Joyce Wang
IS 431
15 December 2017

The Archival Basis of DNA Databases

Abstract

As genetic sequencing technologies become increasingly efficient and affordable, the proliferation and use of DNA records and the implications of maintaining DNA databases becomes an increasingly pressing matter. This paper will first frame the DNA record as a document, focusing on its materiality and types, as well as go into some of its historical uses. The paper will then go on to discuss some of the controversies surrounding the establishment of modern-day DNA databases, putting them in the context of archival theory and ethics particularly with regards to issues of privacy, accessibility, and acquisition.

DNA as a material record

The materiality of DNA is concealed in its microscopic size, necessitating incredibly high-power electron microscopes and complex chemical environments to be rendered visible to the human eye. For many, DNA takes the shape of a colorful, linear double helix, remaining an abstraction no matter how chemically precise or structurally accurate the depiction. Even when represented using complex 3-D modeling software that depict the physical proportions, dimensions, and interactions of the molecule, the physicality of DNA and its orchestration of biological processes does not lend itself to easy documentation. DNA sequencing, starting from its early manifestation in the Sanger method developed in the 1980s to the high-throughput processes of today, is the primary tool used in the generation of the DNA record. By determining the linear sequence of adenine, cytosine, guanine, and thymine nucleotides that comprise a single gene or genome and performing large-scale comparisons with other samples, it becomes possible to identify certain "genetic markers" that could serve identifying, categorical, comparative, or predictive purposes. An individual's genetic sequence is a biochemical record that can be analyzed within certain parameters to describe certain aspects of its subject, whether it be the person's medical condition, ethnic background, or criminal status, in the case of forensics. Thus, DNA falls comfortably into the framework of a "document" as delineated by Suzanne Briet:

"evidence in support of a fact".[1] It should be qualified that the "fact" being referenced here is not absolute, but a belief widely held to be true. As will be discussed later in the paper, the genetic record is subject to varying interpretations and use. Immutable (as far as the sequence is concerned) and difficult to conceal given the human body's penchant for shedding off dead cells, DNA is a ubiquitous and unique document that can paint a very intimate and perhaps novel portrait of the individual in question, in turn reflecting greater sociopolitical beliefs and concerns.

A myriad of DNA records

Despite the materiality of DNA and the experimentally determinable nature of their nucleotide sequence, the generation of a DNA record is not without bias. The term "DNA record" is inexorably vague and does not reflect the variety of genetic documents that get deposited for different purposes in collecting institutions. Today, DNA databases are used in a variety of fields, aiding research in precision medicine, crime scene investigations, and genealogy mapping. As such, the types of genetic information stored in these databases can range from people's copies of a single genetic marker, and in rare cases, to their entire genome sequence. Law enforcement agencies and forensic labs tend to emphasize the documentation of short tandem repeats (STRs), areas in the genome containing a highly variant (polymorphic) number of short nucleotide string repeats that have shown to be very effective at discerning between individuals. The genetic profiles of convicted individuals stored in the massive United State Combined DNA Indexing System (CODIS) are generated from an analysis of 20 select STR locis known as the CODIS Core.[2] On the other hand, popular consumer genetics company 23andMe utilizes a different analytical procedure that genotypes individuals using markers known as single nucleotide polymorphisms (SNPs) – variations in a single nucleotide at a particular location across a genomic population.[3] In a 2016 blog post, 23andMe responded to concerns about law enforcement inquiries into the company's database, stating that "23andMe's

---

[1] Suzanne Briet, Ronald E Day, Laurent Martinet, and Hermina G B Anghelescu, *What is a Documentation?: English translation of the classic French Text* (Lanham: Scarecrow Press, 2006), 9.

[2] Douglas R. Hares, "Selection and Implementation of Expanded CODIS Core Loci in the United States," *Forensic Science International: Genetics* 17 (2015).

[3] 23andMe, "23andPrivacy: Your Data and Law Enforcement," 23andMe Blog (2016).

test focuses on how you are *like* other people, while forensic tests focus on how you are *different* from other people".[4] A genetic record should not be imagined as a read-out of an individual's genetic sequence, but the approximate result of certain biochemical and statistical manipulations under the direction of not-infallible algorithms that satisfy (or give the illusion of satisfying) an institutional and consumer need.

Early friction: DNA records and

Though national DNA databases – largely for forensic purposes – started to emerge in countries such as the UK, Brazil, and United States in the late 1990s, the large-scale acquisition of genetic material from specific groups of people was embedded in a popular post-1960s ideology that DNA could empirically elucidate human evolutionary history. Population geneticists and anthropologists alike were convinced that DNA analysis could produce an anti-racist science free of the blatant prejudice that had defined eugenics and social Darwinism. With this belief in mind, a group of population geneticists at Stanford University launched the Human Genome Diversity Project (HGDP) led by prominent geneticist Luca Cavalli-Sforza. The team sought to conduct an expansive survey of the diversity of the human genome by collecting DNA samples from more than 50 indigenous populations around the world.[5] While these samples were originally acquired with the promise that they would be used to help the medical needs of these communities, it became clear not long after that these documents were becoming another colonialist battlefield waged against minority populations through the artificial creation of genetic distinctions and the patenting of genetic materials for commercial gain on the part of the research group. Historian of science Cathy Gere termed the HGDP's collection of DNA and other biological materials as constituting a "biocolonial archive", a holding of plundered records. (208). The records contained in the database, while themselves likely experimentally precise, existed in an institutional framework that failed to uphold values of transparency and proper representation of the subjects in its records. In separating the record and the human from which it was derived, the HGDP's "biocolonial archive" presents a classical case of the bureaucratic, extractive archive that fails to prioritize social good, particularly within marginalized populations.

---

[4] 23andMe, "23andPrivacy: Your Data and Law Enforcement," 23andMe Blog (2016)..

[5] Jun Z Li, et al, "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation," *Science* 319 (2008): 1.

Formation of the archive: Issues surrounding collection of DNA records

The acquisition of DNA samples for inclusion into a repository remains one of the most controversial aspects in the establishment of DNA databases. According to the Dutch archivist-scholar Eric Ketelaar, the agency of the individual from whom the documents were obtained should determine the documents' value and accessibility for posterity.[6] Records obtained under coercive circumstances and/or in the absence of informed consent should have restrictions in their use and access. That an archive may contain information about a person without their knowledge or consent is nothing new. After all, the covert collection of information has throughout history been the foundation of state control. With information being transmitted with ever-increasing ease, and collected at more and more checkpoints within one's daily life, it seems that the distinction between public and private information has become increasingly liminal.

As can be expected, the issues revolving around the collection of genetic information becomes particularly controversial in forensic and disciplinary settings. The United States Combined DNA Indexing System (CODIS) is a software platform that analyzes DNA records kept in forensic laboratories at the national (NDIS), state (SDIS), and local level (LDIS) which altogether make up the world's largest DNA database. As of October 2017, the CODIS contains the STR profiles of up to 16 million individuals, 13 million of which originate from offenders.[7] The incorporation of DNA records into these three levels, each with their own standards of use and access, occurs in a hierarchical fashion. In most cases, samples are deposited first in the local level of the CODIS database before ascending into the state or national repositories if the individual is convicted or arrested. State repositories are regulated on a state-by-state basis, but many local and regional labs, while also officially subject to federal and state laws, operate with much greater autonomy. A 2013 *New York Times* article (notably published only a week after the Snowden NSA surveillance leaks) reported a rising number of DNA samples that were acquired covertly by police enforcement agencies across the country. Many records are obtained from low-level offenders in exchange for plea bargains, with the justification that data from the NDIS corresponded primarily to convicted criminals who were already going to prison. The *Times*

---

[6] Eric Ketelaar, "The Right to Know, the Right to Forget? Personal Information in Public Archives," *Archives and Manuscripts* 23 (1995): 14.
[7] "CODIS-NDIS Statistics," FBI, https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics.

article also reports on surreptitious feats of DNA collection, with investigators obtaining and retaining for unspecified amounts of time samples from discarded waste, crime scenes, and the victims themselves.[8] The Electronic Frontiers Foundation reports efforts to collect DNA from refugees and asylum seekers as well as their families.[9] The preemptive collection of DNA in such conditions can be seen as a breach of 4th amendment rights, with the mere implication of the person in the database being to some degree a document of guilt.

Even outside the realm of law enforcement, there are numerous occasions in which DNA records could be covertly produced. Hospitals are rapidly amassing blood samples from newborn babies as a result of the many health screenings babies born in the United States now all undergo. In recent years, those residual samples have ended up being stored in a national newborn DNA bank (whereas such samples may have been tossed before).[10] In most cases, parents are unaware of the DNA depositing that occurs after the screening, and efforts to obtain consent are often misleading. For example, many states have an "opt-out" option for screening tests altogether, rather than for the storage of samples in newborn DNA banks.[11] Defendants of the newborn DNA database argue that the samples will be used for the purposes of medical research, although what that precisely entails is largely unknown. Regardless of how the data is used, the generation of the DNA record must be dealt with greater transparency.

The issue of unregulated DNA record formation/collection can be viewed as a matter of appraisal and reappraisal. As stated by Anne Gilliland, one aspect in the proper appraisal of a document for archival keeping involves "assessing the kinds and degrees of bureaucratic, legal, research, cultural, community, personal and intrinsic values that are present in extant records in order to arrive at a decision about their eventual disposition."[12] In the case of forensic DNA databases, the legal and personal valuations of the document may conflict, meaning the individual's desire to keep their DNA record private should be weighed along with the law enforcement's desire to expand their database of potential criminals. In appraising records obtained from newborn screenings, a similar recognition of the need for obtaining proper consent

---

[8] Joseph Goldstein, "Police Agencies Are Assembling Records of DNA." *The New York Times*, (2013).
[9] "Federal DNA Collection," *Electronic Frontier Foundation*, 2012.
[10] Aditi Shah, "Do You Know Where Your DNA Is? Genetic Privacy and Non-Forensic Biobanks." Council for Responsible Genetics (2014): 6.
[11] Ibid.
[12] Anne J Gilliland, "Archival Appraisal: Practising on Shifting Sands," in *Archives and Recordkeeping: Theory Into Practice* (Facet Press, 2014), pp.34.

should play an integral role in the retention of the document. Furthermore, especially in the case of the forensic database, there should be a standard of reappraisal that could better define the retention periods of the DNA documents, particularly when their existence in a particular database could be damaging to the individual. Ultimately, the act of including a genetic profile into a database requires a recognition of the circumstances in which the documents were produced or acquired.

Issues of access

Related to the issue of unregulated DNA collection is that of access. The function of an archive lies in its accessibility, which is dependent on people knowing what types of documents it carries, who is represented, and how they are represented. Even with the case of voluntarily donated samples, the donor is not always aware of what exactly is being kept as well, nor of the potential consequences that could result from its collection and use, and would thus have trouble in accessing the full implications of their materials. Medical DNA databases also contain one caveat that restricts their access compared to traditional archives: the record has usually been "anonymized". Ironically, while this is done to safeguard the identity of the donor, numerous studies have emerged wherein researchers were able to re-identify the owner of the sample using the associated metadata (13).[13] Hence, it is very likely that the record remains accessible to other third parties while remaining inaccessible to the individual who owns it. This imbalance in the technical aspect of accessibility, as well as the oftentimes sheer lack of knowledge of a record's existence, makes DNA databases intrinsically prone to breaches of privacy.

A month ago, Democratic senator Chuck Schumer called for greater scrutiny regarding the privacy policies of popular genetic testing packages such as 23andMe and AncestryDNA (service provided by long-time genealogy website Ancestry.com). It is public knowledge that genealogy services sell the information they obtain to third-party members – in fact, this is their main source of revenue, more so than the actual sale of test kits to consumers.[14] Ancestry.com's privacy statement cites numerous cases in which the company may be compelled to share user's information with third-parties even in non-research related cases.

---

[13] Aditi Shah, "Do You Know Where Your DNA Is? Genetic Privacy and Non-Forensic Biobanks." Council for Responsible Genetics (2014): 13.

[14] wired

Conclusion

DNA records are a relatively new form of identifying document that should be handled with the same considerations as would be applied to traditional archives. However, with regards to issues of appraisal, accessibility, and privacy, they present numerous challenges that will require novel views for archival ethics.

Works Cited

23andMe. 2016. "23andPrivacy: Your Data and Law Enforcement." 23andMe Blog. March 16, 2016. https://blog.23andme.com/23andme-and-you/23andprivacy-your-data-law-enforcement/.

Briet, Suzanne, Ronald E Day, Laurent Martinet, and Hermina G. B Anghelescu. *What Is Documentation?: English Translation of the Classic French Text*. Lanham, MD: Scarecrow Press, 2006.

"CODIS-NDIS Statistics." FBI. https://www.fbi.gov/services/laboratory/biometric-ana.lysis/codis/ndis-statistics (accessed November 19, 2017).

Gilliland, Anne J. "Archival Appraisal: Practising on Shifting Sands," in Archives and Recordkeeping: Theory Into Practice, Patricia Whatley and Caroline Brown, eds. (Facet Press, 2014), pp.31-61.

Goldstein, Joseph. 2013. "Police Agencies Are Assembling Records of DNA." *The New York Times*, (2013). http://www.nytimes.com/2013/06/13/us/police-agencies-are-assembling-records-of-dna.html.

"Federal DNA Collection." 2012. Electronic Frontier Foundation. December 13, 2012. https://www.eff.org/foia/federal-dna-collection.

Hares, Douglas R. 2015. "Selection and Implementation of Expanded CODIS Core Loci in the United States." Forensic Science International: Genetics 17 (Supplement C): 33–34. https://doi.org/10.1016/j.fsigen.2015.03.006.

Ketelaar, Eric. "The Right to Know, the Right to Forget? Personal Information in Public Archives," *Archives and Manuscripts* 23 (1995): 8-17.

Li, Jun Z., Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini

Ramachandran, Howard M. Cann, et al. “Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation.” *Science* 319 (2008):1100–1104. https://doi.org/10.1126/science.1153717.

Millar, Laura A. Archives Principles and Practices, 2nd. ed (London: Facet, 2017), pp.93-115.

Shah, Aditi. 2014. “Do You Know Where Your DNA Is? Genetic Privacy and Non-Forensic Biobanks.” Council for Responsible Genetics.