POPULATION GENETICS

From a database of genomes to a forest of evolutionary trees

Inferring adaptation, migration and population history would be profoundly easier if we could use the genomes that we sequence to infer the true genealogical history of each locus. Two new papers bring us close to achieving this goal.

Kelley Harris

Every gene has a tree-like genealogy that describes how living individuals inherited their DNA through a network of common ancestors. The knowledge encoded in such trees can be transformative; a lineage tree of mitochondria famously proved the recent African origin theory of human evolution[1]. However, there are limits to what we can learn from mitochondrial DNA alone, because it traces only matrilineal history and contains no information about the effect of natural selection on nuclear genes. The histories of nuclear genomic loci encode a fuller picture of evolution but are more difficult to reconstruct from the data, owing to complexity introduced by sex and recombination. In this issue, two groups from Oxford University present powerful new methods for inferring the genealogies of all loci in the genome[2,3].

Speidel et al. describe a new method called 'Relate' that can quickly estimate whole-genome genealogies: the joint collection of genealogical trees across all loci in a sample of thousands of genomes[3,4]. In 4 days, the authors were able to estimate the history of the humans sequenced by the 1000 Genomes Consortium, pooling information across nearby polymorphic sites to estimate how long ago each pair of individuals shared a common ancestor at each locus. It can be computationally expensive to estimate when recombinations between different ancestry groups cause nearby sites to have different histories (Fig. 1), and the previous methods that were able to estimate nuclear genealogies with high accuracy ran only quickly enough to infer genealogies from a few tens of whole genomes at most[5]. For the sake of efficiency, Relate does not calculate the likelihood of every sequence of trees that might underlie the data, but benchmarking studies indicate that the genealogies that it estimates from simulated data are very close to the true ones.

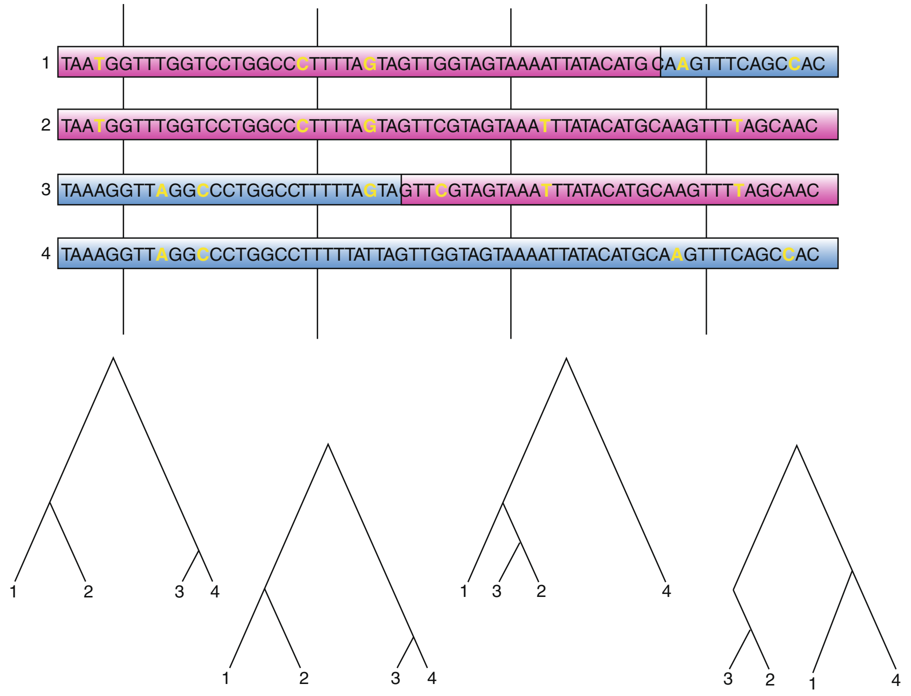
1
TAATGGTTTGGTCCTGGCCCTTTTAGTAGTTGGTAGTAAAATTATACATG CAAGTTTCAGCCAC
2
TAATGGTTTGGTCCTGGCCCTTTTAGTAGTTCGTAGTAAATTTATACATGCAAGTTTTAGCAAC
3
TAAAGGTTAGGCCCTGGCCTTTTTAGTAGTTCGTAGTAAATTTATACATGCAAGTTTTAGCAAC
4
TAAAGGTTAGGCCCTGGCCTTTTTATTAGTTGGTAGTAAAATTATACATGCAAGTTTCAGCCAC
1
2
3
4
1
2
3
4
1
3
2
4
3
2
1
4

Fig. 1 | An example ancestral genealogy sequence from an admixed population containing two ancestry types. Each chromosome is a mosaic of red and blue ancestry types, and the recombination events between ancestries cause the genealogy of these four lineages to vary along this sequence alignment. Polymorphisms in yellow provide information about the history of these sequences.

Ready for a million genomes

The 1000 Genomes Project was the definitive human genome repository of the 2010s, but some new datasets, such as the UK Biobank, are starting to dwarf it in scale. Kelleher et al. anticipate future datasets of millions of genomes and present a tree-building method that will scale to match[2]. The core of the method, called 'tsinfer', is a data structure called a succinct tree sequence, which not only elucidates genealogical structure but also compresses the data orders of magnitude more efficiently than a standard gzipped VCF file. tsinfer works even faster than Relate and scales to even larger datasets, although a tradeoff is that it does not estimate the ages of extinct ancestors. Speidel et al. note that a future hybrid method could achieve the best of both worlds by running tsinfer to hierarchically cluster individuals by relatedness and then using part of the Relate algorithm to infer ancestor ages and refine tree topology accuracy[3].

Streamlining population genetic inference

If we could perfectly infer the genealogy of each site in a sample of genomes, the difficulty level of most problems in evolutionary genomics would be fundamentally decreased. Whenever we want to determine whether something occurred during a population's history, such as whether a particular gene variant helped us adapt to a new environment, it is usually easiest to start by determining how our hypothesis would affect gene genealogies. For example, natural selection tends to make gene trees more unbalanced, with shorter branches than we would expect to see in a neutrally evolving population. After making such a connection between evolutionary processes and tree shape, we must then determine how the predicted change in gene genealogies would translate into a detectable effect on genetic data, and we usually lose information in the process. By working toward eliminating the need for this second step, Speidel et al. and Kelleher et al. have been able to streamline the study of mutation, selection and genetic drift: the trifecta of forces that cause genetic variation to be created, maintained and destroyed.

Genealogical inference explicitly underlies several commonly used methods for inferring how population sizes have changed over time, including the pairwise sequentially Markovian coalescent (PSMC)[6]. PSMC and its successor, the multiple sequentially Markovian coalescent (MSMC)[7], are commonly used because they enable easy inference of the demographic histories of a wide variety of organisms; however, they can analyze fewer than ten individuals at a time, thus limiting their ability to interrogate recent history. Relate infers population-size trajectories that will look familiar to PSMC users but appear smoother and better behaved as a result of more data and more accurate estimates of the ages of common ancestors. Speidel et al. also leverage Relate to refine state-of-the-art tests for natural selection, including the singleton density score[8]. They find novel, complex signals of polygenic selection on traits such as body mass index and blood pressure.

Relate and tsinfer also offer new insights into population structure, both modern and prehistoric. Using genealogical structure inferred from tens of thousands of UK Biobank samples, Kelleher et al. have revealed subtle genetic distinctions among the populations of London, Edinburgh, and their rural outskirts. At the opposite time extreme, Speidel et al. identify variation that probably originated in other species and ended up in the human gene pool, owing to secondary contact with Neanderthals, Denisovans and unknown ancient hominins. They estimate that Neanderthals and Denisovans contributed more than 90% of the archaic introgression found outside Africa, but they find support for a growing consensus in the human genetics community that early humans living in Africa interbred with some other unidentified hominins[9,10].

Although these papers contain too many tantalizing results to summarize here, their farthest-reaching contributions are probably the files of inferred trees that everyone working in population genetics can download and use to improve their own inference methods. The already fast-paced field of population genomics may soon move even faster if method developers can avoid reinventing the wheel that is parsing genes into genealogies. ❐

Kelley Harris
Department of Genome Sciences, University of Washington, Seattle, WA, USA.
e-mail: harriske@uw.edu

Published online: 2 September 2019
https://doi.org/10.1038/s41588-019-0492-x

References

1. Cann, R. L., Stoneking, M. & Wilson, A. C. *Nature* **325**, 31–36 (1987).
2. Kelleher, J. et al. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0483-y (2019).
3. Speidel, L., Forest, M., Shi, S. & Myers, S. R. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0484-x (2019).
4. Griffiths, R. C. & Marjoram, P. *J. Comput. Biol.* **3**, 479–502 (1996).
5. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. *PLoS Genet.* **10**, e1004342 (2014).
6. Li, H. & Durbin, R. *Nature* **475**, 493–496 (2011).
7. Schiffels, S. & Durbin, R. *Nat. Genet.* **46**, 919–925 (2014).
8. Field, Y. et al. *Science* **354**, 760–764 (2016).
9. Hsieh, P. H. et al. *Genome Res.* **26**, 291–300 (2016).
10. Ragsdale, A. P. & Gravel, S. *PLoS Genet.* **15**, e1008204 (2019).

Competing interests

The authors declare no competing interests.