

Hannah Ma & Katie Fifer
cs199r Final Project
May 14, 2007

Privacy Implications of Personal Genomics

Introduction

With the completion of the Human Genome Project, the cost of sequencing DNA has dramatically decreased and comprehensive genetic screening will soon be affordable for the average citizen. Personal genomics refers to the development of affordable procedures to determine “personal genome sequences” and user-friendly applications for analysis of those sequences. Some researchers are aiming to sequence an individual’s genome, or at least the frequently varying parts, for just \$1000 within the next few years (Wade 2006). The ability to inexpensively sequence millions of partial personal genomes could revolutionize both medicine and research, making “personalized medicine” a reality and dramatically increasing the pace of future genetics research. However, such progress is not free of societal concerns. Here we consider the implications for individual privacy, which are both difficult to predict and of the utmost importance to address.

We investigate three of the most salient and interesting aspects of genetic privacy - the anonymity of genetic information, the shared nature of genetic information, and the implementation of protection of genetic information. While there are also a number of social and legal implications - such as genetic discrimination in employment and in insurance - those remain outside the scope of our privacy-oriented project. Many of the issues we discuss remain unresolved and their impact is difficult to anticipate. However, we believe it is critical that leaders in biology, computer science, and public policy make every effort to protect individuals’ privacy before the coming explosive growth of personal genomics.

Background

The genetic makeup of each person is constructed from DNA, a heritable material that dictates an individual’s physical characteristics and traits. Sequences of DNA are subdivided into genes, which each act as the blueprint for the construction of one functional product, usually a protein. A particular DNA sequence is some combination of four subcomponents, or bases (Pearson 2006). The precise arrangement, or sequence, of these bases determines exactly what phenotype, or characteristics, a person will express. If this blueprint is altered, and an incorrect protein or an incorrect amount of a particular protein is made, the result may be an expression of various diseases (“Human Gene Testing” 1996).

Our understanding of the interplay of the thousands of genes making up the average human has seen tremendous advancement over the last half-century. Some diseases can be caused by a change in just one base in an individual’s genome (currently estimated to

be about 6 billion bases long); for instance, sickle cell anemia manifests itself when a person has two copies of a mutated gene (one inherited from the mother, one from the father) responsible for producing hemoglobin, a protein for carrying oxygen in the blood (IHGSC 2004). Other diseases can be caused by sections of repeated DNA (Huntington's) or deleted DNA (Turner's syndrome), and still others are controlled by a whole host of complex gene interaction.

Much of the current focus in genetics research has been on learning more about the sequences that seem to cause or predispose a person to disease. Within the human population there is not a tremendous amount of genetic variation, but scientists have learned that certain locations in the genome vary with much higher frequency than others, and these variations have been implicated in many diseases. There are approximately 10 million sites of more frequent point mutations, or changed nucleotides, commonly known as single-nucleotide polymorphisms (SNPs). Because many genes interact with each other, a mutation in a particular gene may not be a binary determinant of whether or not an individual will have a certain disease, but instead indicates the increased *likelihood* of developing that disease. It will be a long time before scientists fully understand all of the complex interactions of various genes, but they are already able to estimate a reasonable likelihood of disease from the presence or absence of certain SNPs or other mutations, and their understanding is improving at a rapid pace.

Major research efforts have focused on the application of genetic information to medical care, and to some extent it already has been: genetic screening is available for a number of diseases, including cystic fibrosis, Huntington's, and Duchenne Muscular Dystrophy. Many scientists believe that if people are able to sequence their own genomes, medical care will become far more personalized, because an individual's DNA sequence reveals so much more than do his outwardly apparent personal characteristics. Personal genomic information will provide direction for the advanced screening for disease, selecting more effective medications and dosages, and creating better vaccines, which can be tailored to individual needs (Church).

Until now, however, one of the limiting factors for genetic research and its medical applications has been the very small number of available human DNA sequences. The Human Genome Project was finished in 2003, culminating a 13-year international research effort. Researchers sequenced a large percentage of the human genome, but there remain spots that have proven difficult to sequence accurately due to repetitiveness and other ambiguities. Furthermore, this was the result of a multi-million dollar research effort; while a significant scientific accomplishment, it clearly is not a direct model for the sequencing of individuals' genomes ("Human Genome Project Information").

Consequently, the focus of some genomics scientists has been on reducing the cost of sequencing and on determining certain relevant parts of the genome, such as SNPs, to sequence. Personal genome project costs are dropping rapidly. The current cost per subject ranges from \$8,000 to sequence a limited subset of the genome to more than \$200K for a large portion of an individual's DNA. However, some researchers anticipate that the cost of haplotype sequencing ("hot spot" sequencing) will drop to \$1000 per

person within one or two years, as a multitude of corporations, such as Solexa, 454 Life Sciences, and Applied Biosystems, compete to create the cheapest sequencing machines (Wade 2006). These researchers are confident that, consequently, thousands or potentially millions of individuals will be able to have their genomes, or at least significant parts of their genomes, sequenced (Church).

While personal genomics holds great promise for medicine, it also presents unique privacy concerns; “low-cost decoding may bring the genomic age to the doctor’s office, but it will also raise quandaries about how to safeguard and interpret such a wealth of delicate and far-reaching personal information” (Wade 2006). Because scientists understand such a small fraction of genetic sequence information now, it is very challenging to predict the best ways to protect individuals’ information for the future. We hope to predict some of the most important features of what is already becoming a significant privacy-genetics debate and to make suggestions for a model of personal information control.

Anonymity

As our ability to share, store, and aggregate data continues to improve, the ability to keep certain data anonymous becomes increasingly important. Because an individual’s genetic information is so personal and so specific, it seems absolutely critical to be able to protect it from unwarranted access or use. However, it may be impossible to actually make genetic information anonymous, without losing its significance.

Genetic data presents truly unique challenges. Genetic information is unlike other personal information such as fingerprints, in that fingerprints must be linked with other information to be meaningful. However, an individual’s genetic information says everything about his physical characteristics, so it may not need to be linked with anything else to be incredibly revealing. And unlike other data such as a name or social security number, someone cannot change his or her DNA (at least at this point).

It is easy to be skeptical of just how revealing genetic information might become. Currently, scientists only know how a few hundred SNPs manifest themselves as physical traits. However, that is changing rapidly. Even if an individual’s genome could not be linked to his person today, it is quite feasible that science will advance rapidly enough so that within his lifetime it could be. Therefore, the decisions made now about releasing or protecting information may have significant impact that is hard to predict now, and it seems best to be proactive about protecting people’s information.

Perhaps more immediately, if a particular sequence could be linked with some seemingly trivial data – perhaps a zip code, or more generally, a region, it could be easy to link people with their genetic information. In some sense, this seems hypothetical, but in light of other anonymity studies and considering the amount of information aggregated on individuals by firms like ChoicePoint, it does not seem unrealistic. In a famous study of 1990 census data, researchers found that they could identify 87% of the US population solely based on gender, zip code, and full date of birth (Sweeney 2000). Given that

genetic information can indicate things as specific as disease, eye color, age range (based on the length of some repeating sequences called telomeres), and an enormous number of other traits, it might be necessary to link it with very little other information to be able to pinpoint an individual.

This provides a significant challenge both from a technical standpoint and from a research perspective. From a technical perspective, making such data anonymous seems nearly impossible. Many techniques used to make data anonymous, such as perturbation of the data, would significantly diminish the quality of the information, and impede scientists' attempts to learn from it ("Why Your Anonymous Data Isn't"). While we do not have a technical solution to this problem, we hope to see consideration given to the matter from experts in the field. The best means of protecting this data seems to be through some sort of access control, such as public key encryption.

However, access control will not solve some fundamental problems, most notably in research. Currently, geneticists have recognized the importance of privacy, and most studies involving subjects' genetic information operate on a double-blind policy, which prevent the researchers from being able to link back to the subjects and vice versa. However, researchers often publish their findings, and others usually review their datasets. If this data cannot be made anonymous, no matter what double-blind precautions are taken, maintaining subjects' privacy may not be possible. The issue is that there may simply be too much uniquely identifying information in a sequence. We have already seen a similar example involving even less unique information, in AOL's release of search histories, and people became outraged at how easy it was for an amateur to identify the person associated with each history (Kawamoto 2006).

Simply linking a published sequence to an individual may seem like it would not be very significant, since that link was created based on knowledge of the sequence and of that person's physical characteristics alone. In a sense, it seems like no extra information could be gleaned. However, it is possible to imagine linking a sequence to a person based on genes A-D, but then realizing he also carried gene E, something he wanted to remain private, perhaps a gene indicating he would develop some disease. This is a very serious concern, and something that researchers ought to consider in designing their studies. Patient or subject education is of the utmost importance, so that they are fully aware of the privacy risks associated with participating in a particular study. Furthermore, researchers should publish only the most relevant information, as extraneous sequence information may help people to identify the associated person.

While some of this fear may seem far-fetched, it is certainly best to be prepared for the potential consequences of failing to protect people's privacy – as science continues to advance and more genes are understood, individuals' sequence information will become increasingly identifiable. And since the information we publish today really never goes away, it is imperative that we take proactive steps to ensuring a reasonable amount of privacy for individuals starting now. At the same time, this provides an enormous challenge for researchers, and efforts need to be made to see that the valuable discoveries that could be made are not stifled by privacy policies.

Family ties and genetic information

Today some of the most sacrosanct models of health care provision in the United States are changing. The traditional physician-patient relationship is witnessing an increasing fragmentation of care across providers, more geographic dispersion of families, and rising independent patient access to medical information. The idea of the sole primary care provider, who continually oversees the health of an entire family, is becoming as antiquated as its image. In light of these changes, personal medical data - already conceived of as "private" for adult individuals - has become increasingly seen as "privately owned" (in a conceptual, not a legal sense). This can be seen in the clamor for personal medical records, by which patients can easily view diagnoses, transport them to different physicians, and add footnotes or critiques; underlying this organizational methodology is the belief that medical data belongs to the patient.

The rise of personal genomics challenges this trend toward personal ownership of and responsibility toward medical data, an idea itself quite new, as genetic information blurs the lines of what is individual. Genetic information is connected among blood relatives, with each individual sharing typically one-half of his DNA with each parent and sibling. For certain diseases, knowledge of the genetic makeup of an individual may have profound consequences for his children - and may also bring to light previously unknown, perhaps purposely ignored, health issues with his parents and siblings. In light of the genetic interconnections within families, we need to reconsider the current conception of individual medical information.

2.1. New challenges in medical privacy

The conflicts with the sharing of genetic information can be categorized into two broad aspects:

1. The right to know: Who has the right to access a patient's genetic information? Does the patient have the obligation to share results of genetic testing with his relatives?

A certain public ambivalence exists on this issue. In a survey of 200 women, one study found that 100% of the subjects believed that a patient should notify his family members of an easily preventable disease (and 85% held that a patient should tell relatives about a nonpreventable one); however, when asked whether the physician should notify the family, if the patient refused to disclose the information himself, those numbers dropped to 18% and 16%, respectively (Lehmann et al. 2000). Currently, the legal guidelines regarding the latter situation are unclear; most existing guidelines come from the interpretation of state or trial court precedent not specifically written for genetic information. Arguments for the forced disclosure of genetic information to family members have often cited analogous court cases, which typically follow these two lines of argument (Burnett 1999):

On the need to notify third-parties of risk, *Tarasoff v. Regents of California* (1976): the California State Court found a psychotherapist responsible for not notifying law

enforcement of a patient's threats against a third party, whom he later killed. The court held the physician liable for his "failure to warn," and found the criteria necessitating a warning to encompass four parts: whether the patient is engaged in dangerous conduct, whether the physician foresees harm, whether a special relationship is present, and whether the party at risk is identifiable (Burnett 1999). The latter two cases have been interpreted to hold in later cases regarding genetic information - the physician's relationship to his patient, and not necessarily to the third-party, has been deemed sufficient as a "special relationship"; relatives can be defined as the at-risk party - but the first two criteria remain questionable, as they depend upon the amount of genetic determinism known or associated with a mutation.

On the need to prevent or better treat "contagious" disease, *Wojcik v. Aluminum Co. of America* (1959): a New York trial court found a physician responsible for not notifying the wife of a man diagnosed with tuberculosis, preventing her from taking steps to either avoid or treat the disease. While heritable diseases are not "contagious," early detection may still help treat them; hence it can be argued that a physician who fails to notify a relative should be held liable, again for the failure to warn of possible disease.

Consequent genetics-specific cases, or their final state Superior or Supreme Court ruling (*Pate v. Threlkel*; *Safer v. Estate of Pack*), have found that privacy rights (in the legal sense) do not transcend the duty of physicians to warn family members of genetic results that could affect their well-being, given a failure to warn. However, concerns have come from the related issue of confidentiality, whereby physicians are (mostly) legally prohibited from sharing a patient's medical information without his permission. We should keep in mind that these cases have not established "a clear legal duty on the part of the physician to disclose to family members confidential information about genetic test results" (Doukas and Berg 2001). The tension between individual confidentiality and family health results in deadlock without a different conception of the relationship between patients and practitioners, which we will discuss later.

2. The right to share: Who has the right to share a patient's genetic information? Does a patient have the right to freely share his information, given that it may be linked to his relatives'?

The right to share information involves the collision of two other rights: the right of the patient to communicate his information, and the right of his relative to privacy. When a patient reveals information about his genetic makeup, he inherently suggests that, with some probability, his relatives share that data. If the patient publicizes news of a genetically-based medical condition, his actions may have serious consequences for his relatives, psychologically and socioeconomically. For instance, although genetic discrimination is federally prohibited, public knowledge of a person's relative's condition may lead to unwanted attention or subconscious prejudice; regardless, exposing the information violates the relative's right to medical privacy (in layman, not legal terms). This violation is clearer in cases such as Huntington's Disease, where a child's positive test guarantees that at least one of his parents has the same condition, even if they do not yet show symptoms. However, we believe that as predictive technologies become more

sophisticated and widely used, probabilistic information about the likelihood of expressing particular traits or diseases - ie the fact that an individual has a heightened risk for breast cancer - should constitute private medical information.

Some have doubted that much information about a relative could be gleaned from someone's DNA sequence. However, familial linkages are very much possible, and already they have been used in law enforcement. Genetic databases store the DNA information of convicted felons in the United States, and across most developed countries. The databases have been used to not only test repeat criminals for their involvement in a new crime, but to also identify their relatives as potential suspects. In one instance, the DNA found at a murder in North Carolina nearly matched a convicted criminal's; investigators used that evidence to test his brother, who had an exact match and confessed to the crime (Bieber et al. 2006). Hence genetic data collected about convicted criminals (and, in a troubling trend, toward those merely arrested) has been used to gather information about private citizens.

Generalizing from this example, we recognize that allowing individuals to freely share their genetic information does violate the privacy of their relatives (again, in layman, not legal terms). We believe, however, that we cannot legislate or interpret the law otherwise. Unlike law enforcement's use of the genetic databases assembled around convicted criminals (in which they "share" individuals' genetic information for a singular purpose, in the search for a criminal), a patient may desire to freely share the results of genetic testing for many reasons. Hence it will be far more difficult, if not impossible, to come up with effective guidelines for when it is appropriate to share genetic information. Furthermore, a patient's right to free speech - to be able to communicate with their practitioner, as well as to choose to share their condition - is also a critical right.

Consequently, in both of these cases, we choose to remain within the existing legal frameworks. At this point, there are too many unknowns about the true implications of genetic information - and too much vagueness in how it may be used - to provide effective and clear laws. Instead, we believe that there are tools that can help many families voluntarily discuss the effects of personal genomics on all members and preempt many of the challenges of genetic testing.

2.2 The family covenant model

In light of the growing sophistication in genetically-based medical information and diagnoses, the individual patient to practitioner model has grown outdated. It is necessary not to force the results of technological change into an existing and increasingly meaningless model, but to adapt the model to accommodate the technology of personal genomics.

The family covenant model provides a simple but effective framework for thinking about genetic testing in a familial context. Developed by David Doukas in 1991, it works as an agreement among family members and between them and a primary practitioner in regards to difficult medical situations. The family covenant has four main tenets:

- "1. the family is the 'unit of care';

2. the physician is charged with comprehensive family health;
3. individuals in the family are treated within the context of the family;
4. family-based medicine realizes the importance of the biopsychosocial model of medical care" (Doukas 2001).

2.2.1 The family covenant and genetic testing

In the context of genetic testing, the family covenant enables families to discuss and negotiate how the results would be communicated and utilized prior to the test itself. However, the model is more ambitious than a mere one-time agreement; the covenant is conceived as a long-term, flexible document that anticipates potential disagreements, lays out the mechanisms for settling disputes, and accumulates "trust" as different issues are resolved. In Doukas' formulation, the covenant would be administered and mediated by a (shared) family physician. Conceptually, it strives to represent the family as the unit in genetic testing, stressing mutual education such that individual patients are highly aware of the effects upon family members.

The family covenant is especially helpful in defining the questions bundled under the relatives' "right to know" category. With its emphasis on pre-testing negotiation, it helps lay out the extent to which each family member is willing to give up his individual confidentiality; the definition of potential "harm" to relatives that would compel the sharing of medical information; the ways and criteria by which disagreements will be resolved. Finally, by its very nature, it increases the time spent on patient education and the likelihood that they will make informed medical decisions.

2.2.1 Challenges for the model

While Doukas focuses mostly on conceptual issues with his model (for instance, what does a covenant mean when withdrawal is penalty-less?), we would like to discuss some challenges with its real-world application. In particular, it makes assumptions in the following categories:

1. Family dynamics: the family covenant relies on a curiously old-fashioned view of family health care, where families are geographically united and use the same physician, though better communication technologies can certainly simplify those issues. Furthermore, it seems likely that tensions over genetic testing are especially likely to erupt in families with testy or distant relationships.
2. Patient consent: the family covenant provides no mechanism for effectively managing fundamental patient nonconsent, either throughout the process or immediately surrounding a genetic test. (Doukas notes that the patient merely needs to leave the covenant, if he ever changes his mind.)
3. Practitioner resources: primary-care physicians have typically not been trained as genetic counselors and under this model, they are now being asked to operate partly as such, and also to work as family mediators and therapists. These duties require a broader skill set than many physicians have; it also remains questionable whether physicians should have such responsibilities, even if they possess the skills to fulfill them.

The first two issues reflect the inherent weaknesses in a model based on voluntary cooperation. While we wanted to point them out, we do not believe that they diminish the value of family covenants, which can still help many families better conceptualize their shared health information and issues. The third is of more critical importance and should require reflection on the best resource and work allocation. One suggestion has been to educate more counseling professionals with basic genetic information, such that they can serve effectively as mediators in family covenants (Lapham 2001); we believe that this may be more efficient than training physicians counseling skills and diverting them from more traditional responsibilities.

Finally, we need technology to help implement family covenants in a meaningful manner, as families become increasingly less likely to share a doctor's office or bring their own information to only one practitioner. We will discuss one such method below.

Granular access: a model for personal records

At this time, the implications of genetic testing and the significance of certain mutations are yet unknown. The importance of safely storing and managing that data among research subjects has previously been discussed. We believe that that information must also be protected for individuals who are regular patients of the hospital system, particularly given the unknown meaning or consequences of much of that information. We will discuss these limitations in light of personal medical records, which enable a technological solution to these concerns.

Within a comprehensive personal medical record, we believe that patients should have the right to limit genetic information and the results of genetic tests from the rest of their medical records. Individual physicians should be separately and individually granted permission to access genetic information (information that may be critical in an emergency situation could be doubly recorded in the general record). Furthermore, as genetic information becomes cheaper - when more comprehensive personal sequencing becomes common - physicians should not automatically be given access to the entirety of a patient's genetic information. This is because of the heightened consequences of any privacy breaches in patients' genetic data; an extraneous physical or computer copy of a person's genetic sequences may reveal information that could greatly affect the patient in the future. Such a breach may reveal an enormous amount of information, not only about a patient's medical history but with some reference to their medical future. Consequently, we need to think about how to organize and present genetic information to best protect patient privacy.

It seems that a particularly important feature of any system storing genetic information will be granular access control. Each individual has about 6 billion bases of DNA. While it is probably easiest to pass this entire chunk of data to any doctor asking about patient history, it's quite conceivable that individuals would not be best off releasing all of their data. This is especially true given how frequently data is misplaced or stolen. The challenge lies in allowing patients to filter his information in a meaningful way, enabling release of selected portions of his sequence to various people. It seems that

technical solutions should be developed to help the patient filter his information, but this solution should go hand-in-hand with an aggressive educational outreach to patients. Granular access within personal medical records will also be critical in implementing family covenants, allowing selective sharing of information within the family unit. Because patients are usually invested in their own medical care, and already individuals read large amounts of information about their ailments online, it seems feasible and logical to implement such a system - one that relies on patient involvement and education.

Conclusion

In this brief, we sought to highlight some of the most pressing issues resulting from the impending advancements in personal genomics. We also offered some possible solutions or helpful frameworks, to start the conversation on how to incorporate genetic information into our current conception of healthcare. They should by no means be considered a final word or best practice. Ultimately, we believe that education of patients and of experts in a range of fields is absolutely critical as people begin to sequence their DNA. Personal genomics holds great promise for both research and medicine. We must do our best to devise responsible guidelines and laws that will protect individuals' privacy without stifling scientific research and advancements.

Sources

- Bieber, F.R., Brenner, C.H., and David Lazer. "Finding Criminals Through DNA of Their Relatives." *Science* 312, 1315-6 (2006).
- Burnett, Jeffrey. "A Physician's Duty to Warn a Patient's Relatives of a Patient's Genetically Inheritable Disease." *Houston Law Review*. 36 (1999): 559-582.
- Church, G. "Personal Genomics Will Arrive This Year, and With It a Revolutionary Wave of Volunteerism and Self-Knowledge."
http://edge.org/q2007/q07_7.html#church
- Council for Responsible Genetics. <http://www.gene-watch.org/>
- Doukas, David and Jessica Berg. "The Family Covenant and Genetic Testing." *The American Journal of Bioethics*. 3 (2001): 2-10.
- Genomes Online Database. <http://www.genomesonline.org/>
- Golle, Philippe. "Revisiting the Uniqueness of Simple Demographics in the US Population." Palo Alto Research Center.
<http://crypto.stanford.edu/~pgolle/papers/census.pdf>.
- "Human Gene Testing." *Beyond Discovery*. National Academy of Sciences.
<http://www.beyonddiscovery.org/content/view.article.asp?a=239> (1996).
- "Human Genome Project Information." Genomics.energy.gov.
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
- International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome." *Nature* 431 (7011): 931-45 (2004).
- Kawamoto, Dawn and Elinor Mills. "AOL apologizes for release of user search data."
CNet News.com. http://news.com.com/2100-1030_3-6102793.html
- Lapham, E. Virginia. "Family Covenants and Genetic Testing: Utilizing the Skills of Counseling Professionals in Implementing Family Covenants." *The American Journal of Bioethics*. 3 (2001): 1-2.
- Lehmann, Lisa, Jane Weeks, Neil Klar, Lois Biener, and Judy Garber. "Disclosure of Familial Genetic Information: Perceptions of the Duty to Inform." *American Journal of Medicine*. 109 (2000): 705-711.

Pearson H. "Genetics: what is a gene?". *Nature* 441 (7092): 398-401 (2006).

Personal Genome Project (PGP). <http://arep.med.harvard.edu/PGP/>

"SNPs: Variations on a Theme." National Center for Biotechnology Information (NCBI).
<http://www.ncbi.nlm.nih.gov/About/primer/snps.html>.

Sweeney, L. "Uniqueness of Simple Demographics in the U.S. Population." LIDAPWP4.
Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh,
PA, (2000).

Wade, Nicholas. "The Quest for the \$1000 genome." *The New York Times*. 18 July 2006.

"Why Your Anonymous Data Isn't." *Wired Blog Network*.
http://blog.wired.com/27bstroke6/2006/10/why_your_anonym.html