Genomics and Privacy

Frank P. DeVita

Columbia University

May 2016

Introduction[1]

Twenty years ago, the Human Genome Project declared that, "all human genomic sequence information, generated by centers funded for large-scale human sequencing, should be freely available and in the public domain in order to encourage research and development and to maximize its benefit to society" (Human Genome Organization 1996), and these sentiments have been reaffirmed internationally over the years (Human Genome Organization 2003, European Society of Human Genetics 2003),with increased attention toward informed consent (United Nations Educational, Scientific, and Cultural Organization 2003) and privacy protection (Organisation for Economic Co-operation and Development 2009). Genomic sequencing technology has revolutionized our understanding of biology and disease, and allows us to explore the basis of all living things through their nucleic acid codes in the form of character strings (e.g., ACGTTGAC) with computational techniques. With sequencing technology we convert "wet" nucleic acids into digital bits of information that can be stored, analyzed and shared *in silico* to translate discoveries

[1] This paper is a philosophical-legal analysis of issues relating to privacy and sequencing technology. It is structured as a theoretical debate between two opposed parties and explores both sides logically. There is not meant to be a definitive conclusion, but a laying out of the issues and details involved. The views herein are my own unless otherwise specified, and attributions of views to particular sides or entities are merely academic and do not reflect the beliefs, positions, or knowledge of the actual individuals occupying scientific, academic, or government positions hereto alluded for realism.

into practical applications. For example, sequencing can be used to detect the presence of an infectious microorganism by its genetic signature, and DNA fragments associated with identity, health, and disease risk can be detected by minimally invasive blood tests. Sequencing technology thus generates information as sensitive as social security numbers, birth dates and addresses. Therefore, privacy and information security are important concerns as sequencing technology advances and becomes more widely available. As sequencing is increasingly used in the clinic, it generates valuable and actionable information about the molecular genetic features health and disease. This paper explores whether or not this information should be automatically anonymized, standardized and made available to the scientific community for analysis.

Issue

Should clinical genomic data be automatically de-identified and made available to the global scientific community through a central open source portal?

Parties

**Note these are fictional parties. Positions to not represent the views of the individuals currently holding these posts.*

Party A – "Pro"

Director, The Broad Institute of MIT and Harvard

Party B – "Con"

Director, National Human Genome Research Institute

Party A's Position on the Issue

Concise Statement of Party A's Position:

The creation of a global open source clinical genomics portal is a life-saving endeavor, and a major victory for international and local public health. With such a database, genomic information could be downloaded and analyzed by researchers worldwide, fostering parallel discoveries about health and disease with massive sample sizes. The *de facto* anonymization and computational masking of this information protects rights to privacy.

Detailed Rationale for Party A's Position:

Standardized genomics will advance science, improve healthcare, and benefit public health.

Currently, genomics projects are dispersed globally, with leading labs conducting experiments, writing algorithms and analyzing data independently. However, the incongruence of methods and techniques across genomics research nodes presents a big data problem that impedes the strength and clinical translation of genomics discoveries. Some research groups publish their findings in large public databases such as GenBank (USA, http://www.ncbi.nlm.nih.gov/genbank/) and the European Nucleotide Archive (http://www.ebi.ac.uk/ena), while others furnish proprietary public databases such as Memorial Sloan Kettering's cBioPortal (http://www.cbioportal.org). It follows that the results from these independent nodes of research results are structurally different, and data sets generated from different labs across the globe structurally incompatible with one another. As a result the global set of genomics data cannot presently be integrated and analyzed as a

full set, effectively impeding potential groundbreaking discoveries. This ultimately costs lives by slowing down genomics discovery. With a standardized global genomics database, laboratories worldwide could run their proven algorithms on a huge data set, increasing the likelihood for significant, actionable and clinically translatable results. Such large-scale efforts will greatly accelerate genomics research, which has practical implications for therapeutic development, public health, scientific understanding and health policy. (Nature 2003)

Big genomics will save lives through parallel discovery.

Creating a central, standardized open source clinical genomics database will bring together disparate sources of genomic data and will make large scale, global genomics projects possible by putting more data in the hands of researchers. Open data from the Personal Genome Project (PGP) has been accessed by more than 34,000 investigators since 2008 (~1,000 per month per dataset). By contrast, there have only been 7 projects associated with the International Cancer Genome Consortium (ICGC) since October 2011, with an access rate of ~0.00023 investigators per month. (Greenbaum 2011) If the former projects also included clinical and translational components, new genomics data would be made available to researchers and clinicians quickly, creating a reliable and actionable global database. As seen in many cases in engineering and computer science, this type of parallelization greatly increases output. In biotechnology, this is exemplified by massively parallel sequencing. (see Am J Hum Genet. 2009 Aug 14; 85(2): 142–154. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2725244/) In the case of genomics, this will save lives by expanding our knowledge of the connections between genomics, disease and therapy through real-time data generation and big data analysis. Moreover, a standardized and open source format fosters the independent development of tools and algorithms that can be shared and used across the world

because they are built to work with standardized data set. More and wider investigator access will mean more progress in the field, which will ultimately save lives by making information available for research, translational and clinical applications sooner than is currently feasible. This will allow researchers and clinicians across the world to collaborate with one another easily and use the global knowledge base to treat patients and make genomics discoveries in parallel.

Rights to privacy and information security can be reasonably preserved in an open or "free" genomics environment.

In the information era, privacy and security are of paramount concern and genomics is no exception. There are many aspects of genomic information that can be used to identify an individual from their DNA sequence. For instance, single nucleotide polymorphisms (SNPs), other unique genomic variations, and contextual data such as location of consent can be used to triangulate an identity. (Erlich 2014, Figure 1)

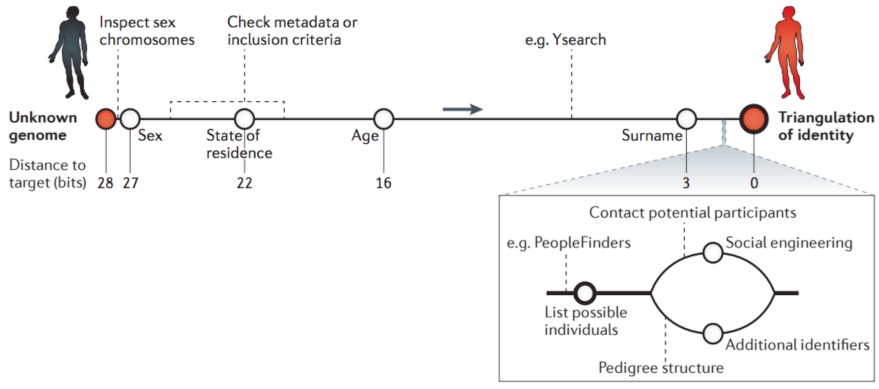

Figure 1: A possible route for identity tracing *(adapted from Erlich 2014):* The route combines both metadata and surname inference to triangulate the identity of an unknown genome of a person in the United States (represented by the black silhouette)…The adversary uses public record search engines such as PeopleFinders to generate a list of potential individuals; he or she can use social engineering or pedigree structure to triangulate the person (represented by the red silhouette).

Any database of genomic information needs to be anonymized and identifying information, genomic and personal, can further be masked prior to import to maintain patients' rights and privacy. (ibid) There are many advanced computational techniques that can be systematically employed to mask sensitive information from genomic data, and specifically calibrated to protect sensitive pieces of genomic information, e.g. single nucleotide polymorphisms (SNPs) and other variants that could identify an individual. Advanced filtering and averaging techniques can also be applied to protect the identities embedded in a raw genomics data set. Such techniques are already implemented in the banking and finance industries, where professionals are constantly manipulating and analyzing large swaths of information associated with individuals, and it is reasonable to assume that forms of these techniques can be ported for genomics research. (Greenbaum 2011) Through these types of precautions and safeguards, we can defend genomics databases against unauthorized or malicious third parties trying to identify an individual from his or her genomic information.

Party B's position on the Issue

Concise Statement of Party B's Position

Creating a central open source clinical genomics database with automatic import violates patients' privacy rights and informed consent protocols, and also introduces a new and serious information security risk into the clinical research environment. A big genomics research program of this magnitude also necessitates the creation of a new international regulatory framework within which countries could create policy and exchange data. These issues must be resolved for international, publicly funded genomics research to be possible.

Party B's Position and Rationale for Why Party A's Position is More Compelling

Party B maintains that a global open source genomics database with automatic data import should not be created because it violates patient privacy, and that the associated political/legal/regulatory hurdles are overtly complex. These are legitimate concerns, however they can be overcome. Privacy concerns can be put at ease by ensuring that the proper legal and administrative frameworks are developed and deployed to protect patients' rights to privacy and informed consent. This may be difficult on a global scale, but the benefits to healthcare, science and public health outweigh these costs. Further, computational masking and encryption can be used to protect sensitive genomic information from malicious parties. In sum, all information would be stripped identifying marks leading back to particular individuals. Practically speaking, the institutions collecting and sequencing clinical samples can integrate redaction steps into their existing data handling and storage processes. Institutions may also choose not to participate in the big genomics effort at all if adherence to its requirements places too much strain on local resources, however this would not prohibit their access to the database. Regarding informed consent, unique documents for any global-scale research projects can be created and presented to any patient considering genetic testing at a participating institution. Alternatively, modified "open consent" protocols can be created to inform patients that, and why, complete anonymity cannot be guaranteed. (Personal Genome Project 2016). The Personal Genome Project states:

> Because we cannot guarantee privacy and we are committed to sharing data for the advancement of science, we feel the most ethical and practical solution is to collaborate with individuals who are comfortable sharing their data without any promises of privacy, confidentiality or anonymity.

(The PGP's full consent form can be accessed at https://my.pgp-hms.org/static/PGP_Consent_2015-05-05_online_stamped.pdf.) Still, a patient could decline participation if they desire. In any case, global open source genomics projects would always ensure that patients whose data are collected and analyzed experience the same rights to informed consent and information security as they would in more traditional clinical research settings.

Detailed Rationale for Party B's Position

Patients must retain rights to informed consent and release of their medical information.

At all times, patients must be able to permit or deny the release of information from their medical records, and this right extends to molecular profiling data. Therefore, some sort of release of information and informed consent procedures must be developed and followed closely in order to create an ethically sound big genomics research program. Patients must maintain their rights to control over the availability and dissemination of their genomic information at all times, and furthermore must be informed of the possible ways in which their genomic information could be handled, stored, used and shared. Patients must also be adequately informed of the research objectives involving the database and the information contributed to it voluntarily. These informed consent discussions need not detail research down to, say, algorithms, but should clearly demarcate the scope and purpose of genomics research efforts using patient data. Moreover, participants volunteering their genomic data must informed of how their data is connected to their identity (Erlich 2014, see Figure 1 above), and how it is protected or masked to block malicious parties seeking to steal identifying information. While it is true that "open consent" protocols in which volunteers acknowledge that information privacy cannot be guaranteed are now being used (Personal

Genome Project 2016), it is not reasonable to assume that such a system can be implemented in large-scale, government funded clinical research without great difficulty due to the links between genomic profiles and other non-genomic medical information. Since the big genomics platform being considered here seeks to leverage clinical data, it is untenable to use a true "open consent" protocol due to inextricable links to medical records. Informed consent in *clinical* genomics is different than informed consent in non-clinical genomics research projects, e.g. the Personal Genome Project. In the former, genomic data is part of the medical record, which also includes general identifying information such as social security numbers. In the latter, open consent is less controversial because the information at risk solely the molecular profile, not one's full medical record. Moreover, a participant's information may be analyzed several times over for different purposes, which further complicates informed consent. However, some middle ground may exist between open and closed consent that both provides anonymity and makes sufficient data available over a long or indeterminate amount of time for different purposes.

Objectives and dynamics of global genomics projects must be clearly defined.

Methods and limits of genomic information sharing must be unified and clearly defined to the agreement of all countries, health agencies, research institutions and participants that would be involved in big genomics. In order to define such parameters, it will likely be necessary to create an international cooperative forum in which members can discuss and define the nuances of privacy protection and data exchange in this context such that patients' rights to privacy of their protected health information are preserved, and any data collection, analysis, sharing and publication meets the ethical and privacy standards of participating health agencies and governments. Big genomics also presents research dynamics that must be well understood by all participating parties. A single big

genomics research project will have a defined objective and "cut" or "slice" the database in a particular way, and future projects with different objectives will likewise cut the database uniquely. This means that a participant's genomic information may be analyzed many times in many different ways. Therefore, patients must also be informed of the fact that future research may probe genomic details more personal than what is found at the surface (e.g. known disease-causing mutations), such as unique genetic variants and other highly specific genotyping analyses that can theoretically be traced to one's identity. (Greenbaum 2011; Erlich 2014, Figure 1 above) A minimum number of 75 independent SNPs, if not fewer, will uniquely identify a person, albeit without being able to phenotype that individual with the limited SNP data. (Lin 2004) It is thus paramount that projects are defined as clearly as possible in advance, and that general explanations of the nature of genomic data projects as sketched above are incorporated into informed consent. It is also paramount that any filtering, scrambling, encryption or de-identification techniques be thoroughly vetted for strength, for it is reasonable to assume that if we can protect information, techniques capable of breaking those protections can be developed. (Erlich 2014)

<u>The regulatory hurdles to a global genomics database are prohibitively complex.</u>

The establishment of a global open source genomics database necessitates the creation of an international regulatory framework for genomics research similar to those that already exist for trade agreements and other international treaties, and will require a complex and technological cooperation among health agencies worldwide. If such a database were to be created, it would require the formation of an international consortium of health, genomics, government and data specialists from each member organization or nation so that issues surrounding information exchange, standardization procedures and policy can be thoroughly debated and agreed upon.

Moreover, research groups across the world exist within different regulatory environments with their own rules pertaining to information exchange and dissemination. In order for global genomics to work, researchers and healthcare agencies involved in data storage and transmission will have to work together to standardize their practices. This kind of international cooperation is seen in global trade agreements and peace treaties, but would be novel for publically funded international scientific and clinical research. Furthermore, global genomics could warrant changes to regulatory law in any home country of a research institution desiring to participate in big genomics projects, possibly necessitating years of legislative work focused on harmonizing legal, scientific and ethical issues internationally.

Party A's Position and Rationale for Why Party B's Position is More Compelling

Party A's position is that a global, open-source genomics database with automatic data import from research institutions is a live-saving endeavor. While it is true that such an endeavor will advance our understanding of genomics, it is of paramount importance that we do not damage rights to privacy or control of personal information in the process of creating such a database. Moreover, we must not place an undue amount of strain on governments and regulatory agencies across the world. In theory, global genomics projects promise to glean massive discoveries about health and disease, but at what cost? If patient data is automatically imported into the database from participating research sites, all patients at those sites must be informed of this activity. De-identification and data masking are not foolproof and it is likely that such a high profile database will be the target of hackers and other malicious parties that have targeted, for example, financial and communications databases in recent years. Genomics discoveries are important to the advancement of medicine, however we have been making those discoveries without automatic

databases and *de facto* data collection. Moreover, global and national genomics consortia have been established through alternate channels, and they are producing results that are changing the way medicine is practiced. The scientific desire to standardize and unify genomics research efforts across the globe is academically ideal, however academic ideals do not surpass our local commitment to protecting the rights of patients. Current genomics programs already allow for translational applications of genomic knowledge, and there is dialectic reasoning for shifting focus from these programs to large-scale projects with inordinately complex regulatory dynamics. Lastly, closed genomics databases may generally protect the personal and genomic information of patients better than open source formats by because they are independently maintained and less vulnerable to breaching. (Erlich 2014)

References

Collins S et al. (2003) A vision for the future of genomics research. Nature 422, 835-847.

Erlich Y and Narayanan A (2014) Routes for breaching and protecting genetic privacy. Nature Reviews Genetics 15 409-421.

European Society of Human Genetics 2003) Data storage and DNA banking for biomedical research: technical, social and ethical issues. Eur J Hum Genet 11: 906–908

Greenbaum D, Sboner A, Mu XJ, Gerstein M (2011) Genomics and Privacy: Implications of the New Reality of Closed Data for the Field. PLoS Comput Biol 7(12): e1002278.

Human Genome Organization (1996) Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing. Singapore: Human Genome Organization

Human Genome Organization (2003) Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. Singapore: Human Genome Organization

Knoppers BM (2010) Consent to 'personal' genomics and privacy. EMBO Reports 11, 416-419.

Lin Z, Owen AB, Altman RB (2004) Genomic research and human subject privacy. Science 305: 183.

McGuire AL et al (2007) The future of personal genomics. Science 317(5845):1687.

The Organisation for Economic Co-operation and Development (2009) Guidelines on Human Biobanks and Genetic Research Databases. Paris, France: Organization for Economic Co-operation and Development

Personal Genome Project (2016) Participation is non-anonymous. http://www.personalgenomes.org/organization/non-anonymous. Accessed April 19, 2016.

United Nations Educational, Scientific, and Cultural Organization (2003) International Declaration on Human Genetic Data, 16 October 2003. Paris, France: United Nations Educational, Scientific and Cultural Organization.