GENETICS

Genealogy at the genome scale

Two teams develop methods for genealogy inference using large genomic datasets.

Genealogies describing ancestral relationships among individuals harbor key information about demographic history and evolution. Although trees are widely used for summarizing genealogies, recombination can shuffle genetic materials between chromosomes and generate discordant trees among loci. In order to encode the genealogies of the whole genome, networks of huge size and complexity are needed, causing a major hurdle for statistical inference using large-scale genomic data.

Several computational methods exist to tackle this challenge. However, the daunting computational burden limits the size of the datasets they can handle. Using different computational strategies, two teams from Oxford University independently developed methods to analyze large genomic datasets for genealogy inference.

Simon Myers and colleagues adopted a two-step approach (called 'Relate') that infers different elements of the genealogies: tree topologies and branching lengths, using position-specific distance matrices and a Markov Chain Monte Carlo (MCMC) algorithm, respectively. Based on the output, quantities of population and evolutionary genetic significance can be further estimated, such as effective population sizes, mutation rates and gene-flow through time. They apply Relate to the 1,000 Genomes Project dataset, and also developed a test for detecting positive selection. "By building its statistics from the genealogies themselves, this makes the test more powerful and sensitive," said Leo Speidel, a member of the team, and Myers. "Despite a lack of strong selection at particular loci, we found directional selection acting on complex traits to be extremely common in humans. The complex, differing patterns among groups and therefore through time were a surprise to us, building on existing findings."

Jerome Kelleher and colleagues developed another inference method ('tsinfer') based on an efficient data structure called a tree sequence, which makes use of the correlation structure between adjacent trees. Using a strategy that infers the ancestral haplotypes, as well as the copying paths for them and the sample haplotypes, tsinfer achieves comparable accuracy with

state-of-the-art competitors and with extraordinary efficiency, as shown by tests using the 1,000 Genomes Project, the Simons Genome Diversity Project and the UK Biobank.

In addition to genealogy inference, "tree sequences, which encode biological history in such fine detail, are also extremely computationally efficient and could help address the spiraling costs of storing and processing DNA sequence data," says Kelleher. The key idea is that the genetic variation in a population is the result of ancestral mutational events. Therefore, storing data indicating where mutations occurred in the genealogies would generate a full picture of the genetic variation, while being much more efficient than storing each sample's genotype. Along this line, they estimate that using tree sequences to store genetic variation can be four orders of magnitude more efficient than the current most widely used format. "To me, this synergy between biology and computing is beautiful and compelling," says Kelleher.

Regarding the next step, both teams are working to improve the tools in several ways. "We focused initially on achieving scale, and are now working to relax some assumptions and make tsinfer more general and robust," says Kelleher, furthermore, "we are continuing to develop tskit (the tree sequence toolkit) and hope that over time an ecosystem of inference methods and analysis tools will grow around this shared, open-source infrastructure." Speidel and Myers hope their work might motivate others to think about leveraging tree-based approaches to answer evolutionary questions, noting, "We also see great potential in combining the innovations of Relate and tsinfer".

Lin Tang

Published online: 31 October 2019
<https://doi.org/10.1038/s41592-019-0639-9>

Research papers

Speidel, L. et al. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).

Kelleher, J. et al. Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).

Guiding New Generations for Decades

Worthington is supporting life science research and discovery, and we have done so for over 70 years.

Authoring technical manuals and guides both in print and digital formats. And, we are cited in thousands of respected scientific journals across the globe.

Sharing Our Enzyme Expertise

Order your free copy of the **NEW Tissue Dissociation Guide**, 18th Edition. Simply go to:

Worthington-Biochem.com/TissueDissociation/default.html

**Come see us at ASCB '19
Booth 526**

Worthington
Biochemical Corporation

800.445.9603 • 732.942.1660