RESEARCH ARTICLE

BIG DATA

Quantitative analysis of population-scale family trees with millions of relatives

Joanna Kaplanis,[1,2]* Assaf Gordon,[1,2]* Tal Shor,[3,4] Omer Weissbrod,[5] Dan Geiger,[4] Mary Wahl,[1,2,6] Michael Gershovits,[2] Barak Markus,[2] Mona Sheikh,[2] Melissa Gymrek,[1,2,7,8,9] Gaurav Bhatia,[10,11] Daniel G. MacArthur,[7,9,10] Alkes L. Price,[10,11,12] Yaniv Erlich[1,2,3,13,14]†

Family trees have vast applications in fields as diverse as genetics, anthropology, and economics. However, the collection of extended family trees is tedious and usually relies on resources with limited geographical scope and complex data usage restrictions. We collected 86 million profiles from publicly available online data shared by genealogy enthusiasts. After extensive cleaning and validation, we obtained population-scale family trees, including a single pedigree of 13 million individuals. We leveraged the data to partition the genetic architecture of human longevity and to provide insights into the geographical dispersion of families. We also report a simple digital procedure to overlay other data sets with our resource.

Family trees are mathematical graph structures that can capture mating and parenthood among humans. As such, the edges of the trees represent potential transmission lines for a wide variety of genetic, cultural, sociodemographic, and economic factors. Quantitative genetics is built on dissecting the interplay of these factors by overlaying data on family trees and analyzing the correlation of various classes of relatives (*1*–*3*). In addition, family trees can serve as a multiplier for genetic information through study designs that leverage genotype or phenotype data from relatives (*4*–*7*), analyzing parent-of-origin effects (*8*), refining heritability measures (*9*, *10*), or improving individual risk assessment (*11*, *12*). Beyond classical genetic applications, large-scale family trees have played an important role across disciplines, including human evolution (*13*, *14*), anthropology (*15*), and economics (*16*).

Despite the range of applications, constructing population-scale family trees has been a labor-intensive process. Previous approaches mainly relied on local data repositories such as churches or vital-records offices (*14*, *17*, *18*). But these approaches have limitations (*19*, *20*): They require nontrivial resources to digitize the records and organize the data, the resulting trees are usually limited in geographical scope, and the data may be subject to strict usage protections. These challenges reduce demographic accessibility and complicate fusion with information such as genomic or health data.

Constructing and validating population-scale family trees

Here, we leveraged genealogy-driven social media data to construct population-scale family trees. To this end, we focused on Geni.com, a crowdsourcing website in the genealogy domain. Users can create individual profiles and upload family trees. The website automatically scans profiles to detect similarities and offers the option to merge the profiles when a match is detected. By merging, larger family trees are created that can be collaboratively comanaged to improve their accuracy. After obtaining relevant permissions, we downloaded approximately 86 million publicly available profiles (*21*). The input data consisted of millions of individual profiles, each of which describes a person; for 43 million of these profiles, the data also included any putative connections to other individuals in the data set. Similar to other crowdsourcing projects (*22*), a small group of participants contributed the majority of genealogy profiles (fig. S1).

We organized the profiles into graph topologies that preserve the genealogical relationships between individuals (Fig. 1A). Biology dictates that a family tree should form a directed acyclic graph, where each individual has an in-degree that is less than or equal to 2. However, 0.3% of the profiles resided in invalid biological topologies that included cycles (e.g., a person who is both the parent and child of another person) or an individual with more than two parents. We developed an automated pipeline to resolve local conflicts and prune invalid topologies (fig. S2) and benchmarked the performance of the pipeline against human genealogists (*21*). This resulted in >90% concordance between the pipeline and human decisions to resolve conflicts, thereby generating 5.3 million disjoint family trees.

The largest family tree in the processed data spanned 13 million individuals who were connected by shared ancestry and marriage (Fig. 1B). On average, the tree spanned 11 generations between each terminal descendant and their founders (fig. S3). The size of this pedigree fits what is expected as familial genealogies coalesce at a logarithmic rate compared to the size of the population (*23*).

We evaluated the structure of the tree by inspecting the genetic segregation of unilineal markers. We obtained mitochondrial DNA (mtDNA) and Y-chromosome short tandem repeat (Y-STR) haplotypes to compare multiple pairs of relatives in our graph (*21*). The mtDNA data were available for 211 lineages and spanned a total of 1768 transmission events (i.e., graph edges), whereas the Y-STR data were available for 27 lineages that spanned 324 total transmission events. Using a prior of no more than a single nonpaternity event per lineage, we estimated a nonmaternity rate of 0.3% per meiosis and nonpaternity rate of 1.9% per meiosis. This rate of nonpaternity matched previous rates of Y-chromosome studies (*24*, *25*) and the nonmaternity rate was close to historical rates of adoption of an unrelated member in the United States (*26*). Taken together, these results show that millions of genealogists can produce high-quality population-scale family trees.

Extracting demographic data

We found that life span in the Geni.com profiles was largely concordant with reports generated by traditional demographic approaches. First, we extracted demographic information from the collected profiles with exact birth and death dates, thereby avoiding the problems inherent in profiles with only year resolution for these events, such as heaping at round years (fig. S4). The data reflected historical events and trends, such as elevated death rates at military age during the American Civil War and First and Second World Wars and a reduction in child mortality during the 20th century (Fig. 2A). We compared the average life span in our collection to a worldwide historical analysis covering

[1]New York Genome Center, New York, NY 10013, USA. [2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. [3]MyHeritage, Or Yehuda 6037606, Israel. [4]Computer Science Department, Technion–Israel Institute of Technology, Haifa 3200003, Israel. [5]Computer Science Department, Weizmann Institute of Science, Rehovot 7610001, Israel. [6]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. [7]Harvard Medical School, Boston, MA 02115, USA. [8]Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02142, USA. [9]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA. [10]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [11]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA. [12]Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA. [13]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA. [14]Center for Computational Biology and Bioinformatics, Department of Systems Biology, Columbia University, New York, NY, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: erlichya@gmail.com

the years 1840 to 2000 (*27*). We found an R^2 value of 0.95 between the expected life span from historical data and the Geni data set (Fig. 2B) and a 98% concordance with historical distributions reported by the Human Mortality Database (HMD) (Fig. 2C and fig. S5).

Next, we extracted the geographic locations of life events by two approaches: an automated geoparsing pipeline and structured text manually curated and approved by genealogists (*21*) (fig. S6A). Overall, we were able to place about 16 million profiles into longitude/latitude coordinates, typically at fine-scale geographic resolution, without major differences in quality between the automated geoparsing and manual curations for subsequent analyses (fig. S6B) (*21*). The profiles were distributed across a wide range of locations in the Western world (Fig. 2D and fig. S7), with 55% from Europe and 30% from North America. We analyzed profiles in 10 cities across the globe and found that the first appearance of profiles was only after the known first settlement date for nearly all of the cities, suggesting good spatiotemporal assignment of profiles (Fig. 2E). Movie S1 presents the place of birth of individuals in the Geni data set in 5-year intervals from 1400 to 1900 along with known migration events.

We were concerned that the Geni.com profiles might suffer from certain socioeconomic ascertainment biases and therefore would not reflect the local population. To evaluate this concern, we collected ~80,000 publicly available death certificates from the Vermont Department of Health for every death in that state between 1985 and 2010. These records have extensive information for each individual, including education level, place of birth, and a cause of death in an ICD-9 code. About 1000 individuals in Geni overlapped this death certificate collection. We compared the education level, birth state, and ICD-9 code between these ~1000 Geni profiles and the entire Vermont collection. For all three parameters, we found >98% concordance between the distribution of these key sociodemographic attributes in the Geni profiles in Vermont and the entire state of Vermont (tables S1 to S3). Overall, this high level of consistency argues against severe socioeconomic ascertainment. Table S4 reports key demographic and genetic attributes for various familial relationships from parent-child via great-great-grandparents to fourth cousins.

Characterizing the genetic architecture of longevity

We leveraged the Geni data set to characterize the genetic architecture of human longevity, which exhibits complex genetics likely to involve a range of physiological and behavioral endophenotypes (*28*, *29*). Narrow-sense heritability (h^2) of longevity has been estimated to be around 15 to 30% (table S5) (*30*–*35*). Genome-wide association studies have had limited success in identifying genetic variants associated with longevity (*36*–*38*). This relatively large proportion of missing heritability can be explained by the following: (i) Longevity has nonadditive components that create upward bias in estimates of heritability (*39*), (ii) estimators of heritability are biased as a result of unaccounted environmental effects (*10*), and (iii) the trait is highly polygenic and requires larger cohorts to identify the underlying variants (*40*). We thus sought to harness our resource and build a model for the sources of genetic variance in longevity that jointly evaluates additivity, dominance, epistasis, shared household effects, spatiotemporal trends, and random noise.

We adjusted longevity to be the difference between age of death and expected life span, using a model that we trained with 3 million individuals. Our model includes spatiotemporal and sex effects and was the best among 10 different models that adjusted various spatiotemporal attributes (fig. S8). We also validated this model by estimating h^2 according to the mid-parent design (*41*) with nearly 130,000 parent-child trios. This process yielded $h^2_{\text{mid-parent}}$ = 12.2% (SE = 0.4%) (Fig. 3A), which is on the lower end but in the range of previous heritability estimates (table S5). Consistent with previous studies, we did not observe any temporal trend in mid-parent heritability (Fig. 3B).

We partitioned the source of genetic variance of longevity using more than 3 million pairs of relatives from full sibling to fourth cousin (*21*). We measured the variance explained by an

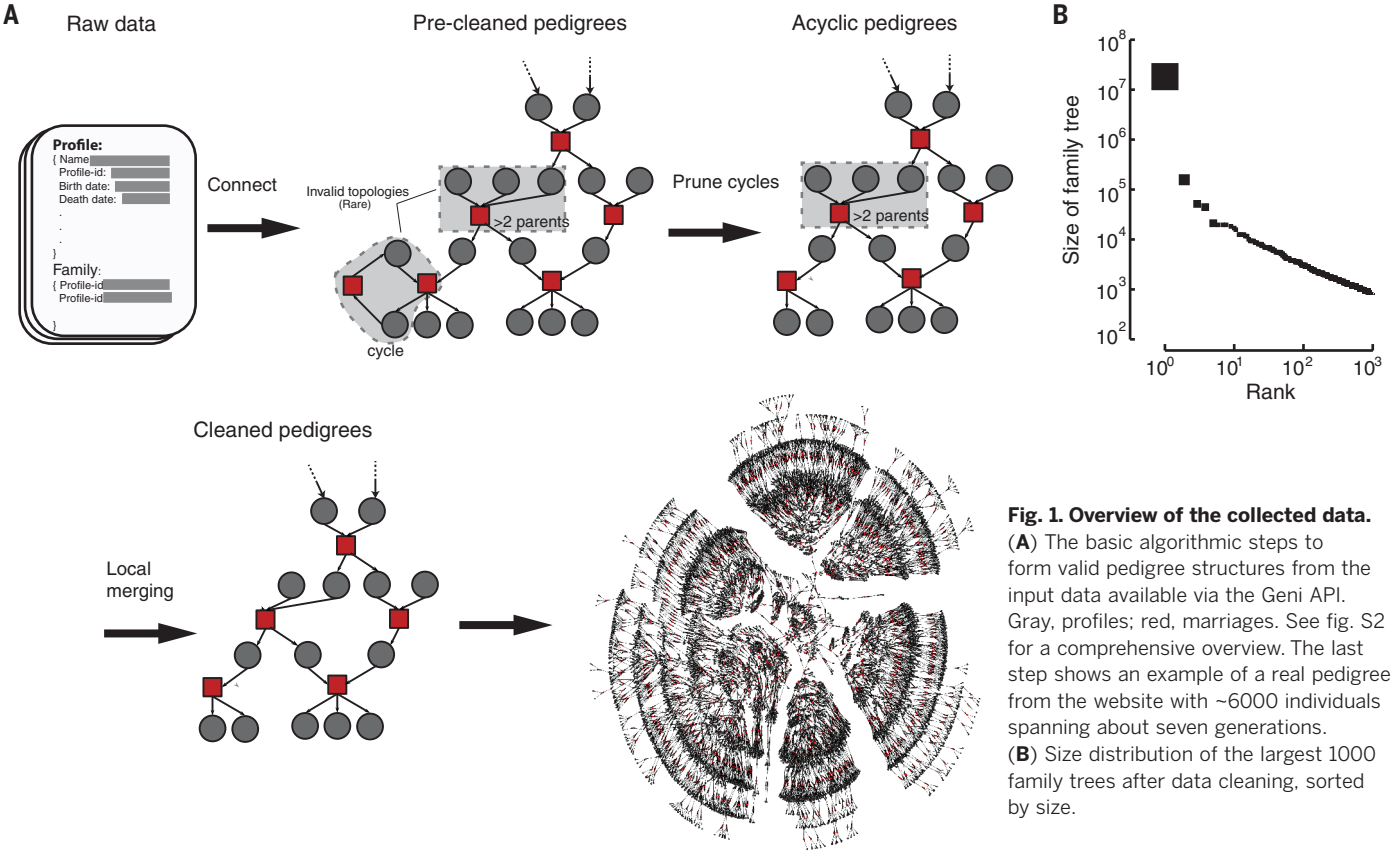

Fig. 1. Overview of the collected data. (**A**) The basic algorithmic steps to form valid pedigree structures from the input data available via the Geni API. Gray, profiles; red, marriages. See fig. S2 for a comprehensive overview. The last step shows an example of a real pedigree from the website with ~6000 individuals spanning about seven generations. (**B**) Size distribution of the largest 1000 family trees after data cleaning, sorted by size.

additive component, a pairwise epistatic model, three-way epistasis, and dominancy (Fig. 3C). These 3 million pairs were all sex-concordant to address residual sex differences not accounted for by our longevity adjustments (fig. S9) and do not include relatives who are likely to have died because of environmental catastrophes or in major wars (fig. S10); this mitigated correlations due to nongenetic factors. We also refined the genetic correlation of the relatives by considering multiple genealogical paths (figs. S11 to S13).

The analysis of longevity in these 3 million pairs of relatives showed a robust additive genetic component, a small impact of dominance, and no detectable epistasis (Fig. 3D and table S6) (*21*). Additivity was highly significant ($P_{additive} < 10^{-318}$) with an estimated $h^2_{sex\text{-}concordant/relatives} = 16.1\%$ (SE = 0.4%), similar to the heritability estimated from sex-concordant parent-child pairs, $h^2_{concordant/parent\text{-}child} = 15.0\%$ (SE = 0.4%). The maximum-likelihood estimate for dominance was around 4%, but the epistatic terms converged to zero despite the substantial amount of data. Other model selection procedures, such as mean squared error analysis and Bayesian information criterion, argued against a pervasive epistatic contribution to longevity variance in the population (*21*).

We tested the ability of our model to predict the longevity correlation of an orthogonal data set of 810 monozygotic twin pairs collected by the Danish Twin Registry (Fig. 3D) (*42*). Our inferred model for longevity accurately predicted the observed correlation of this twin cohort with 1% difference, well within the sampling error for the mean twin correlation (SE = 3.2%). We also evaluated an extensive array of additional analyses that included various adjustments for environmental components and other confounders (figs. S14 and S15) (*21*). In all cases, additivity explained 15.8 to 16.9% of the longevity estimates, dominance explained 2 to 4%, and no evidence for epistatic interactions could be detected using our procedure.

We also estimated the additive and epistatic components using a method that allows rapid estimation of variance components of extremely large relationship matrices, called sparse Cholesky factorization linear mixed models (Sci-LMM) (*43*). This method takes into account a kinship coefficient matrix of 250 million pairs of related individuals in the Geni data set and includes adjustments for population structure, sex, and year of birth. We observed an additivity of 17.8% (SE = 0.84%) and a pairwise epistatic component that was not significantly different from zero (*21*).

Taken together, our results across multiple study designs (fig. S16) indicate that the limited ability of genome-wide association studies so far to associate variants with longevity cannot be attributed to statistical epistasis. Note that this does not rule out the existence of molecular interactions between genes contributing to this trait (*44*–*47*). On the basis of a large number of data points and study designs, we measured an additive component ($h^2 \approx 16\%$) that is considerably smaller than the 25% figure that is generally cited in the literature. These results indicate that previous studies are likely to have overestimated the heritability of longevity. As such, we should lower our expectations about our ability to predict longevity from genomic data and presumably to identify causal genetic variants.

Assessment of theories of familial dispersion

Familial dispersion is a major driving force of various genetic, economic, and demographic processes (*48*). Previous work has primarily relied on vital records from a limited geographical scope (*49*, *50*) or used indirect inference from genetic data sets that mainly illuminate distant historical events (*51*).

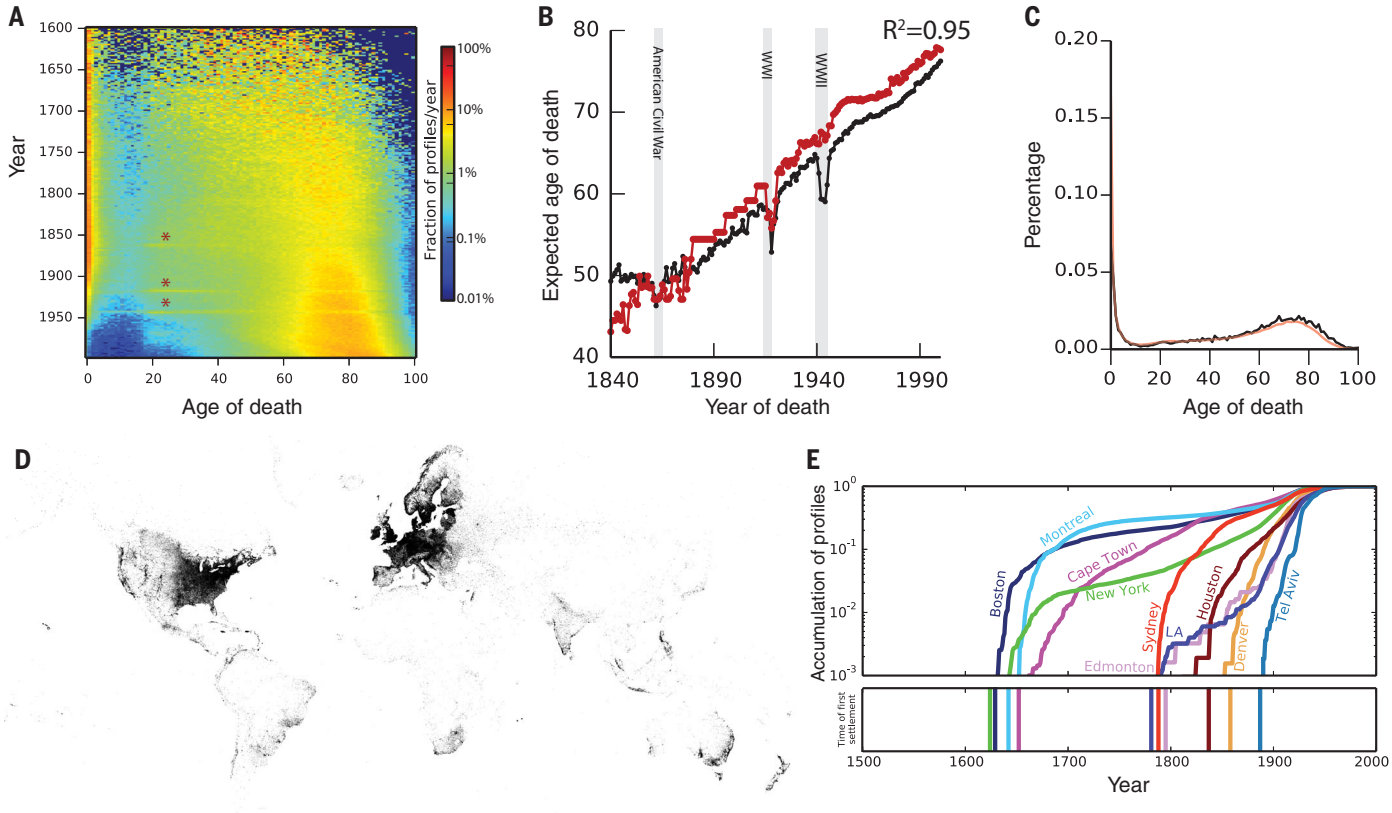

Fig. 2. Analysis and validation of demographic data. (**A**) Distribution of life expectancy per year. Colors correspond to the frequency of profiles of individuals who died at a certain age for each year. Asterisks indicate deaths at military age in the Civil War and First and Second World Wars. (**B**) Expected life span in Geni (black) and the Oeppen and Vaupel study [red (*27*)] as a function of year of death. (**C**) Comparison of the life-span distributions versus Geni (black) and HMD (red). See also fig. S5A. (**D**) Geographic distribution of the annotated place-of-birth information. Every pixel corresponds to a profile in the data set. (**E**) Validation of geographical assignment by historical trends. Top: Cumulative distribution of profiles since 1500 for each city on a logarithmic scale as a function of time. Bottom: Year of first settlement in the city.

We harnessed our resource to evaluate patterns of human migration. First, we analyzed sex-specific migration patterns (*21*) to resolve conflicting results regarding sex bias in human migration (*52*). Our results indicate that in Western societies, females migrate more than males but over shorter distances. Median mother-child distances were significantly larger than median father-child distances by a factor of 1.6 (Wilcox, one-tailed, $P < 10^{-90}$) (Fig. 4A). This trend appeared throughout the 300 years of our analysis window, including in the most recent birth cohort, and was observed both in North American duos (Wilcox, one-tailed, $P < 10^{-23}$) and European duos (Wilcox, one-tailed, $P < 10^{-87}$). On the other hand, we found that average mother-child distances (fig. S17) were significantly shorter than average father-child distances (t test, $P < 10^{-90}$), which suggests that long-range migration events are biased toward males. Consistent with this pattern, fathers displayed a significantly ($P < 10^{-83}$) higher frequency than mothers to be born in a different country than their offspring (Fig. 4B). Again, this pattern was evident when restricting the data to North American or European duos. Taken together, males and females in Western societies show different migration distributions; patrilocality occurs only in relatively local migration events, and large-scale events that usually involve a change of country are more common in males than in females.

Next, we inspected the marital radius (the distance between mates' places of birth) and its

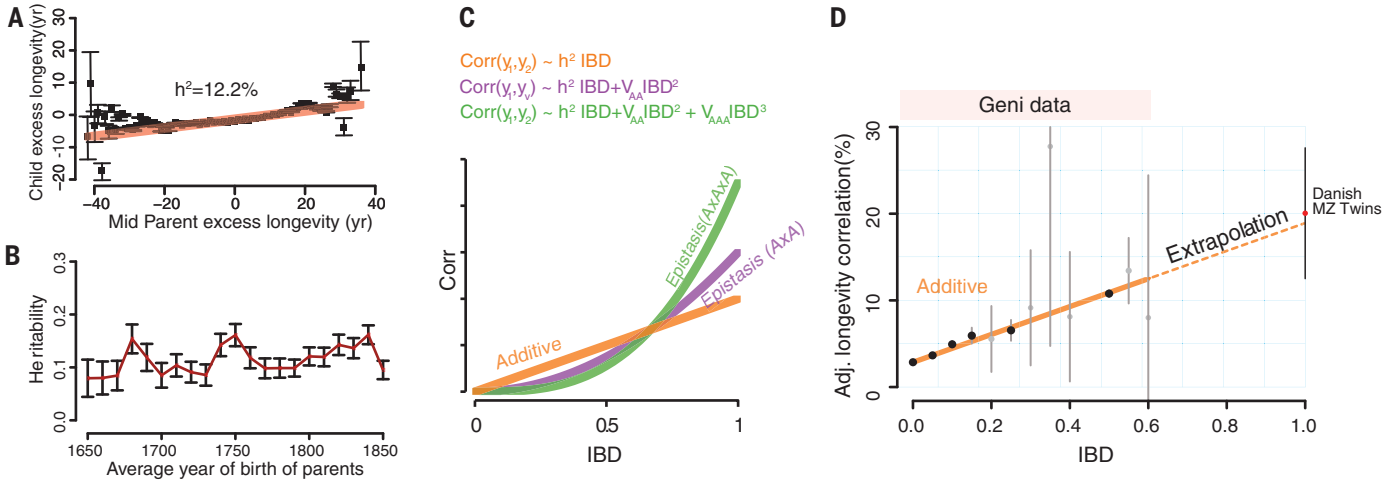

Fig. 3. The genetic architecture of longevity. (**A**) Regression (red) of child longevity on its mid-parent longevity (defined as difference between age of death and expected life span). Black squares, average longevity of children binned by the mid-parent value; gray bars, estimated 95% confidence interval (CI). (**B**) Estimated narrow-sense heritability (red) with 95% confidence intervals (black bars) obtained by the mid-parent design stratified by the average decade of birth of the parents. (**C**) Correlation of a trait as a function of IBD under strict additive (h^2, orange), squared (V_{AA}, purple), and cubic (V_{AAA}, green) epistasis architectures after dormancy adjustments. (**D**) Average longevity correlation as a function of IBD (black circles) grouped in 5% increments (gray: 95% CI) after adjusting for dominancy. A dashed line denotes the extrapolation of the models toward monozygotic twins from the Danish Twin Registry (red circle).

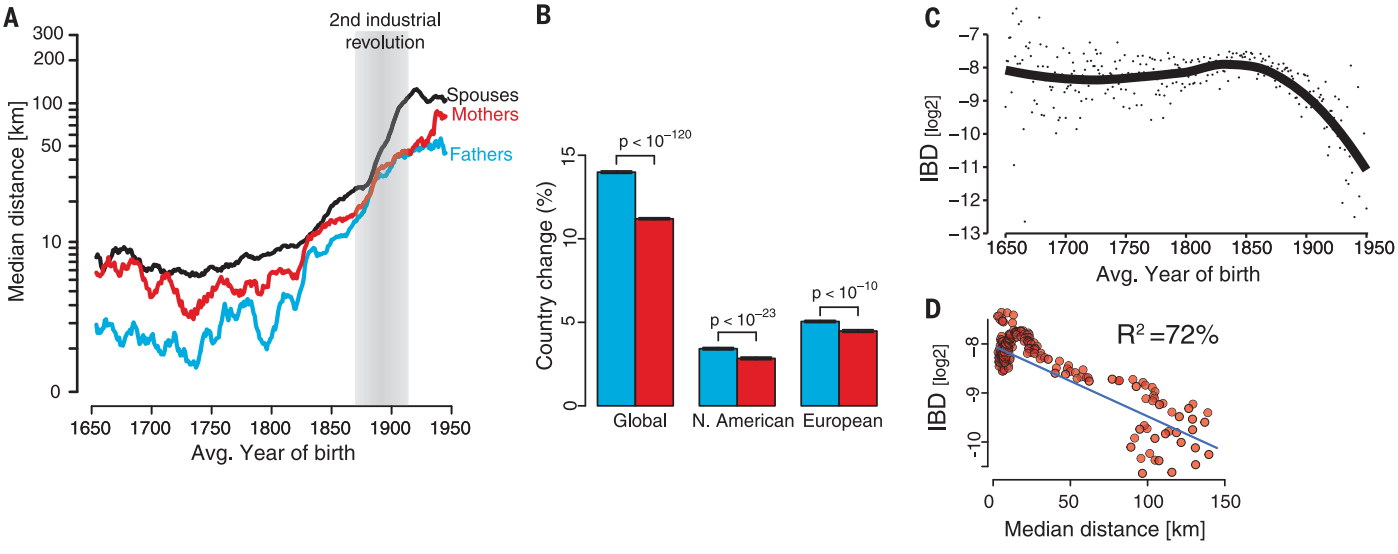

Fig. 4. Analysis of familial dispersion. (**A**) Median distance [$\log_{10}(x + 1)$] of father-offspring places of birth (cyan), mother-offspring (red), and marital radius (black) as a function of time (average year of birth). (**B**) Rate of change in the country of birth for father-offspring (cyan) or mother-offspring (red) stratified by major geographic areas. (**C**) Average IBD ($\log_2$) between couples as a function of average year of birth. Individual dots represent the measured average per year; the black line denotes the smooth trend using locally weighted regression. (**D**) IBD of couples as a function of marital radius. Each dot represents a year between 1650 to 1950. The blue line denotes the best linear regression line in log-log space.

effect on the genetic relatedness of couples (*21*). The isolation-by-distance theory of Malécot predicts that increases in the marital radius should exponentially decrease the genetic relatedness of individuals (*53*). But the magnitude of these forces is also a function of factors such as taboos against cousin marriages (*54*).

We started by analyzing temporal changes in the birth locations of couples in our cohort. Before the Industrial Revolution (earlier than 1750), most marriages occurred between people born only 10 km from each other (Fig. 4A, black line). Similar patterns were found when analyzing European-born individuals (fig. S18) or North American–born individuals (fig. S19). After the beginning of the second Industrial Revolution (1870), the marital radius rapidly increased and reached ~100 km for most marriages in the birth cohort in 1950. Next, we analyzed the expected identity-by-descent (IBD) of couples as measured by tracing their genealogical ties (Fig. 4C). Between 1650 and 1850, the average IBD of couples was relatively stable and on the order of fourth cousins, whereas IBD exhibited a rapid decrease after 1850. Overall, the median marital radius for each year showed a strong correlation (R^2 = 72%) with the expected IBD between couples. Every 70-km increase in the marital radius correlated with a decrease in the genetic relatedness of couples by one meiosis event (Fig. 4D). This correlation matches previous isolation-by-distance forces in continental regions (*55*). However, this trend is not consistent over time and exhibits three phases. For the pre-1800 birth cohorts, the correlation between marital distance and IBD was insignificant ($P > 0.2$) and weak (R^2 = 0.7%) (fig. S20A). Couples born around 1800 to 1850 showed a doubling of their marital distance, from 8 km in 1800 to 19 km in 1850. Marriages usually occur about 20 to 25 years after birth, and around this time (1820 to 1875) rapid transportation changes took place, such as the advent of railroad travel in most of Europe and the United States. However, the increase in marital distance was significantly ($P < 10^{-13}$) coupled with an increase in genetic relatedness, contrary to the isolation-by-distance theory (fig. S20B). Only for the cohorts born after 1850 did the data match (R^2 = 80%) the theoretical model of isolation by distance (fig. S20C).

Taken together, the data show a 50-year lag between the advent of increased familial dispersion and the decline of genetic relatedness between couples. During this time, individuals continued to marry relatives despite the increased distance. From these results, we hypothesize that changes in 19th-century transportation were not the primary cause for decreased consanguinity. Rather, our results suggest that shifting cultural factors played a more important role in the recent reduction of genetic relatedness of couples in Western societies.

Discussion

In this work, we leveraged genealogy-driven media to build a data set of human pedigrees of massive scale that covers nearly every country in the Western world. Multiple validation procedures indicated that it is possible to obtain a data set that has similar quality to traditionally collected studies, but at much greater scale and lower cost.

We envision that this and similar large data sets can address quantitative aspects of human families, including genetics, anthropology, public health, and economics. Our tree and demographic data are available in a de-identified format, enabling static analysis of the Geni data set. We also offer a dynamic method that enables fusing other data sets with our data, based on digital consent of participants using the Geni application programming interface (API) (fig. S21) (*21*). We have been using this one-click mechanism to overlay thousands of genomes with family trees on DNA.Land (*56*). Other projects can use a similar strategy to add large pedigrees to their existing data collection.

More generally, similar to previous studies (*57*, *58*), our work demonstrates the synergistic power of a collaboration between basic research and consumer genetic genealogy data sets. With ever-growing digitization of humanity and the rise of consumer genetics (*59*), we believe that such collaborative efforts can be a valuable path to reach the scale of information needed to address fundamental questions in biomedical research.

REFERENCES AND NOTES

1. R. A. Fisher, *Trans. R. Soc. Edinb.* **52**, 399–433 (1919).
2. S. Wright, *J. Agric. Res.* **20**, 557–585 (1921).
3. A. Tenesa, C. S. Haley, *Nat. Rev. Genet.* **14**, 139–149 (2013).
4. A. Kong *et al*., *Nat. Genet.* **40**, 1068–1075 (2008).
5. J. K. Lowe *et al*., *PLOS Genet.* **5**, e1000365 (2009).
6. D. F. Gudbjartsson *et al*., *Nat. Genet.* **47**, 435–444 (2015).
7. J. Z. Liu, Y. Erlich, J. K. Pickrell, *Nat. Genet.* **49**, 325–331 (2017).
8. A. Kong *et al*., *Nature* **462**, 868–874 (2009).
9. C. Ober, M. Abney, M. S. McPeek, *Am. J. Hum. Genet.* **69**, 1068–1079 (2001).
10. N. Zaitlen *et al*., *PLOS Genet.* **9**, e1003520 (2013).
11. R. Valdez, P. W. Yoon, N. Qureshi, R. F. Green, M. J. Khoury, *Annu. Rev. Public Health* **31**, 69–87 (2010).
12. C. B. Do, D. A. Hinds, U. Francke, N. Eriksson, *PLOS Genet.* **8**, e1002973 (2012).
13. M. Lahdenperä, V. Lummaa, S. Helle, M. Tremblay, A. F. Russell, *Nature* **428**, 178–181 (2004).
14. C. Moreau *et al*., *Science* **334**, 1148–1150 (2011).
15. A. Helgason, S. Pálsson, D. F. Gudbjartsson, T. Kristjánsson, K. Stefánsson, *Science* **319**, 813–816 (2008).
16. J. Modalsli, "Multigenerational persistence: Evidence from 146 years of administrative data" (Statistics Norway, 2016); https://EconPapers.repec.org/RePEc:ssb:dispap:850.
17. J. R. Gulcher, K. Stefansson, in *Encyclopedia of Life Sciences* (Wiley, 2001).
18. L. A. Cannon-Albright, *Hum. Hered.* **65**, 209–220 (2008).
19. L. A. Cannon-Albright, in *AMIA Annual Symposium Proceedings* (American Medical Informatics Association, 2006), p. 1161.
20. V. Stefansdottir *et al*., *J. Community Genet.* **4**, 1–7 (2013).
21. See supplementary materials.
22. A. Kittur, E. Chi, B. A. Pendleton, B. Suh, T. Mytkowicz, *World Wide Web* **1**, 19 (2007).
23. J. T. Chang, *Adv. Appl. Probab.* **31**, 1002–1026 (1999).
24. K. Anderson, *Curr. Anthropol.* **47**, 513–520 (2006).
25. T. E. King, M. A. Jobling, *Mol. Biol. Evol.* **26**, 1093–1102 (2009).
26. P. Maza, *Child Welf. Res. Notes* **9**, 1–11 (1984).
27. J. Oeppen, J. W. Vaupel, *Science* **296**, 1029–1031 (2002).
28. P. Sebastiani, T. T. Perls, *Front. Genet.* **3**, 277 (2012).
29. R. E. Marioni *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **113**, 13366–13371 (2016).
30. P. Philippe, J. M. Opitz, *Am. J. Med. Genet.* **2**, 121–129 (1978).
31. P. J. Mayer, *Am. J. Hum. Biol.* **3**, 49–58 (1991).
32. B. Ljungquist, S. Berg, J. Lanke, G. E. McClearn, N. L. Pedersen, *J. Gerontol. A* **53**, M441–M446 (1998).
33. A. M. Herskind *et al*., *Hum. Genet.* **97**, 319–323 (1996).
34. B. D. Mitchell *et al*., *Am. J. Med. Genet.* **102**, 346–352 (2001).
35. R. A. Kerber, E. O'Brien, K. R. Smith, R. M. Cawthon, *J. Gerontol. A* **56**, B130–B139 (2001).
36. P. Sebastiani *et al*., *PLOS ONE* **7**, e29848 (2012).
37. J. Deelen *et al*., *Hum. Mol. Genet.* **23**, 4420–4432 (2014).
38. G. A. Erikson *et al*., *Cell* **165**, 1002–1011 (2016).
39. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1193–1198 (2012).
40. E. A. Boyle, Y. I. Li, J. K. Pritchard, *Cell* **169**, 1177–1186 (2017).
41. P. M. Visscher, W. G. Hill, N. R. Wray, *Nat. Rev. Genet.* **9**, 255–266 (2008).
42. A. Skytthe, K. O. Kyvik, N. V. Holm, K. Christensen, *Scand. J. Public Health* **39** (suppl.), 75–78 (2011).
43. T. Shor, D. Geiger, Y. Erlich, O. Weissbrod, BioRxiv 256396 [Preprint]. 30 January 2018. https://doi.org/10.1101/256396.
44. W. Li, J. Reich, *Hum. Hered.* **50**, 334–349 (2000).
45. P. C. Phillips, *Nat. Rev. Genet.* **9**, 855–867 (2008).
46. H. J. Cordell, *Nat. Rev. Genet.* **10**, 392–404 (2009).
47. W.-H. Wei, G. Hemani, C. S. Haley, *Nat. Rev. Genet.* **15**, 722–733 (2014).
48. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton Univ. Press, 1994).
49. E. M. Wijsman, L. L. Cavalli-Sforza, *Annu. Rev. Ecol. Syst.* **15**, 279–301 (1984).
50. R. Labouriau, A. Amorim, *Genetics* **178**, 601–603 (2008).
51. K. R. Veeramah, M. F. Hammer, *Nat. Rev. Genet.* **15**, 149–162 (2014).
52. L. J. Lawson Handley, N. Perrin, *Mol. Ecol.* **16**, 1559–1578 (2007).
53. G. Malécot, *The Mathematics of Heredity* (Freeman, 1970).
54. L. L. Cavalli-Sforza, A. Moroni, G. Zei, *Consanguinity, Inbreeding, and Genetic Drift in Italy* (Princeton Univ. Press, 2004).
55. J. H. Relethford, E. R. Brennan, *Hum. Biol.* **54**, 315–327 (1982).
56. J. Yuan *et al*., *Nat. Genet.* **50**, 160–165 (2018).
57. J. K. Pickrell *et al*., *Nat. Genet.* **48**, 709–717 (2016).
58. E. Han *et al*., *Nat. Commun.* **8**, 14238 (2017).
59. R. Khan, D. Mittelman, *Genome Biol.* **14**, 139 (2013).

ACKNOWLEDGMENTS

We thank D. Zielinski, G. Japhet, and J. Novembre for valuable comments, the Erlich lab members for constant support in pursuing this project, and the Vermont Health Department for providing all death certificates. This study was supported by a generous gift from Andria and Paul Heafy (Y.E.), the Burroughs Wellcome Fund Career Awards at the Scientific Interface (Y.E.), the Broad Institute's SPARC: Catalytic Funding for Novel Collaborative Projects award (Y.E. and D.G.M.), NIH grants R01 MH101244 and R03 HG006731 (A.L.P.), and Israeli Science Foundation grant 1678/12 (D.G.). Author contributions: A.G. and Y.E. conducted the downloading, indexing, and organizing of the data; J.K., A.G., M.W., B.M., M.Ge., M.S., and Y.E. developed the procedures to clean the family trees and extract demographic information; J.K., T.S., O.W., D.G., M.Gy., G.B., D.G.M., A.L.P., and Y.E. were involved in analyzing the genetic architecture of longevity; J.K., M.W., and Y.E. conducted the analysis of human migration; and J.K., T.S., O.W., D.G.M., A.L.P., and Y.E. wrote the manuscript. T.S. and Y.E. became employees of MyHeritage.com, the parent company of Geni.com, during the course of this study. The other authors do not declare relevant competing interests. The Geni data set without names is available from Y.E. under the terms described on FamiLinx.org. The code for the API integration is available at https://github.com/TeamErlich/geni-integration-example, the code for Sci-LMM is available at https://github.com/TalShor/SciLMM, and the code to download Geni profiles is available at https://github.com/erlichya/geni-download. The Human Mortality Database (HMD) is available at www.mortality.org. The Danish Twin Registry (DTR) data are available upon request from the University of Southern Denmark (www.sdu.dk/en/om_sdu/institutter_centre/ist_sundhedstjenesteforsk/centre/dtr). The findings, opinions, and recommendations expressed herein are those of the authors and are not necessarily those of the DTR. The Vermont Death Certificate collection was obtained upon request from the Chief of Public Health Statistics, Vermont Department of Health (www.healthvermont.gov/stats).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/360/6385/171/suppl/DC1
Materials and Methods
Figs. S1 to S21
Tables S1 to S6
Movie S1
References (*60*–*79*)

7 February 2017; resubmitted 2 November 2017
Accepted 7 February 2018
Published online 1 March 2018
10.1126/science.aam9309

Science

Quantitative analysis of population-scale family trees with millions of relatives

Joanna Kaplanis, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, Barak Markus, Mona Sheikh, Melissa Gymrek, Gaurav Bhatia, Daniel G. MacArthur, Alkes L. Price and Yaniv Erlich

Science **360** (6385), 171-175.
DOI: 10.1126/science.aam9309originally published online March 1, 2018

Quantitative analysis of millions of relatives

Human relationships, as documented by family trees, can elucidate the heritability of a host of medical and biological parameters. Kaplanis *et al.* collected 86 million publicly available profiles from a crowd-sourced genealogy website and used them to examine the genetic architecture of human longevity and migration patterns (see the Perspective by Lussier and Keinan). Various models of inheritance suggested that life span is predominantly attributable to additive genetic effects, with a smaller component from dominant genetic inheritance. The data also suggested that relatedness between individuals is less attributable to advances in human transportation than to cultural changes.

Science, this issue p. 171; see also p. 153