Exploring Genomes: The need for an accessible map

Josh Tycko

University of Pennsylvania
Research Advisor: Dr. Brian Gregory
December 2014

It's been nearly 14 years since President Clinton stood with Dr. Craig Venter and Dr. Francis Collins and compared the achievements of the Human Genome Project to those of Lewis and Clark. "We are here to celebrate the completion of the first survey of the entire human genome. Without a doubt, this is the most important, most wondrous map ever produced by humankind. Today we are learning the language in which God created life." (Collins p 304, 2010) The massive undertaking was mostly motivated by the perceived promise that the genetic sequence would hold the key to treating innumerable diseases. Genomics has led to an enormous amount of great science, but it would be a stretch to say that the medical promise has come to fruition. Fewer than 60 genetic variants have been proven to be worth using in clinical care (Green 2013). Indeed, the biggest genomics headline of the year was the somewhat gloomy FDA crackdown on direct-to-consumer genetic diagnostic company 23andMe, who are no longer permitted to sell their medical reports (www.23andme.com). Successfully implementing genomic medicine will require the parallel progressive action of several key players, including: health-care providers, clinical molecular geneticists, genetic counselors, policymakers, researchers, translational specialists, electronic medical record (EMR) and other software vendors, and patients.

Here, we will focus on the role of bioinformaticians. According to the director of genomic medicine at Duke, their key next step to make genomic medicine a mainstream reality is the "development of curated genomics databases and means to query them to guide clinical decisions" (Ginsburg 2014). In light of his assessment, one could argue that the field has neglected Clinton's exaggerated, but insightful, metaphor. The human genome could be a wonderful map, but today's genome browsers – the interface through which we visualize the map – are difficult to navigate and give no insights as to where to look next for interesting genomic features. This lack of approachable interface has led to a sense in the public, but also amongst scientists who aren't trained in computational biology, that genomes are massively confusing and inaccessible (www.rickilewis.com, Church 2013). There is a demonstrated need for next-generation genome browsers that make "-omic" data accessible to non-computational biologists, including other life scientists, clinicians, and students.

Part One: Next-Generation Sequencing and the Browser Ecosystem

A well-designed browser's first objective is to rapidly synthesize massive data sets into a more easily digestible visual form; the accelerating growth of these data sets drives the need for next-generation genome browsers. The Human Genome Project (HGP) was a $2.7 billion dollar, international effort to piece together a blended genome of a few individuals (Church 2013). At peak capacity they were sequencing 1000 bases/second (Collins p303). According to ScienceNews (see Figure 1), that number is now closer to 6 million bases/second; moreover, the CIO of Illumina recently said the biggest expense of sequencing a human genome today is storing the data (www.sciencenews.org).

Indeed, since the invention of next-generation sequencing, reading DNA has been growing cheaper faster than storing information in silicon (Figure 2, Stein 2010). Recent estimates say the cost of sequencing a human genome is currently hovering around $5000 and is under $1000 at the few institutions with access to the new Illumina HiSeq X Ten Sequencer (Check 2014).

Lowering costs means there will be a greater number of people with access to raw genomic data, and thus a greater demand for interactive visualizations from which medical meaning can be extracted without computational biology or programming experience. According to recent news, "Saudi Arabia, the United Kingdom, and the United States have all launched projects that in total will sequence about 100,000 individuals. The clinical-sequencing market has been estimated at more than US$2 billion." (Ginsburg 2014) Datasets of that size could drive new research directions for clinical geneticists without computational expertise - if they had an accessible tool to visualize the genomic maps and generate new hypothesis. Moreover, as the number of people with sequenced genomes moves past six figures, there will be increasing demand from non-specialists to inspect their own genomic maps. Recent news from Nature claimed, "although the $1,000 goal is within striking distance, it has not yet enabled the depth of understanding needed to make full medical or biological use of the knowledge derived from ever more genomes. Attacking that problem is the next challenge of genomics." (Nature News 2014) More so than ever, genomics needs accessible, automated data interfaces to enable more minds, trained in different disciplines, to pore over the DNA map.

Today, the UCSC Genome Browser is the most powerful and frequently used browser, and it shares the space with a few other browsers that are similarly tailored to expert users. Fourteen years ago, Jim Kent of UCSC wrote the 10,000 lines of code that assembled the HGP's data as a graduate student (www.ucsc.edu). This program was shortly thereafter released (in September 2000) as the UCSC Genome Browser and has remained the dominant browser, followed by a few others such as Ensembl and the Integrative Genomes Viewer (IGV) (Kuhn 2013). The UCSC

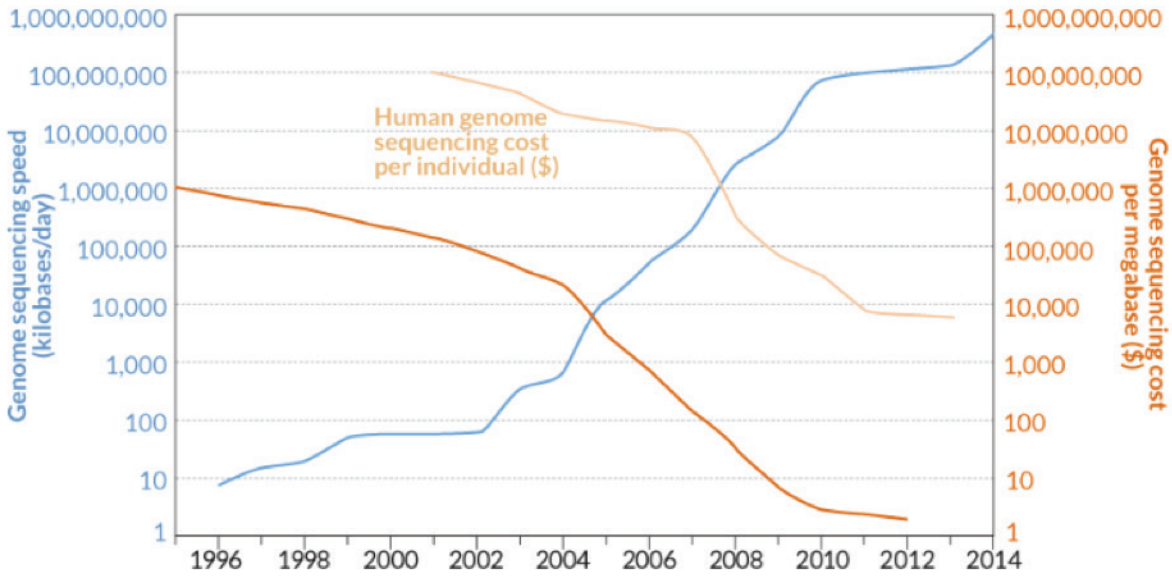

Figure 1: The cost and speed of genome sequencing (ScienceNews). The cost of human genome sequencing is now actually even lower than shown, for institutions with access to the Illumina HiSeq X Ten Sequencer.

Genome Browser serves up to 8 million page requests per week to 185,000 unique IP addresses per month (Kuhn 2013). There are some key differences: UCSC and Ensembl use a client-server model which offers access to very large databases of genomes, IGV uses a server-side system which is much faster than the web-based browsers but lacks the third-party data (Kuhn 2013). However, these prominent browsers are similarly only interested in expert users. In the developer's words, "the basic paradigm of the UCSC Genome Browser is to show as much high quality, whole-genome annotation data as possible and enable researchers to use their expertise to interpret data themselves" (Kuhn 2013). One newer, but less popular, browser is Jbrowse, which "helps preserve the user's sense of location by avoiding discontinuous transitions, instead offering smoothly animated panning, zooming, navigation, and track selection" (Skinner 2009). This open source project stands apart from the dominant browsers by offering an ease of navigation that feels more similar to the standards set by tech products like GoogleMaps (Table 1); browsers like UCSC display portions of the genomes as static images which must be reloaded to move (Skinner 2009). Unfortunately, Jbrowse lacks the useful large reference databases that make UCSC the gold standard.

Above all, the largest problem with these browsers (in terms of enabling widespread genomic medicine) is they are targeted for computational biologists who can code their own tools or use their expertise to select from the nearly 100,000 bioinformatics analytic tools in order to make sense of their data (Mesirov 2014). This paradigm makes sense for bioinformatics researchers who need to carefully work through novel and messy datasets, but does not give the non-computational scientist access to the genomics map. In the words of one cancer genomics expert, speaking at the time genome browsers were moving into cloud computing, "The reversal of the advantage that Moore's Law has had over sequencing costs will have long-term consequences for the field of genome informatics. In my opinion the most likely outcome is to turn the current genome analysis paradigm on its head and force the software to come to the data rather than the other way around." (Stein 2010)

Part Two: Genome browsing for clinicians and non-computational researchers

Genomic data is largely seen as confusing and inaccessible, even to life scientists without computational skills, let alone clinicians and the public. Some find genomics so intractable, they doubt it has anything of value to offer. In 2010, the major direct-to-consumer genomic testing companies, including 23andMe, were brought to a congressional hearing. Congress' investigative wing, the Government Accountability Office, provocatively wrote in their report, "the most accurate way for these companies to predict disease risks would be for them to charge consumers $500 for DNA and family medical history information, throw out the DNA, and then make predictions based solely on the family history information" (Wohlsen 2011). Needless to say, that's not quite true; FDA drug labels already include pharmacogenetic variants for 105 drugs, that is drugs that would be dosed differently or swapped based on the patient's genetic information (www.fda.gov). However, no commercial electronic medical record integrates pharmacogenetic information systematically, and barely any clinicians are trained to use the current genome browsers themselves to make sense of the data (Gottesman 2013). The Electronic Medical Record and Genomics (eMERGE) Network (which includes CHOP) has been developing tools and best practices for integrating EMR's with genomic data which could fill this apparent gap (Gottesman 2013). As it currently stands, the effort of genome sequencing is vastly overshadowed by the effort required to extract medical meaning from the reads. Ricki Lewis, a genetics-focused author, publicly stated the hurdles of interpretation are the reason she is choosing not to get her genome sequenced, at least for now. She noted, "when Stephen Quake, a Stanford University engineer and co-inventor of a DNA sequencing device, laid his genetic self bare in the pages of The Lancet in 2010, interpretation required 32 physicians"

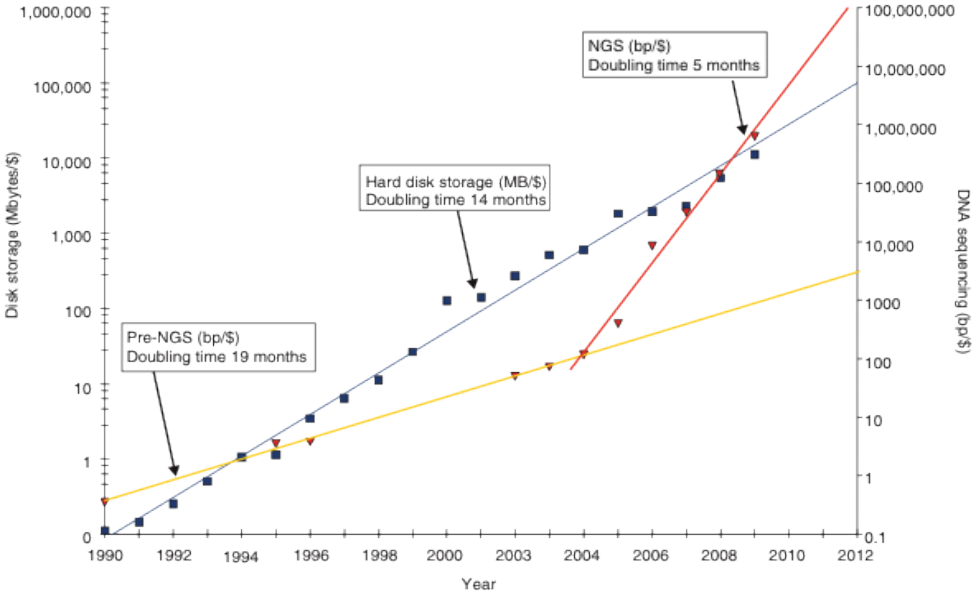

Figure 2: Decreasing costs of hard disk storage and genome sequencing. With the invention of Next Generation Sequencing (NGS), the cost of sequencing is lowering faster than the cost of hard disk storage (Stein 2010).

(www.rickilewis.com, Ashley 2010). Notably, Dr. Lewis does not mention any consideration of the possibility that she would be willing to inspect her genomic map herself, probably because the current interfaces would be unapproachable even with her Ph.D in genetics. There are some actors in this space; Personalis is one company developing a full platform from sequencing to medical interpretation (www.personalis.com). With best practices outlined by eMERGE and other researchers, automated analytic platforms could be developed that make genomic data usable by clinicians, in the form of genomics-integrated EMRs. But, the development of tools like browsers that make the genomic maps themselves more accessible may be farther off.

Outside of hospitals, improved genomic interfaces could accelerate biomedical research and lead to new or improved therapies. A key step here would be improved integration of genomic data with epigenomic, proteomic, and transcriptomic databases as many researchers are more interested in the expression and functions of the cell than its genetic code alone. "Although the generation of some of these [integrated] catalogues has already begun, major advances in technologies and data analysis methods are needed to generate, for example, truly comprehensive proteomic data sets and resources" (Green 2011).

If this were achieved, it is imaginable that all life science experiments would start with an inspection of the genomic map, same as a vacation to an unexplored place starts with a look at GoogleMaps. In my experience in various life science laboratories (specializing in gene therapy, synthetic biology, and epigenetics), this is not at all the case today. For example, a typical gene therapy project starts with a thorough reading of the literature in order to understand the target gene's function before designing a therapeutic vector to deliver the gene to the target cell. What the researcher tries to extract about the target genetic pathway from individual papers could be greatly supplemented by interacting with a next-generation genome browser that displays the target gene in its full context, including its epigenetic marks and related functional enhancers and ncRNAs. The therapeutic vector could then hypothetically be better designed so the delivered gene would interact with the target cell's transcriptional machinery more as it normally would in a human without disease.

A next-generation browser would better meet the needs of non-computational life scientists if it were accessible with natural English language queries. It could machine learn (by training on user's interactions or by integrating knowledge from publications) in order to generate potential hypotheses of interest without the non-computational user necessarily knowing which data tracks would be most revealing of their target gene's genomic context. For example, the browser could detect a binding site for a well-characterized microRNA upstream of the gene of interest and suggest the user inspect the ncRNA seq track in their target gene. This sort of strategy is floating around the bioinformatics literature, particularly since the wealth of data collected by the ENCODE project:

> Additional insights will come from combining the information from different catalogues. For example, analyzing genetic variation within functional elements will be particularly important for identifying such elements in non-coding regions of the genome. To this end, the GTEx (Genotype- Tissue Expression) project (http://www.commonfund.nih.gov/GTEx) has been established to map all sites in the human genome where sequence variation quantitatively affects gene expression. (Green 2011)

Success and failure in translational research are often determined by the scientists' knowledge of and assumptions about the disease's underlying biology; integrating genomic data into the research workflow seems likely to accelerate the development of successful therapeutics.

Synthetic biology, a loosely defined field focused on engineering biology often with genetic parts, could also greatly benefit from browsers accessible to its non-computational researchers. George Church, a synthetic biology scientist and founder of several genomics companies, believes genome sequencing has a purpose far beyond finding disease risk factors. In his book Regenesis: How Synthetic Biology will Reinvent Nature and Ourselves, he has no hesitations stating:

> The real point in reading [human genomes] would be to compare them against each other and to mine biological widgets from them – genetic sequences that perform specific, known, and useful functions. Discovering such sequences

> would extend our ability to change ourselves and the world because, essentially, we could copy the relevant genes and paste them into our own genomes, thereby acquiring those same useful functions and capacities. (Church p73 2012)

Most synthetic biologists do not expect to re-engineer humanity's DNA outside of disease contexts wherein patients lack safer alternative treatments. However, they are very interested in mining biological widgets for use in the production of drugs or bio-fuels, for example. An accessible genome browser would enhance the workflow of synthetic biologists by enabling rapid comparison and comprehension of interesting genetic widgets across organisms.

An accessible genome browser could also impact synthetic biology research while training a new generation of genomic scientists by integrating with iGEM (International Genetically Engineered Machines) – a student synthetic biology research competition. The program is now in its 10th year, and annually includes over 3000 participants in America, Europe and Asia, with estimated annual research expenditures up to $10 million (Vilanova 2014). A 10 year study observed that the finalists in the competition are frequently teams who find new genes in the literature to incorporate into engineered genetic circuits; in my experience these decisions are often made on the basis of a few papers that describe a peculiar phenotype in a bacterial cell, which can then be re-engineered for some useful purpose (Vilanova 2014). An accessible genome browser integrated with other "-omic" datasets and automated analysis could enable teams to mine for larger biological modules, like a negative feedback loop, as opposed to a single widget, like the inducible promoter that lead a German team to victory in 2012. iGEM has been very successful at training and motivating students to pursue life science research, and in return the students have produced useful biological tools (such as non-fluorescent reporter genes) and software. Integrating genomics into this energetic cycle of education and innovation only requires making the data interfaces more accessible to the non-expert user, and could inspire the next generation to explore the genome map.

Conclusion

Given the accelerating output of genomic data, there are ever-increasing numbers of people who could benefit from interacting with genomic maps. However, the current field of genome browsers predominantly caters to expert users, computational biologists, and is unapproachable even to other trained life scientists, let alone clinicians and students. To remedy this problem, accessible genome browsers should be developed with the ease of interface and interpretability of the data in mind as key design factors; they should accept queries with natural language, automatically generate hypotheses of interests by learning from user preferences and published workflows, and automate the standard analysis pipelines for use without programming skills.

Acknowledgements

I gratefully thank Dr. Brian Gregory, Assistant Professor of Biology at UPenn, for his insights, edits, and support.

Works Referenced

1. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. Lancet. 2010;375(9725):1525-35.
2. Burke W, Khoury MJ, Stewart A, Zimmern R. The path from genome-based research to population health: development of an international public health genomics network. Genet Med 2006;8:451–458.
3. Collins, Francis S. The Language of Life: DNA and the Revolution in Personalized Medicine. New York: Harper, 2010. Print.
4. Check Hayden, Erika; . Technology: The $1,000 genome. Nature News. 2014;507(7492):294.
5. Church, George; . Improving genome understanding. Nature News. 2013;502(7470):143.
6. Church, George M., and Edward Regis. Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves. New York: Basic, 2012. Print.
7. Fujita PA, Rhead B, Zweig AS, et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 2011;39(Database issue):D876-82.
8. Ginsburg, Geoffrey; . Medical genomics: Gather and use genetic data in health care. Nature News. 2014;508(7497):451.
9. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011;470(7333):204-13.
10. Green, R. C. et al. Genet. Med. 15, 565–574 (2013).
11. Karolchik D, Barber GP, Casper J, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42(Database issue):D764-70.
12. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinformatics. 2013;14(2):144-61.
13. Manolio TA, Chisholm RL, Ozenberger B, et al. Implementing genomic medicine in the clinic: the future is here. Genet Med. 2013;15(4):258-67.
14. Mesirov, Jill. "Approaches to Genomic Medicine." University of Pennsylvania, Philadelphia. 1 May 2014. Lecture.
15. Macarthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-76.
16. Nature News. How to get ahead. Nature News. 2014;507(7492):273.
17. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19(9):1630-8.
18. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11(5):207.
19. Vilanova C, Porcar M. iGEM 2.0-refoundations for engineering biology. Nat Biotechnol. 2014;32(5):420-4.
20. Wohlsen, Marcus. Biopunk: DIY Scientists Hack the Software of Life. New York: Current, 2011. Print.
21. http://www.rickilewis.com/blog.htm?post=882201
22. http://www.reuters.com/article/2014/05/06/23andme-genetictesting-idUSL2N0NS0Y820140506
23. http://www.scientificamerican.com/article/23andme-is-terrifying-but-not-for-reasons-fda/
24. www.23andme.com
25. https://www.sciencenews.org/article/gene-sequencing-future-here
26. http://www.ucsc.edu/features/genomics/milestones.html
27. http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm
28. http://www.personalis.com
29. http://www.seas.upenn.edu/directory/profile.php?ID=178