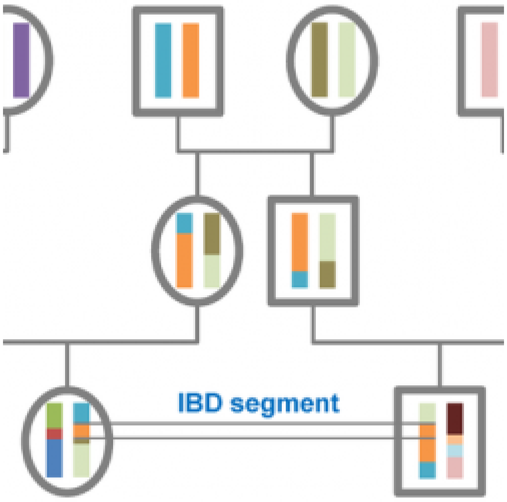
Published on *American Council on Science and Health* (https://www.acsh.org)

Genetic Searches Are Another Nail in Privacy's Coffin

By Chuck Dinerstein — October 11, 2018



Courtesy of GKLambauer [1]

Back in April of this year the media was attracted to the capture of a suspected serial killer through the use of genetic data from the crime scene and GEDmatch, an open-source genealogy website. A new paper in Science on the science behind this form of investigation termed long-range familial research demonstrates that the ability to identify anyone through the use of multiple data sources is not far off. First the science, then the implications.

Science and Search

We are, genetically speaking, mostly the same – 99.9% of our DNA is identical. The remaining 0.1% differs because of more recent changes, due to the DNA swapping of our most recent common ancestors, a term of scientific endearment for our parents. The phrase, “you can choose your friends, but not your parents,” is applicable here too, because genetically that remaining 0.1% is more like your family than it is your friends; genetically those similarities can extend out to 4th and 5thcousins. Individuals that share long DNA segments, segments termed identical-by-descendent (IBD) segments are related. The more similar the segments, the closer the common ancestor. Forensic databases of DNA have existed for years and contain the DNA of those convicted of violent crimes as well as DNA from crime scenes. Because of its particular membership, convicted violent felons, and unknown criminals, it can be used to identify individuals but only if they are very close, siblings or first cousins. The search space within the limited data set is too small. The rise of direct to consumer DNA testing has increased the gene pool, so to speak, tremendously and now we can identify 2ndthrough 4thcousins. The search space of genetic investigation has gotten a lot more personal.

The human genome is quite long and to map it out and facilitate comparisons scientists use a measure termed the centimorgan, cM. It measures the frequency of recombination, the recombination that occurred when your grandparents created your mom and dad. The centimorgan is “a 1% chance that a marker at one genetic locus on a chromosome will be separated from a marker at a second locus due to crossing over [recombination]." While is not a physical measure, like the centimeter, it can be used to map distance. So the methodology is straightforward, segments with a large number of cM in common are more likely to indicate a common ancestor.

The researchers made use of one of these direct to consumer (DTC) databases, containing an estimated 1.28 million individuals. They used the information on one member of the database to search for relatives. They removed members who had IBD segments greater than 700 cM, which would represent “first cousins or closer relationships,” the kind of relations you might find in searching a forensic database. They then searched for members with IBD segments within 30 to 600cM of their references. 60% of the time they could identify a 100cm match or better, equivalent to a 3rdcousin. They replicated their findings using the same techniques on the GEDmatch database, identifying similar segments 76% of the time. Using this form of genetic mapping and matching they reduce their search, from 1.28 million individuals to about 850 – a very significant reduction, but still too extensive a list to effectively consider.

They then winnowed down that list using more conventional, basic demographic information, freely available. If you could localize your target to a hundred mile area, you could eliminate 57% of that list; assuming the target’s age within ± 5years rejected 91% more, and if you knew the target individual’s gender, you narrowed that list of 850 to about 16. Easy enough for a more manual, door to door search. To further confirm their work, they used a known sample from the 1000 Genomes Project [1] again submitting the sample to GEDmatch; after a day’s work, they had successfully reidentified their target.

Implications

First and foremost, as with all genetic information, it is racially based, it does little good to search for an Asian in a dataset of Europeans. In a kind of reverse racial disparity, this search strategy works best with those of Northern European ancestory, they have offered up the preponderance of available DNA and have the bulk of geneologic data. But that said, the researchers believe that the combination of genetic and demographic search will allow them to identify any individual once 2% of the population has volunteered their DNA. While the researchers believe that singularity is rapidly approaching, the MIT Technology Review [2]suggests that 4% of the US population already have provided DNA samples to DTC databases in exchange for ancestory and “health” information.

Ethicists first concern was privacy. Most of these DTC databases, like 23AndMe, protect your privacy through their terms of use, but only to the extent that they can legally do so. That privacy is overcome with a subpoena; a court order that protects your 4thamendment rights, protecting you from unreasonable search and seizure without due cause. For those of us, like my colleague Dr. Wells who believe this is personal healthcare information (PHI) protected by our HIPAA laws, I believe you are mistaken. All of these DTC websites make a concerted effort to distance themselves from medicine and medical judgment, there is no physician-patient relationship there.

The more difficult privacy consideration is the same raised when considering Facebook and Cambridge Analytica. A friend or acquaintance, in allowing access to their data, gave away your data and privacy. This effect is far more powerful when considering genetic information; your 3rd cousin, who you have never met can supply the information needed to find you and there is no legal structure to prevent that from happening.

The authors are concerned primarily with the chilling effect this will have on research, where genetic data is being collected, information that is not currently considered identifiable personal health information, subject to HIPAA's stringent control. This can be corrected, they point out, by a regulatory change issued by Health and Human Services. Their second recommendation is to encrypt raw genotyping files to prevent, or more likely significantly reduce, intrusion by unauthorized individuals. It would still provide a means for law enforcement to access the information, given a valid subpoena assuring individuals of their 4thamendment rights if not their privacy. Both recommendations are worth a greater audience and discussion.

[1] This is a global project that ran until 2015, their goal was “to find most genetic variants with frequencies of at least 1% in the populations studied.”

Source: Identity Inference of Genomic Data Using Long-Range Familial Searches Science DOI: 10.1126/science.aau4832

Source URL: https://www.acsh.org/news/2018/10/11/genetic-searches-are-another-nail-privacys-coffin-13496
Links
[1]
https://en.wikipedia.org/wiki/Identity_by_descent#/media/File:Pedigree,_recombination_and_resulting_IBD_segments,_
[2] https://www.technologyreview.com/s/610233/2017-was-the-year-consumer-dna-testing-blew-up/