Biocuration Workshop Report

Prepared by Jenny Reardon and Jacob Metcalf

Summary of Workshop

On May 17, 2010, the Science & Justice Training Program and the Center for Biomolecular Science and Engineering (CBSE), and the CBSE Research Mentoring Institute hosted a meeting to discuss present challenges and future directions in “biocuration” (the practice of curating, annotating, and/or interpreting genomic information in order to make it medically, biologically, or evolutionarily meaningful). As genomic science accumulates greater and greater quantities of raw data, curation and interpretation have become key issues in the production of medical and evolutionary biological knowledge. This meeting brought the UC Genome Browser staff together with other key actors working on curating genomes to discuss common issues and problems. The meeting addressed two central questions:

1) Who can we trust to biocurate?

2) Will there be public whole genomes?

The first panel addressed the first question, and explored whether and how we can reach a point where there is a trusted source that establishes when a genomic difference associates with a meaningful trait. It considered the various efforts to create trusted sources (like the Human Variome Project, etc.), whether these efforts should or even can be united, and what different factors shape how various public and private efforts address this issue.

First Panel Members:
Elana Silver (Science Manager, Navigenics)
Shirley Wu (Biocurator, 23andMe)
Robert Kuhn (UCSC Genome Browser)
Mark Diekhans (ENCODE Gene Annotation Project, UCSC)

The second panel addressed the second question, taking up the question of whether whole genome sequences will become commonplace, and how this might matter for curatorial work.

Second Panel Members:
David Haussler (UCSC Genome Browser, HHMI)
Misha Angrist (Personal Genome Project and Duke University)
Michelle Cargill (Locus Development)
Malia Fullerton (University of Washington)
The event concluded with a public presentation by Misha Angrist on his upcoming book on the rise of personal genomics (detailed in another rapporteur report).

**Opening Remarks**

In her opening remarks, Reardon observed that although the cost of sequencing is rapidly approaching zero, the price of interpretation of genomic information only continues to increase in cost. Thus, while the early stages of human genomics might have been constrained by the price of sequencing and computing technologies, the current stage may be primarily limited by the considerable challenges (that are not just economic) associated with making meaning out of genomic sequences.

In his opening remarks, Haussler noted that it is important that we keep sight of big questions about human history and identity as we attach meaning to genomic sequences. Although the questions at the center of this meeting are specific, it is critical to attend to the big picture as we move toward medical genomics knowledge and infrastructures. By organizing a group of people conducting biocuration research, it was hoped that we could find important insights into how to come together to address the fundamental and critical questions biocuration poses:

- Is the goal to create personal or general knowledge? To facilitate medicine, basic research or both? Do these distinctions matter to the work of curation?
- To what extent are curatorial practices different when done in different places and contexts?
- Are there tensions between the different approaches or are they mutually supporting and additive?
- Where do the efforts meet and where do they diverge toward different goals?
- How can the legitimacy of these different efforts be established?

The following is a summary of the key themes that emerged from the discussion.

**Why is curation important and how should curation be done?**

Attendees presented a range of opinions on why biocuration is important for producing meaning out of genomic data. Most notably, attendees from private companies emphasized the need to create trust in personal or medical genomics knowledge, while researchers from the UCSC Browser emphasized the need to manage the deluge of raw data. In the case of personal genomics companies, the production of trusted interpretation is a key component of their business model—whether offering tests to personal consumers or medical doctors, it is necessary that curatorial work be guided by criteria that are reliable enough to guide important personal decisions. This creates specific challenges. For instance, results from past genome-wide association studies (GWAS)
often conflicts with old disease candidate genes. When doctors ask about tests for those disease genes, curators need to be able to make sense of varied experimental designs and thus are unable to make arbitrary cutoffs for reliable correlations.

The challenge of piecing together multiple forms of data created with conflicting experimental designs is one reason why personal and medical genomics companies sometimes produce different interpretations of the same genetic data. This conflict has led to attention from the press and government, and, in the words of one attendee, “created an urgent need to think about establishing an external agency to provide oversight and guidance” for curatorial work. Thus, for curators working for personal or medical genomics companies, curation is primarily about creating trustworthy levels of confidence in genomic knowledge—and endeavor that may benefit from external oversight.

In contrast, researchers from the Browser emphasized the need to cope with incredible amounts of data. For instance, the Browser just tripled its data storage to 240 terabytes, roughly 150% of the total bytes of data stored in the US Library of Congress, and this storage will only be sufficient for one year. With such a scale of data, researchers face a daunting task of deciding what it all means, and who or what has the ability to assign meaning. One Browser staff member noted a strong preference for regime of “buyer beware” in which it is incumbent upon users to be experts, and stated that the philosophy at UCSC is to just “throw it all out there.” It was also suggested that anything that comes from a “reliable” lab is worth curating, especially in light of the collapse of the central dogma of genetics and the explosion of biological complexity with the discovery of many modifiers and enhancers. A materialization of the Browser’s philosophy can be found in its curatorial practice of linking out to the original papers and various methods of tracking provenance of data. Thus, in the context of primary research, rather than being experts in interpretation, some Browser staff imagine their curatorial role as one of aggregating and presenting “all the data” (or at least all the data that they consider sufficiently reliable). This model of curation emphasizes the creation of trust through openness and transparency in which users can determine the reliability of data themselves.

The role of established authority

Given these diverging understandings of the role of experts and the purposes of biocuration, the role an authoritative body might play, and what that body should consist of became an immediate question. At several points in the discussion, both the Browser researchers and personal and medical genomics company researchers posited the need for an established authority, but with very different purposes in mind.

One Browser researcher argued that they “desperately need a vetting organization” for determining who ought to have access to certain forms of data (but it was not clear that
The other Browser members agreed with this sentiment). In light of the Browser’s commitment to put all of the data in the public domain, a member of a medical genomics company posed a question about how that data might be used by the Browser’s ever widening range of users. In particular, this geneticist posed a scenario in which the parent of a developmentally delayed child brings her child into a clinic and receives a medical genetics diagnosis from an insufficiently trained postdoc who takes information from the Browser. Although a stated purpose of the Browser is to provide useful genomics knowledge, a certain level of expertise is required to make use of the data that is freely available, and thus such data that is not pre-interpreted can lead to incorrect or harmful medical interventions. In response, this Browser member stated that to cope with such a scenario while maintaining a commitment to open science they require an “umbrella” organization that is designed to decide what information about genetic variation is open to members of the public and medical professionals. This would help protect the Browser from these cases. He also stated that they already had pulled many types of clinical data off of the Browser for fear of misinterpretation and have had to limit their representations to basic biological data. Even still, because the data on the Browser is driven by research, but may be medically relevant, users can get a “glimmer” of knowledge, but aren’t aware of what they need to know to understand that glimmer. Thus protection of patients and other users may require establishing an external authority tasked with vetting users.

The contrasting goals of medical and personal genomics companies led to a different vision of the kind of established authority that would contribute to their curatorial practices. On the one hand, differing curatorial practices may be necessary for offering different products—for instance, Navigenics and 23andMe are targeted at different audiences, and thus may use different interpretive techniques. They may also have different scales for presenting the certainty of the knowledge presented to end-users based on how they expect or intend those users to make use of it. On the other hand, generating and maintaining an audience for their products may require a central authority for establishing criteria for trusted research. Each company has had to develop these criteria individually because of the lack of an external authority, and this labor is expensive and time consuming. One researcher from a personal genomics company indicated that curation is partly about aggregating variation data, and partly about assigning confidence to the meaning of that data in order to offer a buffer between the consumer and the raw data. Where they need more guidance and consensus is how to provide that buffer; their tools will only meet their full potential when they are able to indicate when you “can take something to the bank” and use genome sequencing as a high-confidence diagnostic tool.

In addition to questions about the purpose of an established authority there were concerns about how such an authority could achieve buy-in from the relevant actors. One Browser member offered the example of the Human Genome Variation Society as an example of a group that has tried to achieve the status of a final authority. In this case there are
conflicting personalities and conflicting ideas about the aims of such an authority. Also, the HGVS focuses almost exclusively on clinical uses so would not serve those interested in non-clinical research.

Because high-throughput methods don’t lend themselves well to rare diseases with limited sample numbers these diseases also can get left behind as new methods make it cheaper to resequence and make new data rather than curate and integrate older data. Some databases, such as Decipher and International Standards for Cytogenomic Arrays, do exist to curate these rare diseases.

These various factors have led to the current state in which multiple databases exist, with no central authority to govern them. With major genomics projects, it is often the project that is able to generate funding that becomes the authority.

The role of standardization

Centralized curatorial standards, like all standardizations, would be a double-edged sword. While historically standards have helped facilitate genomics projects, they also exclude forms of knowledge (or bio-ontologies) that do not or cannot fall under the standard. As mentioned above, a curatorial effort that uses GWAS may well leave out other kinds of research (like candidate gene research) that could also be useful but were created with different experimental standards.

The role of the reference genome in curation poses similar problems. The Browser has become the host for the reference genome, an aggregate genome to which genomics data gets annotated. As Browser members pointed out, curation like this involves following moving targets but cannot be done without anointing one version as a reference to which variations can be indexed. One member stated that the reference is not a generalization that is meant to represent all variation, but rather a “fixed point in space that” can be used for comparisons. However, as databases get old that “fixed point” must shift. Because previous researchers relied on whatever coordinate system they were comfortable with there are multiple systems sedimented in older research. Curators that want to use automated programs for pulling data out of older papers face a substantial challenge when accounting for these shifting coordinates.

Yet, the ideal of a single set of terms, categories, or experimental design may not be feasible even if desirable. A Browser member suggested that they will need a thesaurus of genomic research because researchers will want to maintain the history and traditions contained in the many synonyms used throughout genomic research. Another member suggested it may always be the case that new technologies will encourage people to chase new forms of data rather than go back to understand and integrate old data—new technology causes not only increases in the quantity of data, it also creates splits in what the data means.
One of the roles of curation is to create and enforce standard formats for the creation (and dissemination) of knowledge so as to facilitate the movement (and thus growth) of knowledge across domains. For instance, genomics databases and journals could only accept datasets that strictly fit into a transparent format. There is some history of such efforts, such as GenBank, but the success has been rather partial. There is a lot of data that never makes it into GenBank. Without standardized terms those datasets are much harder to mine. The Browser makes some attempt to generate the transparency and traceability that standardization offers by always linking out to publications.

However, while standardization can help produce new knowledge, in some cases it may not be possible (or, at best, incomplete) due to the complexity of biology. Whereas some researchers from personal and medical genomics companies wanted to emphasize the need to provide trustworthy interpretation, one member of Browser insisted that interpretation remains an unsolved problem—he stated that there isn’t “some nice recipe” that is free from active research.

**Manual annotation and computational annotation**

The role computers will play in curatorial work was discussed at several points. Due to the quantity of data, it is impossible to aggregate and correlate all data available in public databases without substantial automation. No longer do curators only deal with the reference genome. They also face the prospect of more genomes, variation data, and a more complex model of gene biology that incorporates regulatory structures as well as genomic modifications such as epigenomic marks (i.e. DNA methylation, histone tail modifications).

Nonetheless, there are curatorial tasks that cannot be performed by machines, especially judgments about the quality of data that require situational knowledges about data provenance. The “gold standard” of manual annotation are teams of humans that are focused on single locus annotations. Human labor is also required for building and tuning algorithms. A member of the Browser team summed up the problem as “computers are fast and dumb; humans are slow and smart.” This slowness is a problem in an era in which processing and sequencing capacities are exploding creating a Sisyphus-like problem in which the mountain keeps getting bigger rather than the rock reaching the top and falling down again. It was noted that we cannot assume humans can always do a better job than computers; rather, humans and computers do certain complementary things well.

A major target for increased automation is data mining from journals. Using machine learning techniques, curators can train computers to scan text for keywords and associate data from multiple sources. However, the effectiveness of such automation is limited by variations in publication formats and data standards. For instance, automated mining
works well for text but not for graphs. Similarly, false or redundant results can become embedded in data as new technologies improve results. For example, RNA sequences are typically identified by the base position of the start codon. However, the most common change in RNA annotation is to make its sequence length longer, which makes identifying RNA boundaries from old data challenging. Similarly, some databases do not actively seek to remove old or incorrect data, such as the Online Mendelian Inheritance in Man (OMIM) database. Automation cannot adequately capture such concerns about provenance.

In some cases, institutions demonstrate a bias toward manual annotation. However, a Browser member argued that computational annotation can successfully do many of the tasks of manual annotators and identify problem cases that require human attention. Given the bias toward human annotation, human driven mistakes can become incorporated into annotations and are prohibitively expensive to correct. In some cases, improperly trained annotators have caused major problems through their mistakes or inconsistent work. Although one possible solution to the shortage of manual annotators is to outsource work to areas like China with adequate infrastructure and education, the costs of dealing with inconsistent work may make this an infeasible option.

The relationship between biocuration and medical genetics

A recurring theme in the discussion was the relationship between genomics and medicine. In light of the interpretive and technical challenges of biocuration, translating genomic information to a clinical setting poses significant problems.

Browser researchers and private company researchers alike argued that there is a need for a new form of education and training in medical genomics, creating a profession somewhere between bioinformatician and medical doctor or genetic counselor. While clinical labs are well equipped to run assays, they often lack the expertise to translate available research—which is often speculative—to decisions about medical interventions. Furthermore, no one understands how to treat unique mutations in a clinical setting—how does one know if that mutation is consequential or benign? Even a well-annotated reference genome is not well suited to resolve such a question. One personal genomics researcher suggested that the Genome Browser is so oriented toward basic research needs that we need a third entity that can map genomics research to current clinical tools. Another suggested that what is needed is a common, underlying database with two browsers: one for researchers and one for clinicians. However, workshops participants also pointed out that clinical problems often lead to research discoveries—such as a broken gene demonstrating something about the gene’s function—and thus wholly separate databases may not serve either clinicians or researchers well.

It was widely recognized that coping with the translation of research genomics to clinical genomics will require some sort of mandate for retraining and/or new types of training
for clinicians. This was in part due to the highly technical nature of genomics knowledge, and the situational or institutional knowledge necessary to interpret speculative research. Additionally, Browser members expressed some dis-ease about the prospect of interacting with clinical populations. This was in part due to their preferences as researchers and lack of appropriate training in interacting with research subjects, but it was also due to concerns about the legal, ethical, and epistemic consequences of providing medical advice.

The role of public genomes and representations of ‘the public’

The rise of personal genomics poses numerous opportunities and challenges and opportunities for curators.

As a member of the Browser team explained, the rise of personal genomics would be helpful for both research and medical purposes. The current model of anonymized genotypic data limits access to phenotypic data necessary to find correlations, but it also limits future interactions between researchers, medical doctors, and patients. Rather than having to either guess at a phenotype based on the available genotype or work through a medical bureaucracy, public genomes would make it possible to easily track relationships between genomes and phenotypes over time.

However, these new opportunities raise new ethical and legal questions—most notably, questions about whether whole public genomes and phenomes may expose human subjects (including extended families or communities) to sometimes unknowable risks and/or harmful disclosure of genotypic and phenotypic information.

Workshop participants expressed a range of views about whether these concerns are legitimate or not. One person suggested that a reason for governments’ reluctance to fund fully public genome research is the public’s “overblown notion” of the power of the genome to predict phenotypes—that is, members of the public have misconceptions about the ability to re-identify and predict highly personal traits from this data. Although he argued that genomics would lead to a reevaluation of what it means to be human, he played down the predictive capacity of genomics on an individual level.

Another researcher noted the significance of maintaining a relationship between scientists and their research subjects that goes beyond “just somebody in a database that you send a note once a year.” He argued that contrary to the common idea that de-identification prevents harm, maintaining that personal relationship may encourage ethical behavior from the researchers and give the research subjects more control over how their samples are used. In one case, Catalona vs. Washington University–Saint Louis, the de-identification of data was used to subvert ownership and remove the samples from
HIPAA protections, possibly weakening patient protections and ending a long-running relationship between a researcher and thousands of patient-donors.

One researcher from a medical genomics firm argued that public genomes will make curation more medically useful by capturing a greater number of variants and creating a stronger basis for quality standards. In some cases, cystic fibrosis and phenylketonuria polymorphisms were misclassified. Having a wealth of public genomes would increase the amount of documented evidence available to make associations between genetic variants and specific diseases. She argued that creating these associations would require a richer representative infrastructure than what the Genome Browser currently provides. The Browser is a “relatively flat structure,” and if you want to examine a complex disease you must be able to associate and examine more than one locus. The risk models that grow out of such complex associations will need to be transparent and included in biocuration. Public genomes will thus require more curatorial work, but will also make genomic data more meaningful.

Another researcher argued that it is necessary to epistemically and ethically re-assess the contextual nature of genomic and phenotypic data. She presented two cases of genomics researchers placing their own genomes and their families’ genomes and medical histories in semi-public domains as examples of how the specificities of family and medical contexts are necessary for making meaning out of genomic data. In the case of James Lupski, his family suffered from a rare heritable genetic disease and they had a richly documented history that could be carefully correlated with their genotypes. Whole-genome sequencing of Lupski identified the mutations responsible for the disease, and using this information to investigate family members revealed previously unrecognized sub-clinical phenotypes. Stephen Quake’s family, on the other hand, suffers from early and often fatal heart attacks, but sequencing his genome identified only hints of clinical relevance that were far less conclusive than in the case of Lupski’s family. This suggests that whole genome sequencing is only as powerful as the depth and breadth of phenotypic data available for correlation. Moreover, the epistemic need for this familial data implicates family members in research decisions, with ethical implications for consent and familial harmony. Contrary to our cultural fantasies of being radically open on our own, it is impossible to compartmentalize genomics data and still have it be meaningful—we will always be making decisions about others. Even highly penetrative diseases require familial knowledge, and the role of relationships between subjects can be easily lost when we focus on the genome as a string of letters. She suggested that our current infrastructures are unable to deal with the consequences for these others, whether they are family members or simply members of our extended population groups.

Genomics, workshop participants also noted, is also creating new forms of relationships between persons. One social science researcher argued that personal genomics creates an emergent form of family relations. For instance, 23andMe’s “Relative Finder” feature suggests other users who may be distantly related, reconstituting what it means to be a
cousin. This is consonant with a long history of social science research demonstrating that family relations are constantly being reinvented. Others discussed the ways in which ancestral genomics has become an entertainment phenomenon, creating new types of relations between people and their distant families. Also noted was the role of personal genomics and the growing reliance on genetic forensics in criminal justice. It is unclear what sort of access to personal genomics data the justice system will have, despite privacy assurances from the personal genomics providers. The case of the history of the HeLa cell line was also discussed, pointing out that even if one consents to donate tissues for use in current scientific regimes, it is impossible to know how those tissues may be used in future technologies.

The relationship between these broader social and ethical issues, and the work of curation, is an area in need of further exploration.

**Action Items**

Since our meeting in May, the issues surrounding biocuration have only gained in prominence and salience. They are, for example, at the heart of what is at issue in the recent government inquiries into Direct to Consumer (DTC) genetic testing. What the Government Account Office report on DTC testing describes as deception of consumers might in many cases be better thought about as deep unresolved epistemic and ethical questions about how to interpret and present genomic data (and how this work of curating the data differs depending on the intended audience). Given this, some action items coming out of the meeting might include:

1) Writing a joint white paper that raises awareness about some of the problems addressed in this workshop. We might think about publishing this in a journal like Science or Nature.

2) Creating themes for subsequent workshops. An initial list of possible topics includes:

   • The Clinical/Bioinformatic Interface: This workshop could explore whether different curatorial infrastructures are needed for basic and clinical research, whether new forms of expertise are needed to deal with the rise of bioinformatics as an important aspect of medicine, etc.) & vice-versa

   • Standardizing Curation? This meeting could explore whether and how curatorial work should be standardized, and explore further the question about whether recognized authorities are needed to govern this work

   • Democratizing Genomics: Both Browser staff and private companies emphasize the need to democratize genomic knowledge, but most Browser
staff mean that it should be open and free and less interpreted, while those
working in private industry often mean that it needs to be trusted and pre-
interpreted, which likely means closed. Are these oppositional visions of
democracy? Are they complementary/compatible visions? Can both visions
exist side-by-side?

3) Conduct empirical research that aims to document different curatorial practices
and their basis in particular research paradigms.