Genomics Privacy in the Age of Personalized Medicine

Dov Greenbaum, Arif O Harmanci, Mark Gerstein, MBB, Yale University.

The issues of privacy and disclosure are two sides of a weighty coin. Computational biologists and other scientists involved in genomic research need to be constantly cognizant of the push and pull of these two important concepts. Clinical genomics research in particular raises a number of particularly poignant concerns as society struggles between invasions of privacy such as recent efforts by the FBI and the NSA, and our own (surprisingly) personal disclosures on social media sites or via apathetic acquiescence to large data collection efforts. With regard to privacy there are numerous computational efforts that have heretofore offered to provide both the robustness of protection and the ease of use to be effective in manipulating the terabytes of data before the genomics researcher. Unfortunately algorithms alone have thus far failed to provide either the necessary strength to foil those intent on obtaining information or the promised agility to manipulate the vast datasets. While technical solutions advance, they cannot stand on their own and this paper proposes and outlines a licensing scheme, similar to those used by professional organizations, that not only enforce a code of conduct and punish those who fail to live up to that code, but also mandate required continuing education to limit the possibility that the code will be violated inadvertently. It is the use of the social and the technological advances together that will likely create not only an environment that fosters research and innovation, but also one that is responsive to privacy needs and norms.

Background;

Researchers and health care practitioners regularly struggle with issues of balancing data disclosure with practical concerns. Often times these are relatively paternalistic concerns. For example, many researchers feel that their average patient lacks the proper knowledge and perspective to deal with complicated medical data. This is not necessarily a wholly unfounded opinion: Clear and actionable medical data often is confounded by statistics, data lag times and biological uncertainties. As such, in many instances, scientists and researchers may choose to provide only the minimum of data regarding diagnoses and prognoses.

Broad and relatively easy access to medical information on the web has provided concerned patients the ability to research their symptoms further, confounding those well-meaning efforts of the data-withholding doctors. But even more so, the advent of cheap and consumer friendly personal genomics will further substantially upset the balance of proper patient care and full data disclosure by providing, often direct to the consumer, medically relevant but statistically complicated data.

The personal genomics industry, an outgrowth of the confluence of diverse technological advancements, is a small but growing sector. As a result of steep declines in sequencing costs [1], genome-wide interrogation technologies, plummeting memory costs and booming computational power, a growing number of companies now offer to collect, analyze and return genomic information direct to the public. In addition to the commercial entities that promise a wide range of dubious to medically relevant information, there are a number of research organizations that also take advantage of these technological developments and economies of scale to collect and process thousands of genomic datasets for research.

However, in addition to concerns that patients provided with their genomic data may either under-estimate or over-estimate their current situation, acting imprudently or, alternatively, inappropriately complacently, there are additional concerns that may be underappreciated by the clinical community: in particular, patient privacy.

Privacy and Genomics

While clinical researchers may otherwise be trained in standard privacy procedures, the necessity of the incorporation of privacy procedures may not be felt as acutely by clinical researchers when dealing with large complex genomic datasets. This may be especially the case when clinical researchers are not interacting with patients directly, but rather just cataloging, annotating, and or otherwise databanking large genomic data sets.

However, even in these situations, patient privacy ought to be considered and maintained with full due care: in contrast to standard medical data, genomic information, often SNP data or even full sequence data, is inherently personally identifiable information. By its very definition, raw genomic data identifies the owner. And, in further contrast to other forms of medical data,

genomic data is largely shared with close family members. Therefore, even in instances where patients have provided broad permissions to use and access their genomic information, care must be taken to persevere patient privacy as the data implicates not only the immediate owner of the sequence but many third-party relatives as well.

What part of genomic data is so revealing? While humans share the overwhelming vast majority of their DNA with most other members of their species, there remains thousands if not millions of potential variations in the genetic data.

These variations, which range, in their impact on our lives from totally irrelevant to life threatening, include mainly, but not limited to, single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and other large scale complex rearrangements like structural variations (SVs). Although many of these variants are commonly shared among large portions of all human populations, particular subsets of variants are shown to be highly discriminative and they can be used as fingerprints. Thus, according to the theories underlying the FBI's Combined DNA Index System CODIS database, and similar systems, just 13 highly variant short tandem repeats, that are otherwise thought to have little to no other medical value, can categorically identify the owner of a DNA sample, i.e., the suspect in a crime [2].

The discovery, cataloging and annotation of DNA sequence variants, however, are only the half of the story. The advancement of sequencing technologies and laboratory techniques has enabled development of many assays to probe, for example, the epigenetic and transcriptomic states of the cell, e.g. gene expression levels and DNA binding protein levels and some of these are medically actionable [3], [4].

These assays generate, like their more simpler sequencing counterparts, large datasets of identifiable information. However, unlike genetic variants uncovered through the more straightforward sequencing route, functional genomics data includes not only legacy data relating to prederminable conditions, traits, sex and race, and the like, but functional genomic data has the potential for being even more intrusive as it can reflect even privately held life choices, for example diet and where one resides.

One can even combine recent technologies to further estimate the genetic variants from other genomic measurements with a high degree of accuracy [5]. For example, the activity levels of some of the genes are very highly correlated with some of the variants, which are referred to as expression quantitative trait loci (eQTL). An example is a large deletion that deletes a gene in whole, which would shut down all the activity of the gene.

Although a lot of the studies concentrate on protecting the DNA sequence variants, the attention on the privacy protection of functional genomics data has been quite limited [5]–[7].

We illustrate Fig. 1 a general setting where sensitive information could be extracted from publicly available databases. In this example, two members of the family have genomic

screening performed, one through a hospital visit and other through a genealogy company. The hospital first de-individualizes the records (by removing the name and other personal information) and releases the data in a database to public with the sensitive information. The genealogy company put the records to an online database, although the database can only be queried. An attacker downloads the public database from hospital. The attacker also downloads several other public databases like voters databases and yellow pages. The attacker uses a prediction algorithm to estimate the genotypes of the individuals in the hospital database and queries the genealogy database to get possible individuals and cross-checks them with the other public databases. By linking multiple databases, the attacker performs a predictive linking attack. Although this is a complicated attack, the attacker can scan many individuals at the same time and can identify a small number of people.

Results:

As we are just learning about the extent to which the sensitive information can be "mined" from the data generated by genomic sequencing technologies, it is not easy to foresee how much privacy breaching information can be extracted.

While it remains necessary to continue to research the technical aspects of extracting and misappropriating private genomic sequencing and functional data, this remains a non-trivial issue.

Any type of genetic data that open to public needs to go through a complicated de-individualization procedure, which removes any sensitive information from the data. This, however, should be done cautiously since the sensitive information may contain important biological content , removal of which may render data useless to the research community.

The formalisms like differential privacy[8] and homomorphic encryption [9] based data analysis establishes the theoretical framework for building secure genomics data sharing systems and policies, applicable especially to cloud based systems. The issue with these formalisms, however, is that it is necessary to define the amount of sensitive information leaked for any type of genetic data and different types of genetic data can be combined in the extraction of sensitive information, which is not well understood yet. In addition, the practicality of these formalisms is still questionable.

Technological solutions should be also studied from the "Big Data Science" aspect of genomics. From this perspective, genomics has huge computation and storage requirements. Cloud based computing infrastructures, with the virtually infinite amount of compute power, presents the opportunity to move the processing and access of the data to the cloud. Many research and clinical institutions already turn to utilizing the cloud based services for genomics analysis rather than financing in house high performance computing services.

While technical solutions may be helpful in promoting and protecting privacy rights of those of us who have identifiable genomic information, they can't be relied upon solely to provide full scale protection. Not only because the natural trade-off between technological protections and accessible data will leave many in the field leaning toward easily accessible and manipulatable data at the expense of privacy, but because every technological enhancement, particularly in areas of privacy and sensitive data, is an implicit invitation to hackers to break through the technological protections.

With this in mind, any technological effort must be accompanied by concomitant efforts both socially and legally to change the underlying concerns regarding the misappropriation of private genomic information.

Cloud computing provides not only increased computing power and efficiency, it could also potentially provide a secure area for storing and accessing data. More importantly, access to the cloud can be monitored and privacy breaches can be sourced and punished, perhaps adding an extra incentive to researchers to be more careful with this data.

Further, funding agencies can promote this centralized cloud of genomic data. With data confined to a single location, access to that data can be made contingent on many aspects, including, knowledge of privacy regulations and best practices. Access can be further made contingent on obtaining a license indicating, as other professional licenses do, a current and continuously updated understanding of the science and the responsible use thereof.

Socially, some of this change is already happening in the form of social media as younger generations create a reality wherein every banality of their life is for public consumption, as long as it fits within 140 characters or can be summarized in 6 seconds of video. In conjunction with a general liberalization of western society, political correctness and a growing acceptance of what was in an earlier generation shunned, younger generations are likely to see a world wherein medical and psychiatric stigmas previously hidden lose their negative connotations. In this world, we might find that heretofore private medical information is freely shared.

However, regulatory changes, complimentary to the social changes are also needed. Laws need to be passed and regulations promulgated that limit the liabilities associated with disclosing genomic information. The Genetic Information Nondiscrimination Act (GINA) is a starting point but broader laws with more bite also need to be passed to protect consumers, many of whom may likely see the genomes of their close relatives become public knowledge as the personal genomics industry grows.

References:

[1] "DNA Sequencing Costs." [Online]. Available: https://www.genome.gov/sequencingcosts/. [Accessed: 07-Mar-2014].

[2] J. Ge, A. Eisenberg, and B. Budowle, “Developing criteria and data to determine best options for expanding the core CODIS loci,” *Investigative Genetics*, vol. 3. p. 1, 2012.

[3] F. W. Huang, E. Hodis, M. J. Xu, G. V Kryukov, L. Chin, and L. a Garraway, “Highly recurrent TERT promoter mutations in human melanoma.,” *Science*, vol. 339, pp. 957–9, 2013.

[4] C. Sotiriou and L. Pusztai, “Gene-expression signatures in breast cancer.,” *N. Engl. J. Med.*, vol. 360, pp. 790–800, 2009.

[5] E. E. Schadt, S. Woo, and K. Hao, “Bayesian method to predict individual SNP genotypes from gene expression data,” *Nature Genetics*, vol. 44. pp. 603–608, 2012.

[6] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, “Identifying personal genomes by surname inference.,” *Science*, vol. 339, pp. 321–4, 2013.

[7] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, “Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays,” *PLoS Genet.*, vol. 4, 2008.

[8] C. Dwork, “Differential privacy,” *Int. Colloq. Autom. Lang. Program.*, vol. 4052, pp. 1–12, 2006.

[9] V. Vaikuntanathan, “Computing Blindfolded: New Developments in Fully Homomorphic Encryption,” *2011 IEEE 52nd Annu. Symp. Found. Comput. Sci.*, pp. 5–16, 2011.

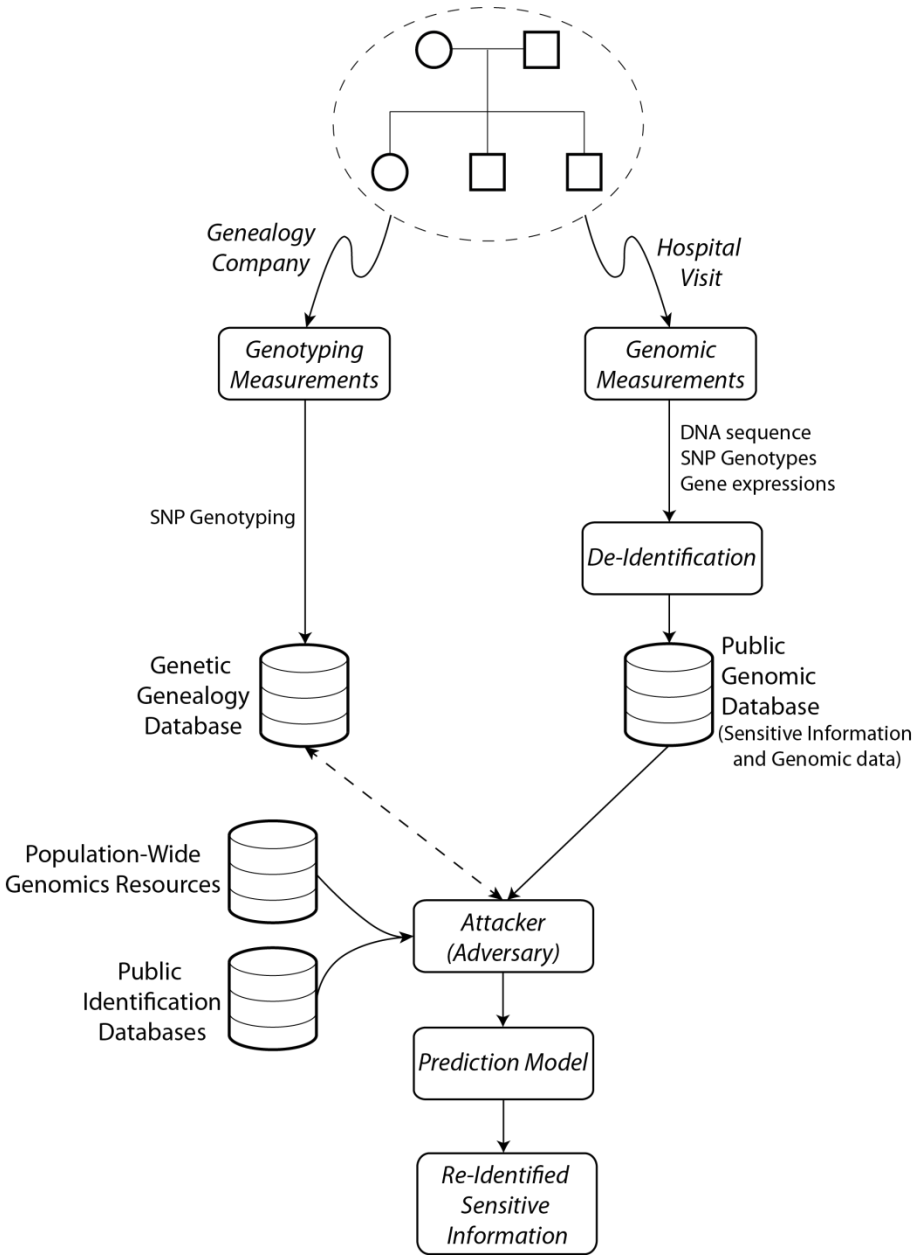

Figure 1: Illustration of the genomics prediction attack: A family with five members is illustrated on top with the tree. One of the members visit a genealogy company (illustrated with the left arrow) and the company performs SNP genotyping. The results are stored in the "Genetic Genealogy Database". Another member of the family visits a hospital where genomic measurements are performed. These can generate different data types like DNA sequence, SNP genotypes, and gene expression levels. The hospital records of this member are de-identified by removal of the name and released in "Public Genomic Database" that contains the health related sensitive information and the genomic data. Attacker downloads the genomic database from hospital. In addition, he downloads the population-wide genomics resources and public identification databases and the genetic genealogy database to perform the prediction based linkage attack to generate the re-identified sensitive information in the hospital records