Nature Reviews Genetics | Published online 26 Mar 2018; doi:10.1038/nrg.2018.16

Cienpies Design/Alamy

HUMAN GENETICS

Population-scale family trees from publicly available data

Constructing extended family trees is costly and time-consuming, especially for larger pedigrees, such as those at the population scale. Now, a study published in *Science* reports the collection of population-scale family trees using publicly available online data from a genealogy website.

First, the researchers obtained 86 million records (of which 43 million had familial information) from Geni.com, a crowd-sourced genealogy website, and refined graphs of the relationships between profiles, using an automated pipeline to remove any invalid topologies (for example, an individual having more than two parents). This automated method had >90% concordance with decisions made by genealogists about invalid topologies. The largest tree in the data set included 13 million people, spanning on average 11 generations between founders and their last descendant. The quality of the tree was also confirmed by assessing genetic segregation patterns of unilineal markers, such as mitochondrial DNA (matrilineal transmission) and short tandem repeats on the Y chromosome (patrilineal transmission).

Next, the authors extracted demographic data from the online profiles, including dates of birth and death and the geographic locations of individuals. Analysis of birth and death dates confirmed that they correlated with historical events (for example, elevated death rates occurred in major wars), and average lifespans closely matched those from historical data. The first historical appearance of individuals in major cities corresponded well with the date of their foundation, establishing the quality of the location data. The authors created a model to detect effects of genetic variance (including additivity, dominance, epistasis and so on) on longevity, which exhibits complex genetics that has been difficult to dissect using genome-wide association studies. Analysis of longevity in 3 million pairs of relatives revealed a prominent additive genetic component (~16%), a lower contribution from dominance (2–4%) and no detectable epistasis. The additive genetic component in these new data was lower than previous estimates (~25%), suggesting that detecting causal genetic variants from genomic data might be more difficult than anticipated.

The authors used their data set to analyse human migration patterns, finding that females in Western societies migrate more than males, albeit over shorter distances. Consequently, large-scale transnational migration events are more likely to involve males than females. The authors also analysed the marital radius — the distance between the birth places of spouses — which, as it increases, is predicted (from the 'isolation by distance' theory) to result in decreased genetic relatedness of couples. The average marital radius changed as expected for migration events during major sociopolitical events, such as increases following the advent of the Industrial Revolutions in 1750 and 1870. Interestingly, there was a 50-year lag between the increase in marital radius after 1800 and the predicted decrease in genetic relatedness of couples, which the authors attribute to changing cultural norms rather than the advent of long-range transportation in the early 19th century.

This study demonstrates the usefulness of collaborations between researchers and the public (via online databases). Large pedigrees created in this way should prove useful for the quantitative assessment of genetics and aspects of public health, following overlay of genome sequencing data onto these trees (the de-identified data are available on FamiLinx.org).

Grant Otto

> "the largest tree … included 13 million people, spanning on average 11 generations"

ORIGINAL ARTICLE Kaplanis, J. *et al.* Quantitative analysis of population-scale family trees with millions of relatives. *Science* https://doi.org/10.1126/science.aam9309 (2018)

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.