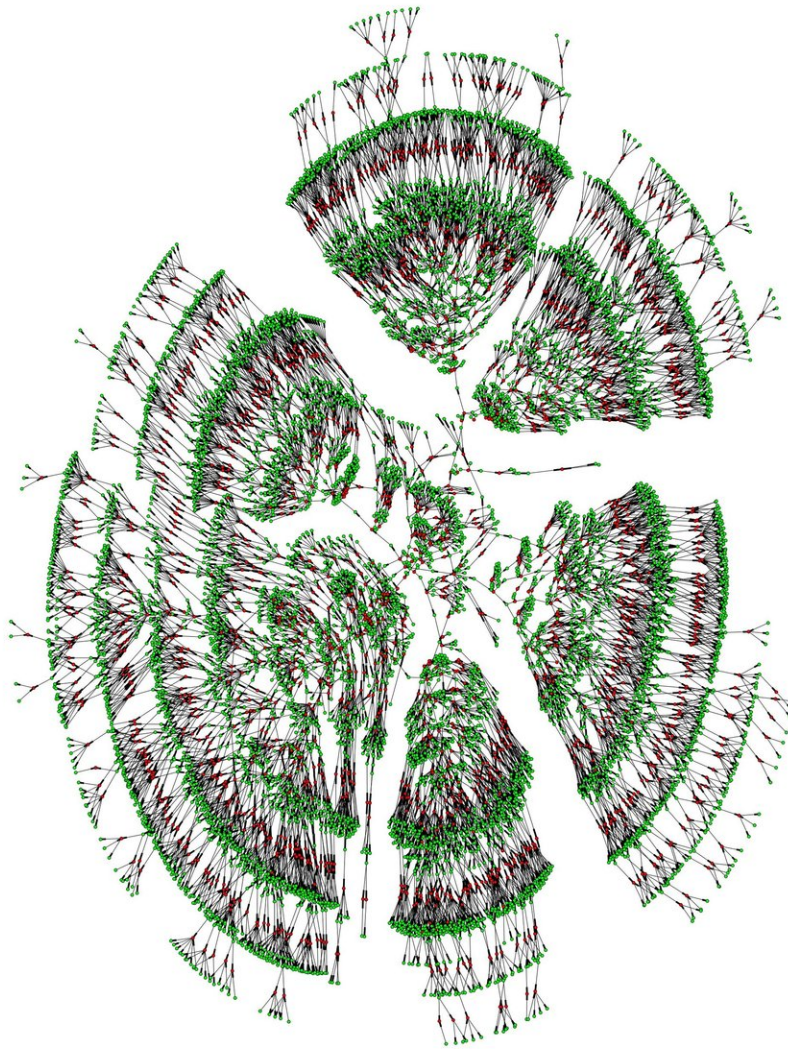


# Crowdsourced family tree yields new insights about humanity

March 1 2018

---



In the above 6,000 person family tree cleaned and organized using graph theory, individuals spanning seven generations are represented in green, with their marital links in red. Credit: Columbia University

Thanksgiving gatherings could get bigger—a lot bigger—as science uncovers the familial bonds that bind us. From millions of interconnected online genealogy profiles, researchers have amassed the largest, scientifically-vetted family tree to date, which at 13 million people, is slightly bigger than a nation the size of Cuba or Belgium. Published in the journal *Science*, the new dataset offers fresh insights into the last 500 years of marriage and migration in Europe and North America, and the role of genes in longevity.

"Through the hard work of many genealogists curious about their family history, we crowdsourced an enormous family tree and boom, came up with something unique," said the study's senior author, Yaniv Erlich, a computer scientist at Columbia University and Chief Science Officer at MyHeritage, a genealogy and DNA testing company that owns Geni.com, the platform that hosts the data used in the study. "We hope that this [dataset](#) can be useful to scientists researching a range of other topics."

The researchers downloaded 86 million public profiles from Geni.com, one of the world's largest collaborative genealogy websites, and used mathematical graph theory to clean and organize the data. What emerged among other smaller family trees was a single tree of 13 million people spanning an average of 11 generations. Theoretically, they'd need to go back another 65 generations to converge on one common ancestor and complete the tree. Still, the dataset represents a milestone by moving family-history searches from newspaper obituaries and church archives into the digital era, making population-level investigations possible. The

researchers also make it easy to overlay other datasets to study a range of socioeconomic trends at scale.

"It's an exciting moment for citizen science," said Melinda Mills, a demographer at University of Oxford who was not involved in the study. "It demonstrates how millions of regular people in the form of genealogy enthusiasts can make a difference to science. Power to the people!"

The dataset details when and where each individual was born and died, and mirrors the demographics of Geni.com individuals, with 85 percent of profiles originating from Europe and North America. The researchers verified that the dataset was representative of the general U.S. population's education level by cross-checking a subset of Vermont Geni.com profiles against the state's detailed death registry.

"The reconstructed pedigrees show that we are all related to each other," said Peter Visscher, a quantitative geneticist at University of Queensland who was not involved in the study. "This fact is known from basic population history principles, but what the authors have achieved is still very impressive."

**Marriage, Migration and Genetic Relatedness** Industrialization profoundly altered work and family life, and these trends coincide with shifting marriage choices in the data. Before 1750, most Americans found a spouse within six miles (10 kilometers) of where they were born, but for those born in 1950, that distance had stretched to about 60 miles (100 kilometers), the researchers found. "It became harder to find the love of your life," Erlich jokes.

Before 1850, marrying in the family was common—to someone who was, on average, a fourth cousin, compared to seventh cousins today, the researchers found. Curiously, the researchers found that between 1800 and 1850, people traveled farther than ever to find a mate—nearly 12

miles (19 kilometers) on average —but were more likely to marry a fourth cousin or closer. Changing social norms, rather than rising mobility, may have led people to shun close kin as marriage partners, they hypothesize.

In a related observation, they found that women in Europe and North America have migrated more than men over the last 300 years, but when men did migrate, they traveled significantly farther on average.

**Genes and Longevity** To try and untangle the role of nature and nurture in longevity, the researchers built a model and trained it on a dataset of 3 million relatives born between 1600 and 1910 who had lived past the age of 30. They excluded twins, individuals who died in the U.S. Civil War, World War I and II, or in a natural disaster (inferred if relatives died within 10 days of each other).

They compared each individual's lifespan to that of their relatives and their degree of separation and found that genes explained about 16 percent of the longevity variation seen in their data—on the low end of previous estimates which have ranged from about 15 percent to 30 percent.

The results indicate that good longevity genes can extend someone's life by an average of five years, said Erlich. "That's not a lot," he adds. "Previous studies have shown that smoking takes 10 years off of your life. That means some life choices could matter a lot more than genetics."

Significantly, the study also shows that the genes that influence longevity act independently rather than interacting with each other, a phenomenon called epistasis. Some scientists have used epistasis to explain why large-scale genomic studies have so far failed to find the genes that encode complex traits like intelligence or longevity.

If some genetic variants act together to influence longevity, the researchers would have seen a greater correlation among closely related individuals who share more DNA, and thus more genetic interactions. However, they found a linear link between longevity and [genetic relatedness](#), ruling out widespread epistasis.

"This is important in the field because epistasis has been proposed as a source of 'missing heritability,'" said the study's lead author, Joanna Thornycroft, a former graduate student at the Whitehead Institute for Biomedical Research, now at Wellcome Sanger Institute.

Adds Visscher: "This is entirely in line with theory and previous inference from SNP [variant] data, yet for some reason many researchers in human genetics and epidemiology continue to believe that there is a lot of non-additive genetic variation for common diseases and quantitative traits."

The dataset is available for academic research via FamiLinx.org, a website created by Erlich and his colleagues. Though FamiLinx data is anonymized, curious readers can check Geni.com to see if a family member may have added them there. If so, there is a good chance that they may have made it into the 13 million-person family tree.

In addition to his position at MyHeritage, a company that allows consumers to discover their [family history](#) through genetic tests and its genealogy platform, Erlich is a computer science professor at Columbia Engineering, a member of Columbia's Data Science Institute, and an adjunct core member of the New York Genome Center (NYGC).

Other study authors are Assaf Gordon, of NYGC and the Whitehead Institute; Tal Shor, of MyHeritage and Technion; Omer Weissbrod of Israel's Weizmann Institute of Science; Dan Geiger of Technion; Mary Wahl of Whitehead Institute, NYGC and Harvard; Michael Gershovits,

Barak Markus and Mona Sheikh of Whitehead Institute; Melissa Gymrek of University of California at San Diego; and Gaurav Bhatia, Daniel MacArthur and Alkes Price of Harvard and the Broad Institute.

**More information:** J. Kaplanis et al., "Quantitative analysis of population-scale family trees with millions of relatives," *Science* (2017). [science.sciencemag.org/cgi/doi ... 1126/science.aam9309](https://science.sciencemag.org/cgi/doi/10.1126/science.aam9309)

Provided by Columbia University School of Engineering and Applied Science

Citation: Crowdsourced family tree yields new insights about humanity (2018, March 1) retrieved 26 April 2024 from <https://phys.org/news/2018-03-crowdsourced-family-tree-yields-insights.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.