JASSs Journal of Anthropological Sciences Open Access
forum Vol. 99 (2021), pp. 1 - 4 e-pub ahead of print doi. 10.4436/jass99010

Methodological challenges and opportunities for inferring human demography

Garrett Hellenthal

UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, UK
e-mail:ghellenthal@gmail.com

Advances in statistical techniques to analyse rapidly emerging large-scale genome-wide genetic variation data, in particular Single-Nucleotide-Polymorphism (SNP) data, have enabled a correspondingly rapid increase in our understanding of the demography of human populations worldwide. These include new insights into the genetic relatedness of different populations, intermixing (i.e. admixture) among them, and population size changes over time. Initial statistics measured genetic differences among populations or genetic diversity within populations (e.g. Jakobsson et al. 2013). Subsequently, techniques emerged to visualise patterns of genetic distance and sub-structuring among populations, such as principal-components-analysis (Patterson et al. 2006) and clustering algorithms (Alexander et al. 2009). Statistical models were then built to infer the processes, such as admixture and population size changes, leading to these observations under simplified scenarios. As examples, prominently used techniques to infer trees relating populations, including relative branch lengths, admixture events and/or population size changes include those that use correlations in populations' allele frequencies (e.g. Patterson et al. 2012) and those that use simulations to match aspects of observed data, such as the distribution of allele frequencies (e.g. Excoffier et al. 2013).

Many of these techniques analyse SNPs independently. Alternative approaches (e.g. Lawson et al. 2012) instead model the correlations among nearby SNPs, so-called "haplotype" information, that result from the process by which DNA is passed down from one generation to another. Several studies have illustrated how these techniques can have more power to infer population sub-structure and admixture relative to methods that ignore these correlations (e.g. Leslie et al. 2015). However, one challenge is extracting reliable haplotype information from genetic variation data collected in ancient human remains (aDNA), a potentially rich source of ancestral information. This is because aDNA is typically of much lower quality compared to DNA extracted from present-day people, making it challenging to make the reliable diploid calls necessary for modelling haplotype information.

A recent breakthrough using haplotype information involves efficient inference of the ancestral recombination graph (ARG) that relates the genetic material of thousands of sampled individuals back in time to a common ancestor (Speidel et al. 2019; Kelleher et al. 2019). Accurate inference of the genealogies comprising the ARG has been a long-standing goal in population genetics, as in principle this captures the entire acheivable information about individuals' genetic history. In practice, these methods have provided opportunities for new insights into a wide range of processes. Current examples include inferring population size changes over time with greater precision, especially in recent time periods within the last 10,000 years, inferring how mutation rates in different types of sites have evolved over time, and unearthing archaic introgression. Importantly, these techniques have been updated recently to incorporate aDNA, enabling inference on how ancient populations fit onto the tree topologies of modern humans (Speidel et al. 2021). Especially promising is that inference on genetic relationships among ancients and moderns was generally robust to analysing aDNA samples with average coverage as low as 0.01-0.1x.

These exciting new techniques set up a powerful new platform for addressing important details of human history for which we currently have limited understanding. For example, we know relatively little about the number and timings of different episodes of introgression of hominin groups carried by various present-day human populations, and the levels of genetic isolation among archaic human groups. In particular while signals related to Neanderthals and Denisovans have been detected in subsets of present-day populations world-wide, the extent of introgression into e.g. present-day African populations is a major focus of ongoing work (Speidel et al. 2019). In addition, we have a limited understanding of older admixture events among anatomically modern human populations, e.g. >3,000 years ago. In particular the most precise current techniques to infer and date admixture rely on studying the sizes of contiguous segments introduced by an admixing source, which present two major challenges that hinder the study of older events. One is that recent admixture occurring in the last ~3000 years appears to be ubiquitous among human groups, and this can act to mask older events when using these approaches. The other is that the lengths of contributed contiguous segments from an admixture event become smaller over time, making segments from older events more challenging to detect. Consequentially, such approaches can only reliably detect admixture events that have occurred less than ~5000 years ago. In theory, aDNA can be used to detect these older events, since they are unaffected by recent admixture, and the signatures of these older events will be more recent in aDNA relative to modern genomes. However, it is unclear whether enough quality aDNA will be recovered for homogeneous populations with the same admixture history to detect such events reliably. In principle, geneaology-based approaches can avoid these issues, by ignoring ancestral relationships in recent time periods and instead focussing on studying the sharing of ancestors in older time epochs.

For such reasons, future methodological developments in this area will likely make increasing use of inferred genealogies. As examples, this may include leveraging genealogies to infer and date admixture, identify segments introgressed from archaic humans, elucidate genetic sub-structure at different time scales, impute missing data, understand how recombination and mutation rates vary over time, and infer selection, with some such advancements already being implemented. As these are emerging, researchers across disciplines will become increasingly familiar with genealogically-based techniques and their output, enabling more widespread use. However, their acceptance into standard practice will depend on the ease-of-use of released software and the magnitude of power increases relative to available programs.

Despite these exciting opportunities, several challenges remain. While current approaches can infer details of genealogies, including time depths, for thousands of samples, more computational resourcefulness will be necessary to scale to current biobank-level data consisting of hundreds of thousands of people. These approaches also currently rely on whole-genome-sequencing (WGS) data, which is scarce relative to array data ascertained at specific SNPs that are available for many present-day and ancient individuals. One potential solution is to first use imputation to infer sequencing information in such array data. However, it is unclear the extent to which this will bias signals towards those in the reference panels used to assist imputation, and the extent to which reference panels will capture haplotype patterns in populations of interest. In addition, distinguishing ancient population sub-structure from archaic introgression will be difficult, especially in African populations. When considering older demographic history, such as ancient admixture events, it is also unclear how much information will be lost by focussing on less recent genealogical events, given the number of branches on the geneaology (and hence information) become more sparse as you go back in time. Finally, genealogical trees are not inferred perfectly, with current approaches requiring model simplifications to scale to large numbers of individuals. For example, the extent to which mutation and/or recombination rates vary over time, and how this affects inference of some parameters, is not entirely understood. Even with perfectly inferred genealogical trees, the processes leading to these trees are not necessarily easy to estimate. In particular disparate demographic processes can in principle lead to similar tree topologies. Therefore, leveraging information from other sources, e.g. anthropological, archaeological and linguistic research, will continue to prove invaluable when interpreting genetic relatedness patterns.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655-1664. https://doi.org/10.1101/gr.094052.109

Excoffier L, Dupanloup I, Huerta-Sanchez E, et al (2013) Robust Demographic Inference from Genomic and SNP Data. PLoS Genet 9:e1003905. https://doi.org/10.1371/journal.pgen.1003905

Jakobsson M, Edge MD, Rosenberg NA (2013) The Relationship Between FST and the Frequency of the Most Frequent Allele. Genetics 193:515-528. https://doi.org/10.1534/genetics.112.144758

Kelleher J, Wong Y, Wohns AW, et al (2019) Inferring whole-genome histories in large population datasets. Nat Genet 51:1330-1338. https://doi.org/10.1038/s41588-019-0483-y

Lawson D, Hellenthal G, Myers S, et al (2012) Inference of population structure using dense haplotype data. PLoS Genet 8:e1002453. https://doi.org/10.1371/journal.pgen.1002453

Leslie S, Winney B, Hellenthal G, et al (2015) The fine-scale genetic structure of the British population. Nature 519:309-314. https://doi.org/10.1038/nature14230

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2:e190. https://doi.org/10.1371/journal.pgen.0020190

Patterson N, Moorjani P, Luo Y, et al (2012) Ancient Admixture in Human History. Genetics 192:1065-1093. https://doi.org/10.1534/genetics.112.145037

Speidel L, Forest M, Shi S, et al (2019) A method for genome-wide genealogy estimation for thousands of samples. Nat Genet 51:1321-1329. https://doi.org/10.1038/s41588-019-0484-x

Speidel L, Cassidy L, Davies RW, et al (2021) Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. Mol Biol Evol, msab174. https://doi.org/10.1093/molbev/msab174