Crowdsourcing big data research on human history and health: from genealogies to genomes and back again

April 12, 2018

By Alon Keinan: with contributions from Alexandre Lussier

About two months ago, Science Magazine invited me to write a Perspective to appear in the same issue as “Quantitative analysis of population-scale family trees with millions of relatives” (AKA the Geni genealogy) by Joanna Kaplanis, Assaf Gordon et al. with Yaniv Erlich and many others. Realizing that this paper will be covered by everyone and their mother, I spent days aiming to come up with unique facts and perspectives, finding myself ensconced in what became research projects, reading dozens of articles, and going down multiple rabbit holes of writing. Nothing that I could have afforded doing within the 30 days I was given to produce the Perspective. Based on other papers I previously discussed with Science editors, though Kaplanis et al. assembled and analyzed genealogies, my mandate from Science was to cover crowdsourced studies in related fields as well, especially in genetics, and the state of these fields in general—past, present, and future. And the icing on the cake: All that in 1250 words (including references!)

Published today in Science (10.1126/science.aat2634), alongside the publication of Kaplanis et al., (of which a First Release appeared online in the interim), is our Perspective, not before removing many parts, frantically cutting down, and editing in several rounds, including with the gracious help of Science Perspective editor, of what has been a much, much longer draft. I thought to complement its publication with the following longer-read version. While not as concise, I find it to be less cryptic to read, more accessible to a general audience, and comprehensive in discussing aspects of that important study and the state of the pertaining fields. If nothing else, I found it worthwhile for being able to use my full-length title (The Hobbit anybody?)

Genealogies are likely the first, centuries-old “big data”, with their construction as old as human civilization itself. Globalization, and the identity crisis that ensued, turned many to online services, building family trees and investigating connections to historical records and other family trees [1]. An explosion has been underway since the beginning of the century in the number and usage of websites offering such genealogical services. About 130 million users combine to have created almost four billion profiles for family members across the three most popular websites of genealogy enthusiasts, Ancestry.com, MyHeritage, and Geni. More recent years have witnessed a similar rapid increase of genetic-based services that address the same need to learn about familial relationships and ancestry. These vast amounts of crowdsourced—and often crowdfunded (as users often pay for these services)—data offers ample scientific research opportunities that would otherwise require expansive collection. In a paper published today in Science, Kaplanis et al. [2, 3] introduce a genealogical dataset based on processing 86 million public Geni profiles. Armed with this crowdsourced dataset, they address fundamental research questions that I will touch upon in the following.

Genealogies constructed and harnessed for research throughout history

The use of genealogies as a research tool is as old as recorded history, spanning early research into human history to modern genetics and medicine. For instance, genealogies have been maintained by the Catholic Church since the 13th century, with a 1563 injunction requiring recording each baptism, marriage, and burial in all parishes. By virtue of monarchical systems, ruling families have also preserved centuries worth of genealogical data. Such genealogies of royal dynasties have recently served to study the impact of inbreeding (marriage between very close relatives) on offspring survival [4, 5].

Centuries of Icelandic genealogical enthusiasts have used both current data and scriptures from early ages to reconstruct their family history. This process was sped up by the establishment of deCODE genetics, leading to the creation of the Íslendingabók, a genealogical database composed of 864,000 Icelanders, including
~300,000 contemporary ones, ~95% of those born since 1700, and some dating back to the settlement of the island in the 9th century. Made freely available to all Icelanders in 2003, this resource has proven an incredible research tool. For instance, a recent study that estimated the relative role of genetics (heritability) in many complex diseases and other traits—e.g., type 2 diabetes, BMI, and number of children—further used the Íslendingabók to contrast individual pairs with equal expected level of genetic similarity but a different likelihood of shared environment. For example, a grandparent-grandchild pair and half-siblings both share 25% genetically, but the latter more often share a household. These comparisons showed that environmental factors (rather than more complex genetic contributions) predominantly explain the relatively low estimated heritability of most traits [6].

Another unique effort of digitalizing genealogies and opening them to the public has been made by The Church of Jesus Christ of Latter-day Saints (the Mormon Church). Since 1921, church employees have constructed genealogies from historical records, later giving rise to FamilySearch.org. With the introduction of personal computers, starting in 1987, hundreds of thousands of volunteers continued the process, which by 2013 reached a rate of creating half a million profiles daily—a true crowdsourcing fit! Of note, the Mormon church genealogy also connects to health care records for at least 700,000 individuals, which offered the earliest opportunities for incorporating detailed family history in risk assessment and for personalized medicine [7].

**Enter the age of genomics**

Not only genetic- and genealogical-relatedness are used synonymously in many other fields, they share a related ancient greek root (genos, meaning family, offspring, among else) and a very old Indo-European root (*genə*; to give birth). With advancements in genomic technologies, genealogy enthusiasts have greeted a new tool to assist in their quest for familial identity. Direct-to-consumer (DTC) genetics, initially based on the small fraction of the history that can be gleaned from one’s Y chromosome and mitochondrial DNA, started taking hold with the launching of the National Genographic’s service in 2005, which reached a million individuals. Several companies and organizations, including those already offering genealogical services, were quick to follow and, with dropping costs, offer DTC services based on whole-genome genetic testing. Each company offers a different menu of services to customers, but they all offer a service that allows users to find relatives, which is based on locating and analyzing stretched of DNA that different individuals seem to share due to having inherited them from the same ancestor (identity-by-descent) [8]. The power of this service increases as more and more users are joining. While still well behind the number of people represented in genealogical services, 16 million individuals have taken advantage of whole-genome DTC genetic testing: Four companies now have over a million consumers each; of note are AncestryDNA, concurrent with their genealogical service, with over seven million and 23AndMe with five.

DNA-based services provide several advantages compared to genealogical-based searches. Importantly, while genealogical services allow family tree merging based on merging different profiles of the same individual, DNA-based relative finding can directly discern relatedness of a level equivalent to distant cousins [8]. For instance, several years after I became a 23AndMe customer, I noticed a new predicted relative. The high percent of DNA shared and the predicted relationship of second cousin would not have been enough to intrigue me.
Rather, it was his last name, the name of my deceased maternal grandfather. A Jewish teenager on the eve of WWII, my grandfather escaped his home in Lodz, Poland towards the Ural Mountains in Russia. Hundreds of family members stayed in Lodz and were murdered in the Holocaust, with few survivors. In 1950, he immigrated to Israel, where his extended family consisted of one brother’s family, the only survivor of six siblings. An elaborate family tree on Geni, contributed by many around the world, corroborates the family history that has been known to my grandfather. Fast forward to 23AndMe, it did not take me very long to conclude that the predicted relative must be a son of my grandfather’s brother. Not before considering other predicted relatives and sharing information with some, did I verify that he is the son of one of my grandfather’s other brothers! Another brother has survived and immigrated to the USA. I have been able to reconstruct another detailed branch of my family tree, a genealogy enthusiast dream and evidence of the powerful, crowdsourced big data at the disposal of DTC genetic testing companies.

Many other “third-party” websites allow users to upload their existing DTC genomic data, broadening the accessibility of crowdsourced data that is otherwise not made public, even though few of these websites obtain consent to use uploaded data for research purposes. Such websites attract users by bringing together customers of different DTC genetic services and offering additional services. For instance, GEDmatch has genetic data of 950,000 users, with an additional 50,000 monthly, and additionally allows the uploading of family trees (~89,000 updating; 3,000 monthly) in standard format that can be extracted from any genealogical service. GEDmatch provides uploaders with new analyses of their genetic data, revealing information beyond the scope of DTC genetic testing companies’ results. However, the goals and terms of service for these websites are often different from those of the companies that generated a user’s data. Consequently, consent and data protection may be different from what customers have come to expect and may have repercussions on the privacy of consumer data. Some even go so far as to request additional personal information, for example—as a prerequisite to accessing a certain result—income and education, which may pose particular risks if accidentally falling into the wrong hands, when combined with genetic and genealogical information.

Crowdsourced genetic research

Large DTC genetic testing companies are using the crowdsourced genomic data in their possession for research purposes. AncestryDNA applied the method for relative finding to 770,000 US customers, detected 500 million pairs of related individuals, and thereby uncovered detailed events in post-colonial North American history [9]. Meanwhile, 23AndMe, based on US customers with additional demographic information, characterized population migrations to the US and within the US, as well as population admixture – for instance, how the DNA of African American populations today is attributed to DNA from populations of African, European, and Native American ancestry [10]. However, only a small portion of crowdsourced research focuses on the history of populations or their general genetic characteristics, as the majority is medically-driven and relies on customers providing additional information. In particular, 23AndMe has geared its research toward the genetic basis of complex diseases and other traits, as evident by the vast majority of the almost 100 scientific articles reported by the company. For example, Pickrell et al. [11] identified genetic risk factors that overlap between any of 42 human traits, of which 17 were newly reported 23AndMe-based studies. 23AndMe further advance their research by sharing summaries of their data with scientific collaborators and pharmaceutical corporations, and through the recently announced funding for their drug development unit. For instance, a recent pharmacogenomic study by Janssen Research & Development analyzed 23AndMe’s customer survey of antidepressant efficacy [12]. deCODE genetics has obtained genomic and medical data for 160,000 individuals in the Íslendingabók genealogy, more than half of Iceland’s adult population. For the last two decades, across >100 scientific papers, many of which very visible, deCODE has contributed to advancements of all the fields
mentioned above more than any other entity, including in the development of groundbreaking methods and tools for improving such studies.

**When big genealogical data meet statistical and population geneticists**

Although the study by Kaplanis *et al.* focuses on analyzing genealogies, it is abundantly clear—by the research questions, threads of thought, and methodologies, as well as the word cloud of the paper's text on the right—that the research has been conducted by statistical and population geneticists. Beyond the size of the analyzed dataset, which allows fine-resolution of analyses, it is unique in the extensive processing and validation of the genealogies, e.g., resolving cases where an individual appears to have three parents [2]. It also benefits from data supplementation by high-tech products such as Yahoo!’s ‘geoparsing’ service that recognizes location names from plain text. This feature allowed the authors to obtain accurate geographical information for an individual’s birth, death, and residence, which are only available as a textual description for older profiles. Of the 86 million profiles, 43 million are connected to other profiles, constituting 5.3 separate million family trees overall, with the average family tree spanning 11 generations. The largest family tree in their dataset consisted of a whopping 13 million individuals, tracing back 11 generations on average and over 20 generations for many. The resulting genealogies allow historical analyses dating back centuries.

Historical research questions (of the Western world) addressed with these data are ones that have been considered previously in population genetics, among other disciplines. Specifically, considering sex-differences in migrations over the last 300 years, the study shows that females migrate more often, except in the case of long-range migrations, such as immigration to a different country [2]. An interesting, detailed analysis involves trends over time between married couples in their relatedness and distance at birth. Most were born less than 10km apart before the Industrial Revolution (1750), followed by gradual increase, and acceleration of that increase with the second Industrial Revolution (1870) to over 100km. However, average relatedness remained mostly the same prior to the second Industrial Revolution (average equivalent to fourth cousins), despite increasing distance. Following this ~50-year lag, relatedness starting decrease in line with distance at birth. The authors hypothesize that recent decreased relatedness is due to shifting cultural norms, rather than increased distance, due to the inconsistent relationship between relatedness and distance [2]. This seems concurrent with popularized writing from the time, which led to 13 US states passing cousin marriage prohibitions by the 1880s [13] (though in question here are more distant relatives). A related study of 160,000 couples in the Icelandic genealogy replicated a previously-observed phenomenon that closer relatedness between a couple (e.g. third or fourth cousins) is associated with higher fertility [14]. By the nature of this relatively homogeneous population, the study could mostly rule out social and socioeconomic reasons that could increase the number of offspring (e.g. preserving wealth within extended families), and concluded that the phenomenon may potentially have a biological basis [14].

The main results of Kaplanis *et al.* involve life span. The resolution of this dataset allows them to discern not only a reduction in life span during WWI and WWII, but also that this reduction was more pronounced for military age individuals. Despite these temporary reductions, I noted that an increase of an almost constant rate was evident, corresponding to an increase in average life span of about 4 years every generation since approximately 1850. Kaplanis *et al.* also conducted a meticulous study of the factors affecting life span,
along some of the same lines of the study of heritability in the Íslendingabók genealogy [6]. They conducted a meticulous study of factors affecting life span, attributing ~7% to gender, birth year, and geography combined. They estimated life span heritability at 16.1 ± 0.4%, lower than most previous studies, although among them, the largest genealogy-based study until now provided a comparable estimate of 15 ± 3% in the Mormon genealogy [15]. Kaplanis et al. further estimate that an additional ~4% of life span is attributable to dominance (where having a single copy of a genetic variant constitutes the majority of the effect of having two) and none to interaction between different genetic variants.

As rich as their study is, Kaplanis et al. only scratch the surface of their resource, now publicly available with individuals anonymized, and so additional insights are likely to come from multiple disciplines and sources. In particular, it may be interesting to reanalyze life span factors focused on very high longevity, including level of heritability. This resource will also allow researchers to revisit questions studied with smaller genealogies, such as the effects of relative marriage fertility and life span. We anticipate that it will be used in numerous future studies across many disciplines, which is already suggested by a sharp increase of visits to the website hosting the data (per Alexa.com) since the article appeared online in Science.

(Big data)²: Genealogical and genetic data unite

Particular promise of crowdsourced genealogical data lies in the combination with genetic data of the same individuals. Kaplanis et al. appreciate the research potential of combining their data with genetic and other data; hence, provide a separate, academic version of their genealogy in which individuals can consent to being identified [2]. DNA.Land, led by Yaniv Erlich, the study’s senior author, already makes use of this. This academic service allows users to upload their DTC genetic data, provide additional information, and now also identify their Geni profile [16], an option selected by thousands of DNA.Land’s 93,000 users. Unlike other third-party websites, DNA.Land operates under an academic Institutional Review Board and provides privacy control that allows users to opt in to academic research. In fact, in partnership with the National Breast Cancer Coalition (NBCC), the website collate information on family history of breast cancer and allow users to share their genomes for study by NBCC [16].

Many studies highlight the power gained from combining large-scale genealogies and matched genetic data. Among these, I find a study by deCODE genetics that just appeared to be a tour de force in this respect. The study reconstructed the genome of an 18th century Icelandic individual from that of his descendants [17]: Hans Jonatan (HJ) (1784-1827) was born in the Caribbean to an enslaved mother of African ancestry and a father of European ancestry. He migrated to Iceland, where he married and had two children. The study mined the genomes of 182 of HJ’s descendants, according to the Íslendingabók, for stretches of African ancestry, which are likely to have been inherited from HJ since African ancestry is rare in Iceland [17]. It then verified which of these stretches are likely to indeed having been inherited from HJ using the genealogy in a number of ways, including the stretches not being shared with individuals not descending from HJ and being shared, and inferred as African, in ancestors who also descent from HJ. Through these unique analyses, researchers managed to infer 38% of the part of HJ’s genome inherited from his mother and used it to deduce that she originated from Central Africa [17].

Bright and more diverse future for genetic and genealogical crowdsourcing

As we enter the era of precision medicine, the opportunities of crowdsourcing become even clearer, with distinct potential when genealogical, genetic, and medical information are integrated. Funding details of large-scale projects such as the National Institutes of Health’s All of Us program have put an effective price tag on the recruitment of each participant, their genetic data, their medical records, etc. In turn, several companies are attempting to resurrect the notion of allowing customers to lease their own information for research use. This may strengthen the potential for research based on crowdsourced data, though the data will no longer be fully
crowdfunded. One such company is the recently-established Nebula Genomics, co-founded by the omnipresent (and Harvard Medical School Professor) George Church, which will soon offer a whole-genome sequencing service at a cost that is potentially subsidized by its leasing [18]. Whether this endeavor succeeds or not, many signs point to tremendous rise in the quantity and quality of crowdsourced data, including:

(i) New customers are joining DTC genetic testing companies (and websites that combine genealogies and genetics) at a rate that is still accelerating: As the number of customers of whole-genome DTC genetic testing just crossed 16 million, it is worth noting that almost two-thirds of them joined since the beginning of 2017 [19]. Based on current rates, this number of customers is predicted to be close to 100 million by end of 2020.

(ii) Given that the crowd of crowdsourced data is predominantly from Europe and North America, which constitutes a mere 15% of the worldwide population, the increase in customer base can be multiplied by reaching out beyond these regions. North Americans and Europeans account for 85% of the profiles in Kaplanis et al., with similar bias observed among other genealogical websites and the major DTC genetic testing companies, AncestryDNA and 23andMe, which are currently selling kits mostly in the Western world. In an effort to push beyond this limitation, these companies are attempting to increase the diversity of their databases through active recruitment efforts, such as the African Genetics project launched by 23andMe in 2016. Perhaps ironically, their products are not available anywhere in Africa, limiting their access to individuals of African ancestry from non-African countries. By contrast, other services offer their products nearly worldwide, including MyHeritage, FamilyTreeDNA, and the Genographic projects (including Geno 1.0, Geno 2.0 and Geno 2.0 Next Generation). While this discrepancy is partly due to local laws and consent, the potential unleashed by integrating worldwide diversity should provide an incentive to overcome these obstacles.

(iii) Another diversity-leaning improvement should come in the form of DTC genetic testing including the X chromosome in their analysis. For consumers, it can shed new light on the nature of predicted relatedness. For medical studies, where X has been practically ignored due to analytical complications [20], its inclusion via appropriate methods can improve them in several ways and, importantly, can provide an important stepping stone toward sex-specific diagnosis and treatment, and thereby toward closing the gender disparity in medicine [20, 21].

(iv) Within a couple of years, the decreasing cost of DNA sequencing technologies will cross a threshold that will make it cost-effective for DTC genetic testing to turn to whole-genome sequencing. This will enable new services for customers, e.g., by tracing mutations along family trees, considering population-specific mutations, and flagging potentially harmful new mutations in specific genes. New opportunities for research will be even more notable and will facilitate considerable improvements in disease risk prediction, diagnosis, and treatment based on genetic and familial information.

(v) Finally, Kaplanis et al.’s exemplary study will motivate others, hopefully including companies, to compile large-scale crowdsourced-based datasets, especially as more data becomes available.

Although many fields make use of crowdsourcing, none is better positioned since all 7.5 billion of us has a family tree, DNA, traits, and medical information to share...

Our work on the Science Perspective piece was supported by grants from the NIH (R01HG006849, R01GM108880).
REFERENCES


19. Regalado, A., 2017 was the year consumer DNA testing blew up, in *MIT Technology Review*. 2018.
