacle to doing a large scale-sequencing project?

GL: I think there will always be a cohort of early adopters, people who enable open sharing, like the PGP project which is open to anyone or Jonathan Willbanks' portable legal consent that allows people to sign up and donate data. There are a lot of similar nascent efforts that are beginning to spread. With the right sort of donors or the right people to back it, a partnership with a disruptive company could make a large scale sequencing project possible. Yes, there are plenty of people out there who don't want to share their data, but in America, which has a population of ~318 million people, I think that you could find 1–2 million people who, with the right kind of engagement, would participate, and then such a project could be doable. For example, say you buy an iPhone, along with a marketing campaign that says donate your DNA and we'll put your genome on this application which allows you to connect up with other individuals, find out variants of interest, upload your phenotypic data and get back predictions, that could excite 1–2 million people to donate. It would require Apple teaming up with a company like 23andMe or a company which is not in the genomic sector but which has deep pockets and engages millions of people, like Facebook. I don't know if that will happen in the next 5 years or not, but it certainly will happen when we have a $100 genome. The problem is that currently, in July 2014, the true cost of a whole genome with good coverage is on the order of $1500–3000 and even more than that with good bioinformatics analysis. So, we are awaiting the development of newer, cheaper technologies for sequencing and interpretation.

Carol Isaacson Barash
Editor-in-Chief
J of Applied & Translational Genomics, Helix Health Advisors, United States
Corresponding author.
E-mail address: cibarash@helixhealthadvisors.com.

Gholson Lyon
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, United States
Corresponding author.
E-mail address: glyon@cshl.edu.

Available online xxxx

Please cite this article as: Barash, C.I., Lyon, G., Open access and data sharing: Easier said than done, Appl. Transl. Genomic. (2014), http://dx.doi.org/10.1016/j.atg.2014.09.008