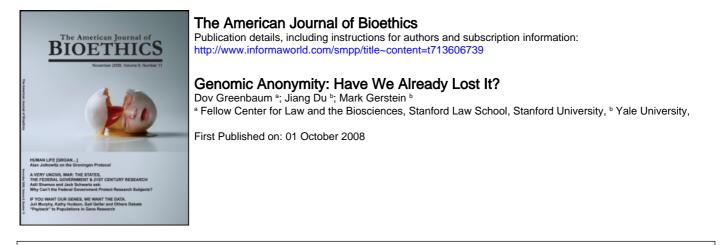
This article was downloaded by: [Yale Univ Library] On: 13 December 2008 Access details: Access Details: [subscription number 788733624] Publisher Routledge Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



To cite this Article Greenbaum, Dov, Du, Jiang and Gerstein, Mark(2008)'Genomic Anonymity: Have We Already Lost It?', The American Journal of Bioethics, 8:10,71 — 74

To link to this Article: DOI: 10.1080/15265160802478560

URL: http://dx.doi.org/10.1080/15265160802478560

## PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

**Open Peer Commentaries** 

# Genomic Anonymity: Have We Already Lost It?

Dov Greenbaum, Fellow Center for Law and the Biosciences, Stanford Law School, Stanford University Jiang Du, Yale University Mark Gerstein, Yale University

Hull and colleagues (2008) discuss the utility of the current regulatory distinction between identifiable and nonidentifiable genomic information, particularly given the seemingly anomalous preferences of their surveyed patient population. As the authors note, this regulatory distinction will become even less meaningful with the proliferation of genomic databases. Particularly as industries such as personal genomics expand — flooding both private and public databases with readily identifiable genomic data - they will effectively prevent an ever-growing number of individuals from remaining genetically anonymous (Lowrance and Collins 2007). In fact, recent research has already shown that individual genomes can be readily identified out of larger mixed groups of publicly available data from genome wide association studies using only a small subset of one's genome (Homer et al. 2008). Once it's known that a person has participated in a genome wide association study, it becomes fairly straightforward to use their or their relative's genomic data — which may well be made available through personal genomics - to re-identify that individual (National Institutes of Health [NIH] 2008).

The general expanse of genomics into our medical system, both through personal genomics and also through other evolving biomedical technologies such as targeted personalized medicine, also raises other non-trivial privacy concerns both for the patient herself but also for her extended family that share much of her genomic complement.

#### THE TECHNOLOGY

The sequencing of the entire human genome was a triumphant coda to the innumerable successes and discoveries of twentieth-century science. And like many of those publicly funded discoveries, genomics has been hastily transformed into the consumer technology, personal genomics. In contrast to the relatively established single-gene testing industry where physicians confronted with a patient's statistical probability of developing or passing on a genetic disorder will send their patient to be tested for that particular condition, personal genomics fundamentally refers to the direct to consumer, data driven, large-scale sequencing, deciphering and open exploration of individual genomes.

The underlying science and technology of personal genomics is the result of a confluence of a number of biotechnological and computational successes. Nobel Prize-winning DNA sequencing technologies gave way to the human genome project that gave us a representative sample of the entire human DNA sequence. To add value to the raw sequence data, scientists have been analyzing and annotating the genome in an effort to catalogue and determine the function, localization, shape and nature of interactions of not only the nearly 25,000 genes and their macromolecular products coded for by our DNA, but also the approximately 15 million sequence variations (around 0.5% of the genome) between individuals, including single nucleotide polymorphisms (SNPs) and copy number variants (Korbel et al. 2008).<sup>1</sup> These polymorphisms — in both coding and non-coding regions of DNA - frequently correlate with genetic diseases, health conditions, or physical characteristics.

Rising computational power with concomitant plunging costs in digital storage and computational speed, coupled with dramatic expansions in sequencing abilities and breakthrough high throughput experimental techniques have helped to turn these successes in DNA sequencing

<sup>1.</sup> Companies include 23andMe (Mountain View, CA), DecodeMe (Reyjkavik, Iceland), Knome (Cambridge, MA), Navigenics (Redwood Shores, CA), Gene Partner (Zurich, Switzerland) and Scientific Match (Naples, FL) or freely available through the Coriell Personalized Medicine Collaborative (Camden, NJ).

Address correspondence to Dov Greenbaum, Fellow Center for Law and the Biosciences, Stanford Law School, Stanford University, Crown Quadrangle, 559 Nathan Abbott Way, Stanford CA 94305-8610. E-mail: dov.greenbaum@aya.yale.edu

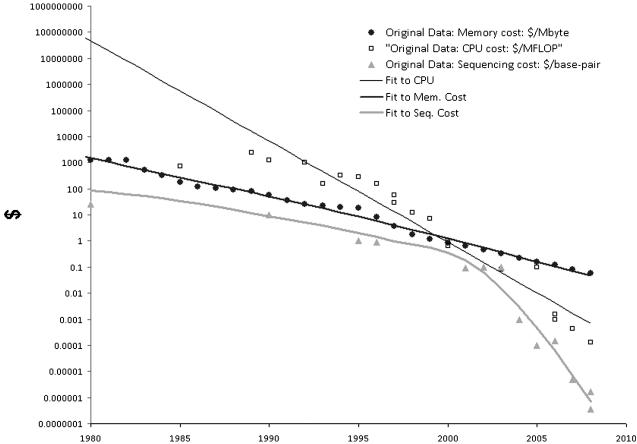


Figure 1. The graph shows how both computational and sequencing costs are pushing toward an affordable and marketable consumer genomics industry. In particular, the falling costs of computer processing unit cycles, data storage, and the plummeting costs of sequencing are shown. The raw data was compiled from a broad range of sources: 1) \$/Mbyte or megabyte reflects falling computer storage costs (Kurzweil 2005).

2) \$/MFLOPS is an acronym for million floating point operations per second, and is a measure of computing performance (data compiled from FLOPS: Cost of computing 2008; Jarvis 2008; Gordon Bell Prize Winners).

3) \$/Base-pair charts the falling cost of sequencing DNA component nucleotide base-pairs. Sequencing cost data from Kurzweil 2005; Cleveland and Devlin 1968.

The raw data points were also fit to curves to better illustrate these falling costs. The fit was done by Locally Weighted Regression (LOESS) regression, with first-order polynomial (linear function) as the local model for the MFLOP and Dynamic Random Access Memory (DRAM) data, and second-order polynomial for the sequencing data.

and genetic variation analyses into a viable, commercializable technology. In fact, sequencing technology has been advancing at a rate even faster than Moore's law — the historic exponential increase in capacity and speed of computer devices (Figure 1). Given the current rate of scientific advancement in genomics, costs for personal genomic screenings will continue to fall precipitously making the technology accessible to an ever-widening audience.

#### PRIVACY CONCERNS

Personal genomics companies provide services ranging from the cataloguing the hundreds of thousands of discrete DNA sequence variations to providing you with your entire diploid 6 billion base-pair genetic sequence, to suggesting potential mates based on your genomic complement.<sup>1</sup> Many also intend to maintain extensive genomic records that can then be used for valuable genomic research. Most personal genomics companies, for policy reasons, describe themselves as recreational services that are not intended to be used for medical purposes. Not only does this designation limit United States Food and Drug Administration oversight of the businesses, but it also allows these services to avoid the supervision of institutional review boards and potentially skirt other federal regulatory safeguards typically enforced when collecting human subject samples for scientific research.

Thus, in contrast to medical records that are traded almost exclusively among authorized doctors, personal genomics will allow equally if not more revealing information to be viewed, traded, and potentially even data-mined, in the online bazaar — with little to no oversight. Personal genomics takes the management over heretofore restricted medical data away from medical professionals, transferring it to the patients themselves, giving consumers unprecedented responsibility and control over their own genomic and medical information.

Genetic information is unique to each individual and unintended disclosure of even a small number of SNPs or other highly variable regions can be used as a reference sample to identify a previously anonymized but publicly available genetic research sample. Revelations of genetic predispositions or disease markers could potentially bring substantial financial harm or social stigmatization not only to the particular individual but also her unsuspecting but genetically similar family members.

Similar to the devastating erosion of online privacy where effectively indelible web pages disclose personal information, confidential emails are rapidly and widely circulated, and surfers unwittingly drop revealing digital bread crumbs, personal genomics undercuts privacy to a new degree. Thus, like Hull and colleagues (2008) has shown with regard to their sample, the current and expected success of the industry is at odds with the regulatory conventional wisdom regarding the public's inhibitions in sharing genomic information.

And, like many users on social networking sites, consumers may not realize how much of their privacy is compromised. But, unlike many of the web 2.0 neophytes who casually and cavalierly post their entire lives online, personal genomics will not only have privacy repercussions for the consumer, but also for any of his relatives; an individual's genome reveals half of the genome of his parents and children and a substantial fraction of his sibling's. Just like posting a picture on MySpace or Flickr can reveal a lot about you and those in the frame with you, when someone shares his genotype, by choice or otherwise, he is exposing substantial private information about himself and his close relatives.

As more people sign on to personal genomics, the remaining unaffected population will rapidly shrink. Consider, for example how email privacy would become negligible if most people made their emails publicly accessible. Even those who would choose to continue to assert some privacy over their email would be unable to maintain that privacy given the high probability that those receiving the emails would put no effort into securing the emails and keeping them private. So too, as more people make their genomic information public, it becomes more likely that someone attempting to preserve genomic privacy will have their genomic information nonetheless effectively revealed through the actions of a family member or close relative. In the extreme, if 90% of the population has their genome sequence the sequence of the remaining ten percent is all but determined.

#### **REGULATION?**

Although we do not currently understand even a subset of the genetic influences on our lives, eventually we will; but by that time it will be too late to retract the genomic data that many of us imprudently uploaded.

Naively one might assume that recent federal legislation such as the Genetic Information Non Discrimination Act (GINA) (H.R. 493), designed to harmonize what was until now a patchwork of state and local laws regarding genetic discrimination, would protect consumers from these privacy concerns. But while insurance companies and employers are prohibited by GINA to ask for genetic information, they are allowed to access freely available information, the type that is produced by personal genomics companies and shared by their consumers, and will likely be collected and indexed by enterprising marketing firms. Further, health insurers and employers are only a small subset of people that can discriminate based on genetics. Life and disability insurance providers, for example, are not included in the current legislation. GINA only limits discrimination, but one can imagine that personal genetic information can be used for a host of other purposes, from unauthorized scientific research to selective dating or just general voyeurism. Thus, a loss of genomic privacy does not only result in the obvious financial or social harms. Invasion of privacy exposes many of our deep-seated secrets or yet unknown genetic frailties to the world; the exposure itself can lead to humiliation and shame. Although the effects, if any, of disclosure may not be immediate given the relative paucity of strong correlative data between genes and disease, nevertheless, as science progresses the descendants of those who shared their genomic data could potentially be substantially affected by their ancestors actions.

### WHO SHOULD REGULATE THE INDUSTRY?

At this juncture there are two possibilities, independent self regulation by the industry or overburdening government regulation. This article suggests the former, as the latter is likely to significantly hamper the industry's ability to innovate and produce indispensable data.

Notwithstanding the possible repercussions to consumers and their relatives in terms of job loss, inability to obtain insurance, or general social stigma that will most likely occur despite the best intentions of Congress to fight genetic discrimination, placing high barriers to acquisition of genomic data through government regulation of the personal genomic industry may chill the use of personal genomics and the concomitant important collection of data for vital research purposes. There is a point at which the complexities of compliance with government regulations effectively serve as a ban on the technology. For risk-averse biotech companies and wary consumers this threshold is often easily met.

And, despite the aforementioned privacy concerns, individuals should be free to share their own genomes, and notwithstanding paternalistic efforts to control the disclosure of genetic information, the government probably does not have a strong enough privacy interest to constrain consumers' free speech. But, without substantial oversight, personal genomic companies might be unable to effectively deal with the varied ethical and moral concerns that might arise, and consumers will belatedly realize the devastating privacy implications for themselves and their families. It is therefore imperative that the personal genomics industry proactively and independently incorporate the tools necessary to protect the privacy of their consumers.

The generic answer to these concerns routinely involves the usage of boilerplate informed consent forms, the ethicist's acknowledgement of the individual's absolute personal autonomy. Typically though, this consent is limited to the acknowledging individual and bounded by the conditions outlined therein, but personal genomics asks the individual to effectively forego complete anonymity and privacy, to extend the reach of the consent beyond themselves to include their family and community members, and to expand it to incorporate information and experimentation not as of yet even imagined. Consequently, the advent of personal genomics also raises issues that could make the current application of informed consent meaningless.

#### CONCLUSIONS

There are no simple solutions. Although, one possible model for the industry might be the financial services industry which has evolved a strong tradition of privacy and protection of information as a competitive strength. As technology pushes forward, substantial privacy issues will continue to rise. And, perhaps, just as the internet changed our perceptions of personal space and privacy, personal genomics will require society to reevaluate our current standards of medical confidentiality and privacy.

#### REFERENCES

Chu, W. L. 2008. Applied Bio sequences a human genome for \$60,000. *MIT Technology Review*, September/October 2006.

Cleveland, W. S. and Devlin, S. J. 1988. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83: 596–610.

Dietrich, F. S. 2007. Sequencing and sequence analysis. Presentation given at Duke University Institute for Genome Sciences & Policy, Genome Academy, September 18, 2007. Available at www.genome.duke.edu/education/genomeacademy/pdf/Fred% 20Dietrich%202007%20Slides.pdf

FLOPS: Cost of Computing. 2008. Available at http://en.wikipedia. org/wiki/FLOPS (accessed September 8, 2008).

Gordon Bell Prize Winners, awarded by the Association for Computing Machinery in conjection with the Institute of Electrical and Electronics Engineers. Available at www.supercomputing.org

Homer, N., Szelinger, S., Redman, M. et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genetics* 44: 167–176.

Hull, S. C., Sharp, R. R., Botkin, J. R. et al. 2008. Patients' views on identifiability of samples and informed consent for genetic research. *American Journal of Bioethics* 8(10): 62–70.

H. R. 493, Genetic Information Nondiscrimination Act of 2008, now Public Law No: 110-233.

Jarvis, J. F. 2008. *Computation history & technology*. Available at http://knology.net/~johnfjarvis/histcompnotes.html

Korbel, J., Kim, P. M., Chen, X. et al. 2008. The current excitement about copy-number variation: How it relates to gene duplications and protein families. *Current Opinion in Structural Biology* 18: 366– 374.

Kurzweil, R. 2005. The singularity is near. New York, NY: Viking.

Lowrance, W., and Collins, F. 2007. Identifiably in genomic research. *Science* 317: 600–602.

National Institutes of Health (NIH). 2008. Modifications to genomewide association studies (GWAS) data access. Available at: http:// grants.nih.gov/grants/gwas/data\_sharing\_policy\_modifications\_ 20080828.pdf (accessed September 29, 2008).

Perkel, J. M. 2006. My own private genome. The Scientist 20(2): 67.

74 ajob