HHS Public Access

Author manuscript

*Science*. Author manuscript; available in PMC 2021 May 27.

Published in final edited form as:

*Science*. 2018 April 13; 360(6385): 153–154. doi:10.1126/science.aat2634.

# Crowdsourced genealogies and genomes:

## Genealogical study provides insight into history and life span and heralds crowdsourced genetic research

**Alexandre A. Lussier**[1], **Alon Keinan**[1,2]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA.

[2]Cornell Center for Comparative and Population Genomics, Center for Vertebrate Genomics, and Center for Enervating Neuroimmune Disease, Cornell University, Ithaca, NY 14853, USA.

Genealogies are likely the first, centuries-old "big data," with their construction as old as human civilization. Recent renewed interest led to the largest genealogical websites (Ancestry.com, MyHeritage, and Geni) amassing 130 million users who generated billions of online genealogical profiles, offering ample research opportunities that would otherwise require extensive recruitment. On page 171 of this issue, Kaplanis *et al.* (1) showcase the research potential of this type of crowdsourced data, studying genealogies based on processing 86 million public Geni profiles.

An important research tool throughout human history, genealogical studies reach from anthropology to modern genetics and medicine. Of note are centuries of Icelandic genealogical enthusiasts combining family information with early-age scriptures. The founding of deCODE genetics sped up the process to create an online genealogy of 864,000 Icelanders, the Íslendingabók, by 2003. It has proven an incredible research tool, for example, in studying the relative roles of genetic heritability and shared environment in many complex diseases and other traits (2). Another unique genealogy has been constructed from historical records by the Mormon Church since 1921. Continued more recently by hundreds of thousands of volunteers, they created about half a million new profiles daily, with many connected to health care records (3)—a true crowdsourcing effort.

Entering the age of genomics, genealogy enthusiasts greeted a new tool. Direct-to-consumer (DTC) genetic testing companies all provide a service for finding relatives, while obtaining powerful, crowdsourced genome-wide data for millions of individuals. AncestryDNA and 23andMe applied their data to study migrations, structure, and admixture of U.S. populations (4, 5). However, most crowdsourced genetic research is medically driven, focusing on the genetic basis of complex traits. For example, a recent study describing how genetic risk factors are shared across many traits included analyses of 23andMe customers for 17 of the traits (6).

alon.keinan@cornell.edu.

Kaplanis *et al.* demonstrate the potential of large-scale genealogies, although with no genetic data, but at the hands of statistical and population geneticists. Extensively processing and validating genealogical data, they compiled 5.3 million genealogies, including one with 13 million individuals that often depicts at least 20 generations.

They analyze relatedness and distance at birth between married couples. Distance for most was less than 10 km before the Industrial Revolution (1750), followed by a gradual increase, which then accelerated to over 100 km after the start of the Second Industrial Revolution (1870). Average relatedness remained the same (equivalent to fourth cousins) prior to the Second Industrial Revolution, when it began decreasing in line with increasing distance. Kaplanis *et al.* postulate that recently decreased relatedness is due to shifting cultural norms, rather than increased distance, because of the inconsistent relationship between relatedness and distance. This appears concurrent with popular writing from the time, which led 13 U.S. states to pass cousin marriage prohibitions by the 1880s (7) (although more distant relatedness is in question in Kaplanis *et al.*). A related study considered 160,000 couples in the Íslendingabók to show that increased couple relatedness (equivalent to third or fourth cousins) is associated with higher fertility that is not explained by socioeconomic influences on number of offspring, and hence is claimed to have a potential biological basis (8).

The main results of Kaplanis *et al.* involve life span. The resolution of the data set allows them to discern not only that average life span decreased during World War I and World War II, but also that the decrease was larger for individuals of military age. Despite these major events, life span appears to have increased at an almost constant rate of ~4 years per generation since ~1850. 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 (9). Kaplanis *et al.* 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.

Despite extensive analyses, Kaplanis *et al.* only scratch the surface of their resource, which is publicly available, stripped of personal information. It may be interesting to reanalyze life-span factors focused on very high longevity, and revisit other questions previously studied with smaller genealogies. The resource may benefit many disciplines, with unique promise in the combination with genetic data of the same individuals, an opportunity that led to large investments in DTC genetic services by the companies with the largest genealogical websites.

DTC genetic data are not publicly available, but Kaplanis *et al.* provide an academic version of their resource where individuals can consent to being identified. It can be used on websites to which participants upload their genetic data, as Kaplanis *et al.* implemented in DNA.Land, which, for example, collates family history of breast cancer and allows users to contribute their genomes to the National Breast Cancer Coalition (10). In a recent study, deCODE genetics highlighted yet again the power of large-scale genealogies with matched

genetic data. They reconstructed an ancestor's genome by mining descendants for inherited genetic fragments, which they tested via unique genealogical analyses (11).

One critical limitation of available crowdsourced data is that the "crowd" is mostly from 15% of the worldwide population that comprises Europe and North America. The overwhelming majority of DTC genetic testing customers are from these regions, as are 85% of the profiles in the Kaplanis *et al.* study. Partly due to local laws and consent, the potential unleashed by integrating worldwide diversity should provide an incentive to overcome these obstacles. Another shortcoming is the underutilization of the X chromosome by DTC genetic companies for both customer services and medical research (12). Its inclusion via newly developed analytical methods may improve these and, importantly, provide a key step toward closing the gender disparity in disease diagnosis and treatment (12).

The era of precision medicine heralds a greater potential for crowdsourcing, with distinct opportunities when familial, genetic, and medical data are integrated. Funding details of large-scale endeavors such as the U.S. National Institutes of Health *All of Us* program have put an effective price tag on the recruitment of each participant, their genetic data, and medical records. Recently founded companies, in turn, are attempting to resurrect the option for participants to lease their data to researchers. This may increase the potential for research based on crowdsourced, although not fully crowdfunded, data.

Beyond explosive growth of DTC genetic testing services (of ~16 million current customers, almost two-thirds joined since early 2017), whole-genome sequencing will likely become a cost-effective DTC choice within 2 to 3 years. This will enable tracing, and flagging of potentially harmful, de novo mutations in families and allow crowdsourced genetic research to more substantially advance disease risk prediction, diagnosis, and treatment.

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

## REFERENCES

1. Kaplanis J et al., Science 360, 171 (2018). [PubMed: 29496957]
2. Zaitlen N et al., PLOS Genet. 9, e1003520 (2013). [PubMed: 23737753]
3. Knight S et al., Hum. Hered 81, 1 (2016). [PubMed: 27424187]
4. Han E et al., Nat. Commun 8, 14238 (2017). [PubMed: 28169989]
5. Bryc K et al., Am. J. Hum. Genet 96, 37 (2015). [PubMed: 25529636]
6. Pickrell JK et al., Nat. Genet 48, 709 (2016). [PubMed: 27182965]
7. Paul DB, Spencer HG, PLOS Biol. 6, e320 (2008).
8. Helgason A et al., Science 319, 813 (2008). [PubMed: 18258915]
9. Kerber RA et al., J. Gerontol. A Biol. Sci. Med. Sci 56, B130 (2001). [PubMed: 11253150]
10. Yuan J et al., Nat. Genet 50, 160 (2018). [PubMed: 29374253]
11. Jagadeesan A et al., Nat. Genet 50, 199 (2018). [PubMed: 29335549]
12. Editorial, Nat. Med 23, 1243 (2017). [PubMed: 29117171]