

Research Note

Who Needs ID: DNA-fication in the Modern World

By Malavika Ravi

Copyright © 2015, ASA Institute for Risk & Innovation

Originally Written: December 2014

Edited & Published in ASA Newsletter: April 2015

It is hard to find an industry that does not rely heavily on technology to run the operational, strategic, and tactical aspects of business. The most interesting aspect of information technology, to me, is its potential to revolutionize healthcare. One of my many undergraduate courses was bioinformatics, the intersection of informatics and biology, which focuses on utilizing computer analysis on genomic data. Each organism has a genome, a complete set of DNA, which contains the information needed to build and maintain that organism.¹ In humans, the genome is three billion DNA base pairs long. Scientists sequence and analyze the genomes of humans in order to help identify diseases that are caused by genetic mutations, or variances found when individuals inherit genes from their parents or are exposed to environmental factors. I had the opportunity to mimic scientists and analyze public genomes stored in computer databases using software I created. It opened my eyes to how easy it was for those with no research experience to extract important personally identifiable information. Just like other industries that rely on information technology, there is a catch-22 that is inherent when software and personally identifiable information meet: “new technology [poses] new risks.”² Scientists, lawmakers, and even the public must weigh the potential for these databases to cause undue harm to the DNA donors against the potential for these databases to help the collective. This decision must be made as soon as possible as some scientific communities have begun to reject privacy all together.

In March 2014, the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany committed a serious blunder. One of the foremost research organizations in the world allowed Lars Steinmetz and his team of researchers to publish the genome sequence for one of the most famous cell lines in the world, HeLa, extracted from Henrietta Lacks. The scientific community was outraged. On one hand, this genetic blueprint provides data to the much needed field of genomics. Scientists are able to use advanced tools to

¹ U.S. National Library of Medicine, “What Is a Genome?” *Genetics Home Reference*. 24 Nov. 2014. Accessed Dec. 2014 <<http://ghr.nlm.nih.gov>>.

² Baase, Sara. *A Gift of Fire: Social, Legal, and Ethical Issues for Computing Technology*. Pearson Education Limited. Sep. 2012.

extract information from the genomes that will help diagnose, treat, and sometimes prevent diseases.³ On the other hand, the very tools which are used to benefit humans can be used to identify the participants who are tied to the genomic profile. The EMBL denied that this double edge sword existed claiming that the sequence published did not reveal anything about the Lacks family or even about Henrietta, herself. Yaniv Erlich, known as “the Genome Hacker” believed otherwise tweeting “Nice lie EMBL!”⁴ In fact, a few scientists decided to prove Erlich right. They uploaded the published genome to a site called SNPedia, a Wikipedia site used for translating genetic data. Within minutes, the site produced a report that contained personally identifiable information (PII) about Lacks and her family.⁵ The EMBL were forced to admit their folly; it was possible to garner information about the Lacks family from the genome. They quickly pulled the genome from their public database. However, Erlich believes this move was fruitless. He states, “People don’t realize it’s impossible to hide genetic information once it’s out there.”⁶ Erlich would know, his study published in January 2014 proved exactly that.

The ability to link a genome profile in an anonymous database to a real-world donor was not supposed to be possible. Yaniv Erlich and his team from the Whitehead Institute for Biomedical Research at M.I.T. were able to uncover the names of fifty people whose anonymous genome profiles were available online in a free-access database known as the 1000 Genomes Project.⁷ The 1000 Genomes Project was created to help researchers find genes associated with different human diseases.⁸ The creators of the 1000 Genomes Project never imagined that an ingenious team would develop an algorithm to connect donor profiles to surnames. Erlich and his team created an algorithm which focused on extracting genetic markers from the Y chromosome. The short tandem repeats (STRs) on the Y chromosome (Y-STRs) are passed from father to son with very little mutational change from one generation to the next.⁹ The team used the Y-STR along with the age and location metadata attached to each genome profile in the database to help refine the search of possible DNA matches. They were left with a potential pool of ten thousand men who matched the age metadata and lived in Utah at the time of the DNA donation. The team then used free genealogy sites, Ysearch and SMGF, to connect the Y-STR markers to the surnames. Erlich and his team published their findings to much furor. In response, the 1000 Genomes Project has made small strides to increase donor anonymity such as stripping the age metadata attached to the genome profiles. However, teams inspired by Erlich have begun decoding genomes to detect DNA variants that identify characteristics such as hair, eye color, and facial features. They plan to use this data to filter through public databases that contains these identifiers to reveal the identity of the donors. This study highlights

³ Collins, Francis S. “The HeLa Genome: An Agreement on Privacy and Access,” *The NIH Director*. 7 Aug. 2013. Accessed Dec. 2014 <<http://www.nih.gov>>.

⁴ Hayden, Erika Check. “Privacy Protections: The Genome Hacker.” *Nature*. 8 May 2013. Accessed Dec. 2014 <www.nature.com>.

⁵ Skloot, Rebecca. “The Immortal Life of Henrietta Lacks, the Sequel,” *The New York Times*. 23 Mar. 2013. Accessed Dec. 2014 <www.nytimes.com>.

⁶ Ibid.

⁷ Ahmed, Abdul-Kareem. “Unhidden Traits: Genomic Data Privacy Debates Heat Up.” *Scientific American*. 14 Aug. 2013. Accessed Dec. 2014 <www.scientificamerican.com>.

⁸ Ferguson, William. “A Hacked Database Prompts Debate about Genetic Privacy,” *Scientific American*. 5 Feb. 2013. Accessed Dec. 2014 <www.scientificamerican.com>.

⁹ Ibid.

how fast technology has evolved and how scientists are struggling with these new advancements.

The first human genome was sequenced roughly eleven years ago, with the help of over two hundred scientists, costing the U.S. government approximately three billion dollars.¹⁰ Sponsored by National Human Genome Research Institute (NHGRI), the Human Genome Project was able to publish all three billion base pairs that comprise the human genome. After the end of the Human Genome Project, the cost of genome sequencing declined exponentially roughly following Moore's law. When next generation sequencing entered the market seven years ago, the cost to sequence the whole human genome dropped dramatically. Today, a full genome profile hovers around five thousand dollars.¹¹ Since then, sequencing costs have dropped dramatically and it was estimated to cost approximately five thousand dollars to sequence the whole human genome one year ago. The NHGRI has launched the new Advanced Sequence Technology award to spur innovators to create a thousand dollar sequencer. Illumina claims to have reached this goal and supplies these fast machines to research centers, pharmaceutical companies, academic intuitions, and biotechnology companies. With cheap but powerful technology, doctors around the world now have the ability to quickly sequence genomes to save lives.

The Children's Mercy Hospital in Kansas City is saving the lives of newborns one genome sequence at a time. A two month year old boy has spent his entire life in their neonatal intensive care unit (NICU) at the brink of death. Doctors were mystified and could not determine the cause of his abnormalities.¹² They warned his parents that his life may be cut short. Refusing to accept that as an answer, geneticist Stephen Kingsmore and his team used rapid sequencing technology to quickly identify differences between the genomes of the boy and his parents. They were able to find a rare mutation that was linked to a disease that results in an overactive immune system which damages the liver and spleen.¹³ This diagnosis helped the doctors tailor a remedy which helped lower the boy's immune response. The boy is now at home and is expected to live a long and healthy life. This rapid genome sequencing technology has been used by experts to successfully help diagnose and treat forty-four sick infants. The team is now faced with determining what to do with the extra information they gleaned that is unrelated to the diagnosis. This information could potentially shed light on diseases that may help other children. So while this story highlights the benefits of genome analysis, it raises the question: should individuals have an expectation of privacy when it comes to their genetic material and is it considered property?

Genome privacy revolves around the fundamental concepts of ownership and authorized use. Traditionally speaking, a person has property rights or ownership of his/her body. This is why donors must consent before any or all parts of their body are given away. However, if the individual is in anyway involved with a criminal investigation, he/she can expect no privacy. DNA evidence found at a crime scene or DNA extracted in the context of a criminal prosecution does not violate the Fourth Amendment. Once convicted, the U.S. Federal Bureau of Investigation has a database of all "samples from convicted

¹⁰ Hayden, Erika Check. "Technology: The \$1,000 Genome," *Nature*. 19 Mar. 2014. Accessed Dec. 2014 <www.nature.com>.

¹¹ Ibid.

¹² Reardon, Sara. "Fast Genetic Sequencing Saves Newborn Lives." *Nature*. 30 Sep. 2014. Accessed Dec. 2014 <www.nature.com>.

¹³ Ibid.

offenders, crime scenes, victims of crime, and unidentified human remains”.¹⁴ The court believes that defendants have no expectation of privacy. But what happens when an individual is not a felon? The consensual donation of genetic matter for the purposes of research either for academic or commercial applications raises multiple issues and with genomic policies still in a nascent stage, subjects are at risk.¹⁵ For the past few years, the media has been in an uproar about database breaches that contain social security numbers, credit card information, and other personally identifiable information. It is time for the public to be as concerned about the privacy and security of human genomic data.

There are a myriad of potentials way to abuse the information collected from genetic materials and without privacy laws, this is difficult to stop. Baase defines privacy as having three key aspects: “freedom from intrusion...control of information about oneself, [and the] free from surveillance...”¹⁶ Genomic information falls under the second category of privacy. Individuals are concerned that their genetic information will be used to screen potential employees or even used to deny, limit, or cancel insurance policies.¹⁷ Though the U.S. government contends this is a small concern since only a limited number of cases have been brought to their attention.¹⁸ They do admit that this may be due to the fact that it is difficult to uncover and document the discriminatory use of data. The federal government has attempted to prevent genomic discrimination with the Genetic Information Nondiscrimination Act of 2008 (GINA). GINA supplements the Privacy Act of the Health Insurance Portability and Accountability Act (HIPAA) and the Americans with Disabilities Act (ADA). GINA and HIPAA state that health insurance companies cannot use the genetic data for their own testing in either the capacity of providing samples to researchers or sponsoring researchers to test samples.¹⁹ They can however use the genome to define life insurance policies and long-term disability care.²⁰ GINA and the ADA state that employers cannot ask prospective employees for genetic information (medical examination) during the pre-employment stage. However, at the pre-placement state, the employer can require a medical examination to be conducted. The employee has no right to be told what they are getting screened for, how the test will be used, or what the test results are. Furthermore, if the employer decides to withdraw an offer of employment, the employee will not be told why. It is against ADA policy to withdraw a conditional offer based on medical reasons but if the company fails to disclose a reason for unemployment, it is difficult to take up legal action. At the state level, approximately half the states have legislation that prohibits genetic discrimination in the workplace.²¹ Each state has its own spin regarding genetic testing and many of them fail to adequately define the what, how, and when genetic material can be used. The need for proper federal regulations is

¹⁴ Nerko, Charles J. “Assessing Fourth Amendment Challenges to DNA Extraction Statutes after *Samson v. California*.” *Fordham Law Review / Edited by Fordham Law Students*. Nov. 2008.

¹⁵ Curley, Robert A. Jr. & Lisa M. Caperna. “The Brave New World Is Here: Privacy Issues and the Human Genome Project” *Defense Counsel Journal*. Jan. 2003.

¹⁶ Basse.

¹⁷ Curley and Caperna.

¹⁸ “Privacy and Progress in Whole Genome Sequencing.” Presidential Commission for the Study of Bioethical Issues. Oct. 2014 Accessed Dec. 2014 <bioethics.gov>.

¹⁹ “Summary of the HIPAA Privacy Rule.” *Health Information Privacy*, U.S. Department of Health & Human Services. Accessed Dec. 2014 <www.hhs.gov>.

²⁰ Ferguson.

²¹ Curley and Caperna.

increasing especially as government, medical, and research organizations begin to work together to create large databases that can be breached with genome profiles stolen and sold to the highest bidder.

Last year, seventy leading medical and research organizations spanning forty one countries, including the NIH, has declared their intent to form an alliance that will build a shared database of genomic and clinical data. They hope that this huge database will uncover genetic links that have so far proven to be elusive.²² The gene variants will help pinpoint diseases such as cancer and diabetes and will usher the advent of a new type of personal medicine; one where physicians use each patient's genome sequence to prevent diseases and to help customize treatment plans. The most troubling aspect is that the alliance has no plan for how to protect the privacy of genetic donors. If this database is breached, the genome can be used not only against the donor but also against their relatives who have not consented to the study. This is not the only database individuals should be concerned about.

23andMe is marketed as a personal genome service that allows their customers to learn more about their health. The company stores all the genomic data in a database that contains metadata that volunteers submit about themselves. Their current privacy policy states that they only share aggregate information about user's genomes to third parties and not personal genetic information.²³ However, just like any corporation they protect themselves by having "a broad use of the [user's] personal information."²⁴ In this case, 23andMe reserves the right to use their customers' personal information, which includes their genome, to try to sell them products and services. As a private corporation they have the ability to change their privacy policies on a whim which gives them the ability to sell their customer's DNA to the highest bidder. If that is not ominous enough their website warns, "Genetic information that you share with others could be used against your interest."²⁵ 23andMe state that they only use the information for their own research purposes of finding genetic anomalies. The extremes between privacy and access have forced scientists to rethink which side they lean towards.

The biggest challenge with genome data is the balancing act between access and privacy. On one hand, researchers want to make sure the data is widely available without restrictions. On the other hand, researchers understand that restricting how the data will be used and increasing privacy and confidentiality will draw more donors to the study.²⁶ Jeffrey Khan, a professor of bioethics and public policy at Johns Hopkins Berman Institute of Bioethics, believes that "[i]f privacy can't be guaranteed, then the focus should be shifted to mitigating data misuse."²⁷ It is not enough for researchers to simply inform people that there is a risk of re-identification. Many individuals who sign up for genome based research studies will have a hard time fully grasping how their data might be used in the future especially when researchers themselves are still finding out. Informed consent is no longer sufficient enough to

²² Ahmed.

²³ Collins, Francis S. Collins & Margaret A. Hamburg. "First FDA Authorization for Next-Generation Sequencer," *New England Journal of Medicine*. 19 Dec. 2013.

²⁴ Haynes, Allyson W. "Online Privacy Policies: Contracting Away Control Over Personal Information?" *Penn State Law Review*. 2011.

²⁵ Collins and Hamburg, "First FDA Authorization for Next-Generation Sequencer."

²⁶ Hayden, "Privacy Protections."

²⁷ Ferguson.

cover the potential repercussions to donors. Lawmakers along with scientists, both in public and private organizations, need to focus on way to minimize the risk to individuals who are willing to share their genetic material.

The first task lawmakers can focus on is determining if genetic material is considered uniquely private or if it should be lumped together with other sensitive information in individual's medical records. If genetic material is considered to be to be common medial information, then the existing confidentiality laws with HIPAA and ADA should suffice and the only amendment needed would be to explicitly state this within those acts. However, the best course of action would be for the federal government to support genetic exceptionalism, or the idea that genetic material is unique, because of its distinct abilities.²⁸ Genetic material is the source of all medical information and similar to a personal medical record, can be stored and accessed without the permission of the owner. However, genetic material goes above and beyond the information found in a medical record, it has to ability to positively identify an individual's future risks. Because of the unique features of genomes, the genetic material extracted from an individual should follow a new set of laws.

Lawmakers should focus on affording special privacy laws to both the sequence information as well as the sample that is extracted from the individual. It is important to include the sample since once a physician or researcher extracts a genetic sample, they will have the ability to conduct tests on the sample in the individual's absence. The individual will be powerless and cannot stop what they do not know. Currently, in the U.S., genetic tissue samples of about twenty million people is being collected and stored each year in collections ranging from two hundred to ninety-two million samples.²⁹ The ease and proliferation of genome databases make them a gold mine for researchers who are focused on solving which genetic factors contribute to what diseases but at the same time the mass quantity places thousand if not millions of individuals at risk. The genetic sequences could be tied to individuals in the real world and be used to potentially harm them. Not only are the samples a risk but the large genome databases are a huge privacy concern for individuals.

Researchers need to focus on what it means to have a large central repository of genetic material. The data should be set to be off limits to users while allowing trusted and verified researchers blanket permission to use the dataset. One such method is to only allow trusted users would be to encrypt the genome data before it is placed into the database and only allowing those who poses the decryption key to work with it freely.³⁰ Additionally, instead of presuming that a breach will never happen, it is best to be proactive and safeguard against them. All metadata that contains identifying characteristics such as age, location, or year of extraction should be removed from the database. The power to malign individuals with the genomic sequence itself is very powerful; hackers should not have the ability to take it one step further. Lastly, researchers and scientists should be subject to the same codes of conducts, ones with enforceable

²⁸ Roche, Patricia A. & George J. Annas. "Protecting Genetic Privacy." *Nature*. May 2001. Accessed Dec. 2014 <www.nature.com>.

²⁹ Ibid.

³⁰ Ibid.

penalties, which lawyers and doctors are subject to.³¹ Individuals trust the confidential relationships they have with these professionals because they take an oath to protect privacy and have penalties for breaching it.³² With the advent of cloud-based genome databases such as the one Google recently launched, these safeguards will help minimize the risk individuals take to help further humanity.

It is the age-old battle between privacy and social good. Researchers contend that there is always some amount of privacy that individuals are willing to give up in order to further national security.³³ In the internet age, this translates to genomic privacy and the social good of providing individuals with better healthcare. However, with a few simple policies and regulations, individuals will not have to give up all their privacy rights in order to help their fellow community members. Laws and policies should recognize individual genetic rights such as the right to determine who has access to their genetic information and the right to determine who has access to their genetic samples. They should also prohibit the unauthorized use of genetic samples to solve crimes or determine paternity. Furthermore, these laws should be enforced and instituted with severe penalties in order to deter violators. With these simple actions, the world will be one step closer to establishing a standard for genetic privacy.

³¹ Church, George et al., "Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection," *PLoS Genet* 2 Oct. 2009.

³² Ibid.

³³ Ahmed.