Genomics Data 2 (2014) 49

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data collaboration will lead precision medicine

I was excited when Bart Wacek, a publisher at Elsevier, approached me several years ago to discuss creating a 'big data' journal. At the time I was at Illumina getting ready to launch BaseSpace and believed strongly that storing data centrally and driving universal reporting formats would help to both democratize access and promote innovation. I still do.

The timing of *Genomics Data* could not be better. Today the brightest minds coming out of top undergraduate and graduate programs are going into data-driven fields. Out here in Silicon Valley anybody who can correctly spell "machine learning" or "Bayesian" is practically guaranteed a job, if not venture funding. As DJ Patil, formerly of LinkedIn fame and now with big data startup RelateIQ has noted, there are today more data jobs posted on LinkedIn (61,000) than engineering jobs (53,000) [1]. That numerous institutions have begun offering data science curriculum is no coincidence, and suggests the trend is here to stay [2].

This bodes well for genomics in general, and in particular for the diagnosis and treatment of disease. As the genome research discipline has matured, the core analytical tools that were once centralized at genome institutes and leading academic medical centers (AMCs) have become ubiquitous, much to the credit of those organizations. This has freed academic and medical practitioners from developing tools and has instead shifted the analytical focus to where it ought to be: the patient.

Of course, tools are important. A surgeon wouldn't be able to perform surgery without a scalpel, and the same holds for building a better *de novo* assembler, a faster aligner, or a more intuitive variant interpretation engine all to gain insight from raw data. But focusing on building the best scalpel won't transform the way a surgeon treats her patient, or help her to determine whether surgery might be avoided in the first place. To deliver innovation that will guide those decisions, private companies, hospitals, and AMCs are starting to use 'omics' data both in clinical research and at the point of care. They are doing so by leveraging complex datasets to determine what tests to order next, what courses of treatment to administer to a patient, and how similar patients responded to that treatment course (how we define "similar" here is an interesting discussion in and of itself).

Recent technological and clinical advances suggest that a transition from research to mainstream is underway. A harbinger for this was a study just published out of the Mother Infant Research Institute at Tufts University that used massively parallel sequencing to detect fetal chromosome abnormalities an order-of-magnitude more accurately than the current standard of care [3]. With the cost of sequencing continuing to decline, these early successes will lead the way to standardized clinical sequencing across a range of diseases [4].

Undoubtedly the data challenges will shift as our understanding of the genome matures and a greater importance is placed on integrating multiple 'omics' data with an assortment of clinical indicators. Things like family health history, symptoms, medications, pathology reports, and patient outcomes can be as important to an accurate diagnosis and treatment as the underlying genetic makeup [5]. As the power of a wider breadth of data is realized, the network that results from numerous institutions sharing data across a range of diseases will have large statistics to drive insight [6]. The data challenge will be different but no less complex.

This will not simply happen on its own; hospitals, diagnostic labs and clinical researchers will need to actively collaborate. And yet, a very serious public discussion is underway about the acceptable use of data and how best to protect individuals from unintended consequences. Is it OK to collect and report metadata (this data exists in a clinical setting just as it does when we surf the web at home)? If data technology makes a mistake in how it analyzes or reports an individual's information, how do we fix the technology and how do we correct the public record?

Data has come under a heightened level of scrutiny in all aspects of life, and that's important. But rather than shy away from the debate we can step forward and responsibly support a collaborative paradigm. Our ability to realize the potential of precision medicine is at stake, and open access forums like Genomics Data will need to stay vigilant of the changing data landscape to provide platforms and formats that are as relevant tomorrow as they seem today.

After all, great young minds are counting on it.

References

[1] http://blog.relateiq.com/making-world-better-one-data-scientist-time/.
[2] http://datascience101.wordpress.com/2012/04/09/colleges-with-data-science-degrees/.
[3] D.W. Bianchi, et al., DNA sequencing versus standard prenatal aneuploidy screening. N. Engl. J. Med. 370 (2014) 799–808.
[4] https://www.genome.gov/sequencingcosts/.
[5] M.M. Segal, et al., Evidence-based decision support for neurological diagnosis reduces errors and unnecessary workup. J. Child Neurol. 29 (2014) 487–492.
[6] T.A. Manolio, et al., Implementing genomic medicine in the clinic: the future is here. Genet. Med. 15 (2013) 258–267.

Jason Blue-Smith
Real Time Genomics, Menlo Park, CA 94025, United States
E-mail address: jason.bluesmith@gmail.com.

http://dx.doi.org/10.1016/j.gdata.2014.04.001