JASs Journal of Anthropological Sciences
Open Access
*forum* Vol. 99 (2021), pp. 179 - 182
doi. 10.4436/jass.99009

# Complexity in human ancestral demography

**Aylwyn Scally**

*Department of Genetics, University of Cambridge, Downing St, Cambridge, CB2 3EH, UK*
e-mail:a.scally@gen.cam.ac.uk

Demography – the size and structure of populations, and the movement of individuals within and between them – is central to how we describe and understand the events of human evolution. But while archaeology can reveal the presence of people at a certain time and place, and genetic data may indicate their shared ancestry with others, we rarely have direct evidence for how many people there were or in what groups they lived. Therefore the inference of these and other aspects of the past using demographic models is a focus for many genetic and archaeological studies.

Traditionally, models of ancestral demography have emphasised the branching and divergence of populations around the world, representing human demographic history as a tree. Thus, the default picture has been one in which the relationships between human populations mirror those between species in the tree of life. This way of thinking emerged to some extent from older racially-motivated ideas about anthropological diversity, and was partly reinforced (particularly outside genetics) by the fact the first sources of genetic evidence for human population history were mitochondrial and Y-chromosomal DNA, which as single genetic loci were fully described by genealogical trees. However there are often good reasons to use simple models even when exploring complex phenomena, and tree models have the advantage that they are relatively easy to implement and to interpret in terms of many questions of interest. In this context, putative episodes of gene flow between populations (which are incompatible with a tree-like demography) have generally been examined in terms of model fit and thus treated implicitly as exceptional events. Sometimes this is made explicit by casting admixture as an alternative hypothesis in contrast to the 'null' model of tree-like population divergence.

However it is now clear that admixture, gene flow and migration have been ubiquitous in human demographic history, and not infrequent events (Korunes and Goldberg 2021). In light of this, trees become more difficult to interpret and their parameters less meaningful. What, for example, does the inferred split time between two populations mean in the context of gene flow after divergence, or when there have also been interactions with multiple other populations? To deal with these and similar questions, more complex models and approaches have been developed. In particular, two kinds of model are predominantly used in paleogenomic studies to investigate demography and genetic ancestry, namely 'isolation-migration' (IM) models and admixture graphs.

An IM model comprises one or more discrete populations, each with an explicit size, and a branching topology describing the history of relationships between them, including divergence events (leading to isolation) and gene flow (due e.g. to migration). Gene flow may be represented either as pulses of admixture with no extension in time or as intervals of continuous flow at a specified rate, constant or otherwise (Fig. 1A). IM models are graph-like, in that when restricted to pulsed gene-flow events and piecewise-constant population sizes the model has a straightforward graph representation. However with continuous gene flow and continuously varying population sizes things become more complex, and the parameter space considerably larger. IM models can be inferred from data quantifying the genetic differences within and between sampled populations, and in general the inference (particularly of population sizes) is improved by including more genomes, although the genetic coalescent process limits this as one looks further back in time. Different implementations (e.g. Gronau et al. 2011; Hey et al. 2018;

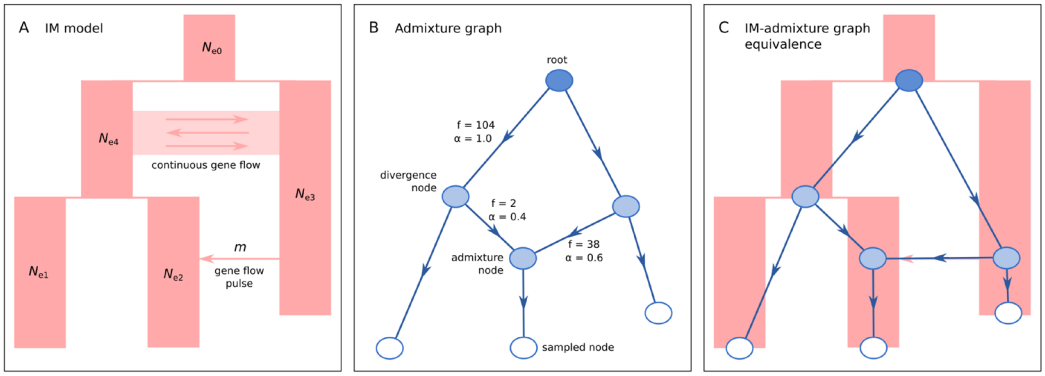


***Fig. 1 - A -** An IM model, comprising finite populations (solid blocks, labelled by size) connected by divergence and gene flow events (horizontal lines). The vertical dimension represents time, with earlier events at the top. In the model shown there are two gene flow events: an earlier continuous episode and a later pulse of admixture; to fully specify the model, migration parameters (rates or fractions) need to be given for each.*
***B -** An admixture graph is a directed acyclic graph in which each edge points forwards in time and leaf nodes represent sequences (or collections of sequences) for which we have data. Edges are labelled by the strength f of genetic drift along them (typically in units of FST / 1000) and by their proportional contribution α to the downstream node (with contributions from input edges at any node summing to one). Implementations of this structure may vary; for example in one common implementation, edges with α < 1 'admixture edges' are restricted to have f = 0, necessitating the inclusion of additional nodes and edges.*
***C -** An IM population of effective size Ne, extending for a time T, corresponds to an admixture graph edge with drift parameter f = 1 − exp(−T / 2Ne), and the proportion of migrant genomes in an IM population following a gene flow pulse corresponds to the input edge flow contribution in the admixture graph. Under this equivalence, leaf populations are related by the same genetic correlation structure. Note however that due to its explicit representation of population sizes and times, an IM demography also models aspects of genetic data not represented by an admixture graph. Similarly, continuous gene flow in an IM model has no straightforward equivalence in an admixture graph.*

Kamm et al. 2020) have taken a wide variety of approaches to the problems of model specification and parameter inference, and IM models are also the demographic framework used by coalescent simulators such as msprime (Kelleher et al. 2016).

An admixture graph describes the genetic ancestry of a set of samples in terms of correlations between their sequences (Patterson et al. 2012). One can think of it as representing the flow of shared ancestry information (in the form of genotype correlation) from ancestors to descendants, information which flows and dissipates along the graph edges, diverges or converges at internal nodes, and terminates at the leaf nodes, where we sample it (Fig. 1B). Unlike IM models, admixture graphs are not strictly demographic models, in the sense that they do not explicitly represent population sizes and timescales. However for any admixture graph one can construct a family of equivalent IM models (Fig. 1C) whose populations correspond to admixture graph edges, and where divergence and admixture events correspond to admixture graph nodes. Thus the structure of an admixture graph tells us about demographic history, albeit from a perspective determined by the samples chosen as leaf nodes.

Both IM models and admixture graphs are valuable tools for investigating complex demography, and are widely used with a variety of algorithmic approaches. However, except where there are very few sampled sequences or taxa, inference is typically carried out only on the non-topological parameters of the model (such as effective population sizes, times, drift and admixture parameters).

Indeed branch length parameters can often be inferred efficiently, for example by taking advantage of the conditional independence of subgraphs given their ancestral nodes. But as yet no equivalent approaches exist to optimise over the vast space of possible topologies, so in general the topology itself is arrived at by a combination of practitioner expertise and heuristic methods. Ultimately this is problematic, because it means our ability to reconstruct ancient demography may be strongly influenced by prevailing ideas and prior expectations.

It should be noted that topology inference was already a problem even with strict tree models. The number of rooted tree topologies for n taxa is given by (2n – 3)!!, so for example with 10 taxa there are already over 34 million topologies to consider, many of which will be indistinguishable unless the data are sensitive to differences in branching deep within the tree. This problem is even more acute in an admixture graph or IM model, where the space of topologies is much larger.

This is a key challenge for ancestral demographic inference, and thus for the field in general. Demographic models often form the basis for conjecture about the relationships between ancient peoples, sometimes including hitherto unobserved 'ghost' populations. For most time periods and in most regions it is no longer reasonable to make the default assumption of no gene flow between human populations, and while ancient DNA can be very informative for these questions, its availability is not something we can rely on as a matter of course. Depending on what samples we use and which topologies we consider, our best-fitting models may exclude important components of the true demographic history. It is difficult to know whether, for example, inferred common ancestry is due to direct interaction between certain ancient populations or an indirect relationship via other unobserved groups. Similarly, a putative ghost population may indeed have existed as a prehistoric society, or may be an abstraction of something far less coherent. Such distinctions can be important when relating genetic inferences to archaeological, historical or other evidence.

Methodologically we would like to be able to fully sample the space of graphs for different sample choices. Among other things, this would allow us to use statistical and machine learning approaches which have been applied successfully to many other complex inference problems, including in evolutionary genetics (Schrider and Kern 2018). There are also relevant approaches in other fields, particularly machine learning, where directed acyclic graphs (DAGs – a category which includes all the graphs described here) are very widely used, and the question of learning graph structure is a long-standing one (Kuipers and Moffa 2018). Possibilities may also be found elsewhere within population genetics, where much work has been devoted to inference of the ancestral recombination graph (ARG), representing the common genetic inheritance of one or more sampled chromosomal sequences. Importantly, an ARG is a connected DAG with essentially the same topology as an admixture graph, and hence the same algorithms used for generating, sampling and inferring ARGs can in principle be applied.

We have focused on graph-like models, which represent demography in terms of populations and gene flows, as they are widely used and map naturally onto theoretical concepts in population genetics. But populations are themselves a much-simplified abstraction. If complexity was no object, a more realistic depiction might represent individuals living, moving and interacting within a spatial landscape, with the process of genetic inheritance governing the transmission of haplotypes to their offspring. Due to the complexity involved, such models have yet to be developed or implemented on a scale sufficient to investigate human ancestry and demography. However, some theoretical and practical progress has been made, for example in exploring models with explicit spatial-genetic representations, which have provided valuable insights into the nature of genetic evolution in a spatial domain (Barton et al. 2010; Bradburd et al. 2018; Al-Asadi et al. 2019; Battey et al. 2020). These and similar approaches have the potential to open up new possibilities for evolutionary inference.

For now, even while topology inference remains prohibitive for large datasets, users of demographic models should still consider how they might better explore and communicate the robustness of their

inferences. This might include, for example, more systematic ways of presenting the range of models considered and greater emphasis on the motivation and reasoning behind the expert choices involved. Some recent studies have indeed begun to do this, particularly where key findings depend on the structure of the graph (e.g. Ning et al. 2020), although the discussions remain abstruse for non-expert readers. Elsewhere some datasets (or key subsets of data) are small enough to allow a more intensive computational investigation of the model space, and useful progress has recently been made on ways to facilitate this (Leppälä et al. 2017; Hey et al. 2018; Yan et al. 2020; Molloy et al. 2021). Such approaches should perhaps be standard where feasible. We can hope that increases in computing power will gradually enlarge this category, and with the approaches mentioned above, or others unforeseen, will enable us to better investigate the complexity of human ancestral demography.

## References


Al-Asadi H, Petkova D, Stephens M, et al (2019) Estimating recent migration and population-size surfaces. PLoS Genet 15:e1007908. https://doi.org/10.1371/journal.pgen.1007908

Battey CJ, Ralph PL, Kern AD (2020) Space is the place: Effects of continuous spatial structure on analysis of population genetic data. Genetics 215: 193-214. https://doi.org/10.1534/genetics.120.303143

Barton NH, Kelleher J, Etheridge AM (2010) A new model for extinction and recolonization in two dimensions: quantifying phylogeography. Evolution 64:2701-2715. https://doi.org/10.1111/j.1558-5646.2010.01019.x

Bradburd GS, Coop GM, Ralph PL (2018) Inferring continuous and discrete population genetic structure across space. Genetics 210:33-52. https://doi.org/10.1534/genetics.118.301333

Gronau I, Hubisz MJ, Gulko B, et al (2011). Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43:1031. https://doi.org/10.1038/ng.937

Hey J, Chung Y, Sethuraman A, et al (2018) Phylogeny estimation by integration over isolation with migration models. Mol Biol Evol 35:2805-2818. https://doi.org/10.1093/molbev/msy162

Kamm J, Terhorst J, Durbin R, et al (2020) Efficiently inferring the demographic history of many populations with allele count data. J Am Stat Assoc 115:1472-1487. https://doi.org/10.1080/01621459.2019.1635482

Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol 12:e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Korunes KL, Goldberg A (2021) Human genetic admixture. PLoS Genet 17:e1009374. https://doi.org/10.1371/journal.pgen.1009374

Kuipers J, Suter P, Moffa G (2018) Efficient Structure Learning and Sampling of Bayesian Networks. arXiv preprint arXiv:1803.07859

Leppälä L, Nielsen SV, Mailund T (2017) Admixturegraph: an R package for admixture graph manipulation and fitting. Bioinformatics 33:1738-1740. https://doi.org/10.1093/bioinformatics/btx048

Molloy EK, Durvasula A, Sankararaman S (2021) Advancing admixture graph estimation via maximum likelihood network orientation. bioRxiv 2021.02.02.429467. https://doi.org/10.1101/2021.02.02.429467

Ning C, Fernandes D, Changmai P, et al. (2020) The genomic formation of First American ancestors in East and Northeast Asia. bioRxiv 2020.10.12.336628. https://doi.org/10.1101/2020.10.12.336628

Schrider DR, Kern AD (2018) Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet 34:301-312. doi: 10.1016/j.tig.2017.12.005.

Yan J, Patterson N, Narasimhan VM (2020) miqoGraph: fitting admixture graphs using mixed-integer quadratic optimization. Bioinformatics 37:2488–2490. https://doi.org/10.1093/bioinformatics/btaa988