

Sometimes hidden but always there: Assumptions behind demographic inference from ancient DNA data.

Liisa Loog*

Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH,
United Kingdom

*Corresponding author: L.L. (liisaloog@gmail.com)

Abstract

Demographic processes directly affect patterns of genetic variation within contemporary populations as well as future generations, allowing for demographic inference from patterns of both present day and past genetic variation. Advances in laboratory procedures and sequencing and genotyping technologies in the last decades have resulted in massive increases in high quality genome-wide genetic data from present day populations and allowed retrieving genetic data from archaeological material, also known as ancient DNA. This has resulted in an explosion of work exploring past changes in population size, structure, continuity and movement. However, as genetic processes are highly stochastic, patterns of genetic variation only indirectly reflect demographic histories. As a result, past demographic processes need to be reconstructed using an inferential approach. This usually involves comparing observed patterns of variation with model expectations from theoretical population genetics. A large number of approaches have been developed based on different population genetic models that each come with assumptions about the data and underlying demography. In this article I review some of the key models and assumptions underlying the most commonly used approaches for past demographic inference and their consequences for our ability to link the inferred demographic processes to the archaeological and climate records.

Key words: Population genetics; Statistical modelling; Demographic modelling; Ancient DNA; Population history; Archaeology

Introduction

Genetic information from present day individuals has been used for past demographic and evolutionary inference for decades. However, recent advances in sequencing and genotyping technologies have reduced the cost of generating genetic data substantially, allowing for large high-quality population-wide datasets to be produced and used for addressing questions about demography ranging from the timing of recent human “Out-of-Africa” event [1] to mate choice patterns and structure of genetic variation within present day populations [2].

This reduction of cost per base pair sequenced combined with advances in specialised laboratory protocols for degraded genetic material [3–5] has also allowed large scale sequencing of genetic data from archaeological material, also known as ancient DNA, and resulted in an explosion of work exploring changes in population size, structure, continuity and movement in last few thousand years.

While some aspects of past demography have simple predictions for genetic patterns of variation, such as the biological sex [6] of a past individual or familiar genetic relationships between individuals [7], most demographic processes leave behind a more convoluted signature. In fact, the patterns of genetic variation across individual genomes are a result of accumulated effects of different past demographic processes combined with the stochasticity of inheritance.

As a result, demographic processes need to be inferred by comparing patterns of observed genetic variation to theoretical or (in case of more complicated demographic scenarios) simulated model predictions, both relying on the population genetics theory with its roots in the work by R. Fisher, S. Wright, J. B. S. Haldane and others in the 1920s and 1930s.

Different inferential approaches require assumptions about the data and the underlying demographic processes, that in some methods need to be made explicitly whereas in others they are implicit. These assumptions, and the level to which they are met, can heavily affect the suitability different frameworks for testing competing demographic hypotheses as well as the interpretation of such demographic modelling. In this article I will review the assumptions underlying the most common approaches for past demographic inference as well as their consequences, especially in relation to our ability to link the inferred demographic processes to the archaeological and climate records.

Pattern-based approaches to demographic inference

Phylogeographic inference is an approach to reconstruct population histories that was especially popular in the early days of past demographic inference using genetic data (e.g. [8,9] but also [10]02/03/2020 15:16:00). Here, a phylogenetic tree is constructed based on the substitutions (mutations) in a single locus (usually a non-recombining part of the genome like the mitochondrial DNA or the Y chromosome). All sampled individuals are assigned to haplogroups and haplotypes, which are the branches and sub-branches, or lineages, on the inferred tree. The root of the tree corresponds to the most recent common ancestor of all the samples. When the mutation rate is known (or can be estimated), it is possible to estimate the dates of each branching point of the tree, using the accumulation of mutations along branches in the tree as a "molecular clock". Inferences about the past are based on the phylogenetic relationship between different haplotypes or haplogroups, their estimated splitting times, and their distribution in space and time.

The key challenge with inferring the past from these trees lies in that events in the single locus phylogenetic trees do not generally correspond directly to population-level events as they are stochastic outcomes of given population histories. As a result, very different demographic scenarios can give rise to qualitatively similar gene trees and distribution of haplotypes [11]. Although studies relying on data from single loci often lack the statistical power to test complex population histories, especially when exclusively from modern populations [12–14], robust inference can in principle be achieved when using sufficient data (e.g. samples from before, during and after the demographic event of interest) and explicit statistical modelling that accounts for the randomness of individual loci and allows to formally consider the likelihood of different alternative demographic scenarios (e.g. [15,16]). Although ancient DNA alone already provides additional resolution that often enable researchers to exclude less likely scenarios (e.g.[17–19]), without explicit demographic modelling the conclusions can be easily steered by the subjective biases of a particular researcher [14].

The latter also applies to neighbour joining trees based on many more loci or even whole genomes: Such algorithms are designed to join samples based on genetic similarity without explicit consideration to demographic processes such as gene flow, genetic drift, isolation and identity by descent, let alone more complex scenarios combining these processes, that may have caused the observed similarities and differences between populations. Therefore, the resulting tree does not in itself provide sufficient evidence to conclude that the studied populations have split like the tree suggests. Nevertheless, these approaches are extremely

useful for identifying issues with data quality, such as sequencing batch effects and other artefacts resulting from data generation, as well as for generating alternative hypotheses or demographic scenarios, to be formally tested within a hypothesis testing framework.

Descriptive methods for inferring population structure

Principal Component Analysis (PCA) is a commonly used technique for assessing the genetic similarity between individuals and the extent to which populations form distinct clusters. In a PCA of genetic data, each locus is treated as an independent variable (dimension) and genetic variation among samples is reduced into main axis of variation - principal components (PCs), using the covariance matrix. PCA analysis of modern human populations has been shown to reflect geographic relationships between populations on different scales: on a global scale, populations from the same sub-continent tend to group together [20], and across Europe the genetic locations of samples in the PCA show striking similarity to their geographic locations [21].

In ancient DNA literature, modern populations are typically used to define the PCs and ancient individuals are projected onto them. As a general rule, only the first few axes of variation (PCs) are presented, representing a small proportion (typically only few percentages) of total variation present in modern populations. However, the older the sample, or otherwise more distantly related to the modern populations used in the analysis, the more likely it is that more of its variation ends up in an orthogonal axis to the ones presented/used for inference. This problem can be exacerbated by the projection method as individuals with more missing data are more likely to end up closer to the average of the modern data points. In principle, it would be possible to overcome this problem by turning the analysis around: Using ancient samples to define PCs and projecting modern samples onto this variation, in order to investigate the genetic similarity between modern and ancient populations. However, this would require a representative high-quality ancient DNA sample from each time slice and geographic region of interest.

The main problem with inferring population history from PCA is that it lacks an underlying population genetic model, and several different scenarios can result in similar distribution of samples on PCs [22]. For example, a population can appear different in a PCA because it has been separate from other populations for a long time, or because a recent population bottleneck has caused extensive drift in allele frequencies within that population [23]. Therefore, it is not

possible to directly relate the inferred distances between samples on PCs to demographic processes behind the observed variation.

Similar criticism applies to tools commonly used in demographic inference using modern and ancient DNA such as STRUCTURE [24], ADMIXTURE [25], Finestructure, and Chromopainter [26]. These tools vary in their technical implementations but are all designed to identify major genetic clusters and express each sample as a mixture of these clusters. For example, people living in central America today can be modelled as admixture between present-day Native American, European and African populations [27]. The underlying model of such approaches assumes that (present day) individuals are a product of admixture of distinct “source” groups that existed in the past. Problems with such approaches may arise when this cannot be assumed a priori and the inferred statistical clusters are misleadingly taken as evidence of existence of “ancestral” or “source” populations, when in reality, the existence of such clusters could be explained by multiple distinct demographic histories [28]. Unless more complex demographic scenarios are explicitly considered, it is inherently not possible to identify which demographic scenarios have resulted in such clusters using clustering approaches alone. However, sometimes it is possible to use such tools to test competing demographic scenarios, provided that it is clear that the tested scenarios are expected to produce distinct clustering patterns (e.g. [29]).

In summary, statistical approaches such as the Principal Component Analyses and neighbour joining trees (regardless whether based on a single locus or whole genomes) as well as clustering tools such as tools such as ADMIXTURE can be very useful for summarising and visualizing complex population genetic data (e.g. for generating alternative hypotheses or identifying problems with data quality). However, the translation of such patterns into explicit demographic scenarios is less than straightforward as such tools lack a formal demographic model as well as a hypothesis testing component. This results in inference that can be easily steered by subjective interpretation of individual researchers, unless the demographic scenarios of interest can be expected to produce vastly different patterns of variation.

Methods for inferring spatial barriers to past mobility

From theoretical population genetics and ecology it is known that reduced mobility, for example caused by geographic or cultural barriers, can increase differences in allele frequencies between populations. For pairs of populations, this has traditionally been captured by F_{ST} , the ratio of the variance of allele frequencies between sub-populations to the variance

in the whole population [30]. In an ideal setting, such as Wright's infinite island model [31] where populations are located on an infinite square lattice, F_{ST} is directly related to the number M of migrants per generation between neighbouring populations in the lattice [$F_{ST} = 1/(1+2M)$], such that no migrants yield $F_{ST} = 1$ and "infinite" migration rate (i.e. a panmictic population) yields $F_{ST} = 0$.

Several tools have been developed to extend this principle. CircuitScape [32,33] is commonly used in landscape population genetics to test the effect of geographic features on population connectivity. It approximates the effect of gene flow as current through an electric circuit, where F_{ST} values between pairs of populations correspond to pairwise voltage measurements in the circuit. By solving the corresponding electrostatic equations, the program estimates gene flow along all paths. Recently, this tool was used to estimate different routes during the initial colonisation of Australia [34].

Pagani et al. [1] developed an analytical technique, based on Gaussian kernel interpolation, to study past barriers to gene flow by quantifying spatial gradients in allele frequencies. They applied it to human whole genome data across Eurasia and found evidence of major mountain ranges and deserts acting as barriers to gene flow. Petkova et al. [35] introduced EEMS (Estimated Effective Migration Surfaces), a tool based on computational geometry designed to deal with spatially irregular patterns of data. In this framework space is represented as a polygon subdivided into triangles. Each of the triangles is associated with a local movement rate that is constant within the triangle. Finally, the method uses classical population genetics theory to find movement rates that corresponds to observed pairwise differences in allele frequencies among samples.

Mathieson et al. [36] applied the EEMS approach to 116 samples dated to be older than 7000 BP in order to investigate population structure in European hunter-gatherers. These methods rely on the assumption that observed genetic similarities and differences between samples are a function of physical isolation between them, so that genetic differences between samples of different age (resulting from genetic drift and mutations) can cause spurious geographic barriers to be inferred in temporally heterogenous datasets. However, this issue can be mitigated by making sure that the temporal genetic differences are small enough compared to the geographic differences considered.

Sampling bias

Bias in sampling is an issue that can potentially affect all population genetic inference from pattern-based qualitative descriptions of data to explicit demographic modelling. As a general rule, approaches for past demographic inference from genetic data rest on the assumption that individuals are sampled randomly from a population so that the observed differences between samples is representative of the whole population. This can pose a significant problem for demographic analyses based on individuals from museum collections or archaeological specimens where sampling is largely influenced by a number of non-random factors such as preservation, as well as excavation and sampling locations and periods that are not always spatially or temporally randomly distributed.

The issue of non-uniform sampling can be dealt with using approaches that explicitly consider spatial and temporal location of samples. For example, Loog et al. [37] developed an extension to the F_{ST} measure that unifies spatial and temporal distances into a single metric. They showed that this metric directly informs on past levels of mobility and applied it to genome-wide data from a set of spatially and temporally sparsely distributed ancient Europeans.

However, even such methods can give misleading results if the sampled individuals are closely genetically related. For example, consider a situation where related individuals are more likely to be subjected to similar funerary practices or be buried in a similar location. If these practices or locations were more likely to be excavated or result in better preservation, sampled individuals could end up being more related than would be expected by chance, leading also to biased downstream demographic analyses. Although familial relatives can be identified and removed from subsequent demographic inference, more subtle sampling biases can be very difficult to detect and account for in the analyses.

Methods for inferring demographic histories from explicit models

As discussed above, approaches such as PCA and Admixture are descriptive, and narratives based purely on such results do not constitute a formal analysis of population history. In order to achieve robust demographic inference, alternative hypotheses need to be formulated as different demographic scenarios. The likelihood of different demographic scenarios can then be formally compared by calculating the probability of observing the data given each scenario. Here the level of detail described in each demographic scenarios is heavily constrained by the available data so that more information rich data can be used to test increasingly more nuanced demographic scenarios.

Tree-like population history models & the history

In order to understand the properties of modern methods for inference of past population dynamics, and some of their weaknesses, it is useful to review the concept of population in population genetics: Conceptually, either directly or indirectly, most population genetic models treat populations as lineages (or taxa) within a phylogenetic framework. That is populations are distinct, homogenous entities; with a history that can be represented as a tree, where branch points represent splits between ancient populations, and leaves of the tree are extant or archaeological populations for which samples are available.

Several methods of varying degree of complexity exist for constructing the phylogenetic tree representing the joint history of populations. Here, I will review the approaches most commonly used in the inferences about past demography, as an exhaustive enumeration of all population genetic modelling methods would be beyond the scope of this paper.

Several computationally intensive methods have been developed to estimate past population sizes, gene flow between different branches in the population tree and, if the mutation rate is known, date of divergence between populations from sequence data (e.g. IMA [38], fastSIMCOAL2 [39] and G-PhoCS [40]) or allele frequency data (e.g. $\partial a \partial i$ [41]). For example, Freedman et al. [42] used the G-PhoCS tool and seven high quality genomes from present day wolves and dogs to formally compare demographic models involving different divergence times, ancestral population sizes, and rates of post-divergence gene flow between the different branches of the Canid phylogenetic tree (each sample representing a different branch).

The approaches based on the phylogenetic framework rely on the assumption that the relationship between populations can be represented as, essentially, a phylogenetic tree, i.e. as abrupt splits between different branches of the tree, followed by independent evolution with potential for subsequent episodes of gene flow between them. As a result, such approaches are not well suited to testing past demographic scenarios where the relationship between populations has historically been more complex than can be represented in a splitting tree model (i.e. groups in close geographic proximity).

A different but related issue is that such approaches do not scale well to large histories, with complex historical interactions as they require all potential demographic events and interactions to be explicitly defined. As a result, a potentially large number of a priori modelling decisions about a demographic process which is, as a general rule, unknown are required for the inference, making them challenging to apply to many archaeological and anthropological questions.

This has created a need for inference methods that require fewer a priori assumptions, and as a result, are sometimes informally referred to as “data driven” or “model free” approaches. Although these approaches do not require a predefined demographic model, they make a number of assumptions about the possible genetic histories of samples within and between populations. As I will illustrate with examples below, such approaches are far from model free.

Population genetic inference for estimating past population sizes

One such approach has been developed for estimating changes in past population sizes. Coalescent theory predicts that gene lineages coalesce with a rate dependent on the effective population size (the size of the population contributing offspring to the next generation). This provides a unique relationship between the size of a population through time and the distribution of time to the most recent common ancestor for pairs of sequences. Li and Durbin [43] developed PSMC (Pairwise Sequentially Markovian Coalescent) approach to implement this principle and reconstruct a detailed history of population size from a single high quality (i.e. with sufficient sequencing depth to resolve heterozygotes) genome sequence. Extensions of this method have been subsequently developed to additionally handle multiple samples from a single population for an increased resolution [44,45] and to infer the history of gene flow between pairs of populations [45].

Li and Durbin [43] used PSMC to infer population size change curves through time from high-quality genomes from Africa, East Asia, Europe. They inferred a severe reduction in population size for non-African individuals around 10-60 thousand years ago) and attributed this to out of Africa population bottleneck.

In these approaches the inference of effective population size changes through time rest on two main assumptions: a) that mutation and recombination rates are known and b) that the population from which the sample was drawn has been panmictic throughout history, i.e. that all contemporary pairs of individuals within that population are equally likely to mate, regardless of geographic separation or other factors. Uncertainty of both mutation and recombination rate estimates results in large confidence intervals for estimated population sizes and time scales of inferred changes (when this is taken into account, frequently on single mutation and recombination rate values are considered in the analyses).

Deviations from the assumption of panmictic populations can have more obscure consequences for the inferred histories. For example, Mazet et al. [46] showed that when PSMC is applied to samples from spatially structured populations, it tends to infer changes in the effective

population size with time also when the true population size has been stable. This can be understood by considering the case of a local sub-populations, where each sub-population is panmictic and is connected to its closest neighbours by migration (gene flow). For short time scales, the ancestors of an individual sampled from one of the sub-populations will tend to belong to the same sub-population, and the inferred effective population size will reflect the size of this sub-population. On longer time scales migration will cause the ancestors to be spread out over a larger geographic area, with increasingly rare contacts (and, as a consequence, lower coalescent rates). From this reduced coalescence rate the PSMC will infer an increasing effective population size, even when there has been no change in it. In this case, the detailed shape of the inferred population size history will depend on the size and number of sub-populations as well as the migration rate between sub-populations [46]. Although magnitude of this effect will largely depend on the level of historic structure present in populations, Mazet et al. [46] suggested that estimates of population size that ignore population structure should be interpreted as estimates of past population size and population structure as it is not clear which aspect is being captured for any given demographic history.

F-statistics and admixture graphs

Another class of methods that do not require a detailed relationship between populations to be explicitly modelled are built on so-called f -statistics [47]; reviewed in [48] in further detail) and have been used extensively in the ancient DNA literature to test hypotheses of the relationship between ancient and modern populations. These statistics also use classical phylogenetic view of a population (described above) and rely on the assumption that genetic drift occurs independently in each population, i.e. on each branch of the tree. Pairs inferred to share more genetic drift are expected to share more of their demographic history. The f_2 statistic measures the amount of drift along both lineages since the divergence from a shared ancestral population in the tree. The f_3 statistic measures the amount of shared drift between two pairs of samples relative to an out-group population. As a result, the f_3 statistic is often used to quantify the level of shared drift between modern-day populations and archaeological (ancient DNA) samples, and to identify the closest present-day relatives to an archaeological sample.

For example, Rasmussen et al. [49] used the f_3 statistic and a sample from an ancient North American individual directly associated with the Clovis tools to test the two competing hypotheses about the origins of the Clovis culture. They found that present day Native

Americans from Central and South America share the greatest amount of genetic drift (highest f_3) with the ancient Clovis individual, followed by individuals from Central Eurasia and only then individuals from Europe, suggesting that the demographic history of people associated with Clovis culture involved a more recent split from East-Eurasian populations than European populations and subsequently supporting the hypothesis where people practicing Clovis culture arrived to the Americas from Asia (via Beringia) over the European origins (Solutrean) hypothesis.

The f_4 statistic estimates the amount of shared genetic drift between pairs of populations, and can be used to statistically test hypotheses of gene flow between sub-populations [47], or hybridisation between populations of different species [50]: if the shared drift is zero, the two pairs of populations must belong to different parts of the tree (i.e. separated by a more ancestral node in the tree). Conversely, if the f_4 statistic is significantly different from zero, the null hypothesis of no separation between populations or species is formally rejected and it is concluded that gene flow or hybridisation occurred. The related D statistic (the f_4 statistic divided by a positive scale factor that makes the D statistic have a range between -1 and 1) is often used in these tests, often referred to as ABBA-BABA tests, measuring the proportion of two gene-tree configurations across the genome.

The f_4 statistic can also be used to statistically test whether different ancestral tree topologies are compatible with the observed data. Frameworks such as TreeMix [24], AdmixtureGraph [48] and qpGraph [51] build on this principle to construct taxonomic trees, using genome wide data and bootstrap analysis to assess the significance of population splits.

Because of their conceptual simplicity and ease-of-use the f -statistics have been hugely popular. They have been successfully employed in answering many long standing archaeological and anthropological questions and have led to a number of important new insights, (e.g. Neanderthal admixture to the present day human gene pool [50] or demographic origins of modern European populations [52]).

However, care needs to be taken when interpreting such results as the analysis can be confounded by incomplete lineage sorting caused by ancient population structure [53], i.e. the genetic drift in the branches of the phylogenetic tree has not been completely independent, but instead influenced by geography.

Population genetic inference for testing population continuity through time

A common question that can be addressed with ancient DNA is whether a present population is directly ancestral to a past population or if the latter has experienced admixture from external

populations. Testing this is not straight forward because alleles present in any generation represent a randomly drawn sample of alleles in a previous generation and, as a consequence, allele frequencies can change considerably even under a scenario of complete population continuity (no gene flow from an external source), a phenomenon usually referred to as “genetic drift” in the population genetic literature. What is more, the expected level of drift greatly depends on the population size: because of the law of large numbers, genetic drift is smaller in large populations and larger in small ones.

Formal tests of continuity are usually (directly or indirectly) based on comparison of observed allele frequency differences between ancient and modern populations with a distribution of expected differences under a null hypotheses of population continuity at a given size (or sizes) (e.g [54,55,49,56,57] examples). Here the difficulty lies in specifying a meaningful null-model, as very few population histories involve complete isolation from neighbouring groups. As demonstrated by Silva et al. [58], the threshold level above which continuity is rejected largely depends on expected level of gene-flow with neighbouring populations. Here a high level of continuous gene flow from an external source can leave a genetic signal indistinguishable from a more sudden population replacement.

One way to overcome this problem is to use inference methods that allow explicit modelling of gene flow (e.g. in [49,15,58,16]). External lines of evidence, such as archaeological, linguistic and geographic information could be used to inform on the expected levels of gene flow to build a more realistic null model for testing continuity. Additionally, a dense serial sampling of a geographic location can help distinguishing between models with more continuous gene flow over longer time periods from models that involve instantaneous replacement.

Linking demographic processes to the archaeological and climate records

Placing demographic events inferred from the genetic data to well-defined geographic areas and/or time periods is important for investigating ecological, cultural or climatic drivers of demographic processes and for testing spatially and temporally explicit archaeological and historical hypotheses. However, different population genetic approaches differ substantially in the extent to which they allow doing this.

Timing of demographic events

The methods discussed above, including the frameworks that build on the f -statistics, such as TreeMix, AdmixtureGraph and qpGraph, use genetic drift to determine branch lengths in the

inferred population phylogeny, which depends on both time and population size (genetic drift is stronger in small populations than in large ones). This means that there is no absolute time scale associated with any node in the tree (or graph), and nodes in different sub-trees have no well-defined temporal order. Thus, in these trees (or graphs), only the leaves are populations for which known dates exist, while the internal nodes of the tree (or graph), corresponding to ancestral populations and associated demographic events, are abstract in time, making it challenging to link them to archaeological or environmental information.

Approaches such as TreeMix, AdmixtureGraph and qpGraph model admixture by introducing directed gene flow between pairs of branches in the inferred population tree. Sometimes this is not sufficient to explain the pattern of genetic variation in the data, and it is necessary to introduce hypothetical “ghost” populations that contributed to some past admixture events. Similarly to the internal nodes of the tree, dates and the geographic locations of these “ghost” populations are unknown.

Dated ancient samples can help anchoring the internal structure of the tree and provide a way to infer approximate times of internal nodes in the tree and subsequently add some resolution to the inference. However, the extent to which this is effective depends, on the availability of samples in close temporal and geographic proximity to the nodes and ancestral populations of interest.

Explicit modelling of past populations offers a way around this problem. Rasmussen et al. [49] introduced a maximum likelihood method for testing whether an ancient sample can be considered directly ancestral to a given modern sample. This test is based on coalescent theory and does not require explicit modelling of population size changes. However, it does assume an instantaneous split between the population of ancient sample and the ancestral population of the modern sample, with no gene flow between the two groups following the separation. Posth et al. [15] created a temporally explicit coalescent model of the history of late Pleistocene populations in Europe based on the mitochondrial tree of directly dated individuals. The combination of a powerful maximum likelihood framework and samples from both before and after the last glacial maximum allowed them to test the hypothesis of population turnover during this time period using mitochondrial DNA alone.

Linking demographic processes to space

Even with temporally explicit modelling, a key obstacle for interpreting the genetic history of past populations is that the inferred ancestral populations, as well as the “ghost populations”

discussed above, lack well-defined geographic locations. Thus, even though approximate temporal boundaries for these populations can sometimes be informally inferred using genetic data from closely related dated ancient samples, geographic areas remain largely unknown when there is not sufficient geographical coverage of ancient samples to inform about the boundaries of these populations. As a result, it is often challenging to compare the inferred demographic history to the geographically explicit archaeological record.

Overcoming this problem calls for demographic models to be spatially as well as temporally explicit, with geographically well-defined subpopulations through time. Usually this is done by defining demographic processes in terms of local population changes and migrations, where past and present populations are typically represented network of sub-populations (demes) with explicit locations in space. Thus, such explicit simulation methods allow modelling complex population history outside of the phylogenetic framework. In these frameworks, complex spatial patterns can emerge from simple local demographic processes, allowing to explicitly model the effects of population structure and other more subtle or emergent patterns of genetic variation with greater ease. Crucially to archaeological inference, such simulation modelling can incorporate information from various sources (e.g. archaeological, anthropological, demographic or linguistic data and/or climatic and geographic information). For example, estimated population densities from radiocarbon [59,60] or paleoclimate data [61,62] can be used to inform on the relative population densities through time, and presence or absence of different fossils and material cultures can be used to constrain possible geographical ranges. However, such simulation methods require explicit defining of assumptions underlying the demographic model e.g. connectivity of different demes. In particular, the explicit nature of such models makes them represent very specific demographic scenarios, therefore care must be taken that a wide range of plausible scenarios are represented in the analysis. Here it is especially important to make sure that the considered models capture the key patterns of variation observed in the data, for example using the descriptive statistics and other tools described above on simulated data from these models.

Despite their flexibility and theoretical advantages, only relatively few studies have so far used spatially explicit models to reconstruct demographic history due to the fact that such simulation approaches can be complicated to set up and computationally demanding to perform. Early studies used linear steppingstone models to represent founder effects during the expansion of Anatomically Modern Humans out of Africa [63,64] and the levels of shared genetic variation between humans and Neanderthals [53]. Warmuth et al. [65] used a spatially explicit model of

Eurasia to infer the origin and timing of horse domestication. The SPLATCHE2 framework [66] uses a two-dimensional spatially explicit model, which has been used to study interbreeding between humans and Neanderthals [67]. Eriksson et al. [61] used a global spatial model to link late Pleistocene human demography to climate, and this model was later used to test the hypothesis of whether Eskimo-Inuit populations in the Arctic derive from the same population as the original founders of Native American populations [68]. Loog et al. [16] used a spatiotemporally explicit framework to reconstruct wolf demography in the last 50,000 years using mitochondrial genomes. In the past, researchers have relied on inhouse custom simulation code, difficult to adapt to different demographic models and/or data but powerful, easy to use simulation tools such as msprime [69] (for coalescent demographic modelling) & SLiM [70] (forward in time, and can accommodate very complicated demographic scenarios, including genetic selection, and ecological interactions) have been recently created and are now publicly available for researchers to use.

The complex relationships between populations in such models typically mean that formal likelihoods cannot be calculated analytically but can be estimated by comparing descriptive statistics of simulated and observed data using the Approximate Bayesian Computation (ABC) [12,71]. Thus, unless parameters can be constrained using independent data (e.g. historical, anthropological or archaeological information), it is usually necessary to consider a very large number of value combinations for each parameter in order to make reliable inferences about past, adding to the computational cost. Another potential disadvantage of an explicit modelling approach is that it is often time consuming to set up and calibrate (especially for testing more involved demographic scenarios) compared to using a descriptive approaches or summary statistics for demographic inference.

Conclusion

Demographic processes directly affect patterns of genetic variation between and within both populations and individuals. The large body of population genetic theory and mathematical modelling developed to describe these patterns has allowed researchers to take advantage of genetic information from both present and past populations for powerful past demographic inference.

The starting point of this inference is quantification of genetic variation patterns. These patterns can provide some insights and be used for formulating hypothesis about the past demographic processes. However, care must be taken that samples and their relatedness to each other are

well representative of all populations of interest. Patterns of genetic variation, especially when ancient DNA data is included, can provide valuable insights. Although it might be tempting to weave a compelling story based on striking patterns alone, because different demographic scenarios can result in similar spatiotemporal patterns of genetic variation, it is important to formally quantify the likelihood of different demographic scenarios.

Population genetic frameworks that allow formal testing of alternative demographic scenarios usually represent demographic history as a phylogenetic tree, where populations are as independently evolving branches (lineages), sometimes connected by gene flow. Certain methods require additional assumptions (such as the relationships between population, timing of the changes in population sizes and the direction(s) of gene flow) to be explicitly defined. But as these parameters are often unknown, the application of such approaches is limited to more tractable demographic histories. A number of non-parametric approaches have been developed that, on one hand, allow more flexibility but, on the other, come with a number of implicit assumptions that needs taking into consideration when interpreting the results.

In general, population genetic tools that allow formal testing of competing demographic scenarios range from very simple models, with many generalising assumptions, to very complex ones that require explicit modelling decisions. As a general rule, the ability of formal analyses and hypothesis testing to distinguish between complex scenarios is heavily constrained by the amount and type of genetic data available. Very simple models may lack key aspects of demography or not provide sufficient resolution, missing out on important phenomena or leaving the details to be filled by a post-hoc narrative. Complicated models, on the other hand, require much more information (data) to robustly distinguish between different demographic scenarios, or external lines of evidence (such as climate or archaeological information) to guide parameter ranges in the model. Thus, the right tool for the job is determined not only by the questions but also by the available data.

Ancient DNA can provide valuable snapshot of patterns of past genetic variation and has therefore substantially increased the resolution with which different models of past demography can be tested. Although ancient DNA data has been extensively used to address questions about past demography, it's full potential has not been harnessed fully: currently very few population genetic approaches allow formal inclusion of dates and geographic locations of ancient samples at the hypothesis testing phase of the inference, not only missing out on the opportunity to gain additional power to detect more subtle demographic changes, but also leaving the timing and the locations of the inferred demographic events approximate at best.

However, recent developments in simulation (e.g. SLiM [70] and msprime [69]) and analytical tools (e.g.[37,72,73]) will not only make linking past demographic events to the archaeological, historical and climate records more straightforward and robust but will also allow direct inclusion of data from these lines of evidence.

Acknowledgements

I am grateful to Mark Thomas (UCL), Andrea Manica (U. Cambridge) and Anders Eriksson (U. Tartu) for valuable discussions on population genetic modelling using ancient DNA data and Anders Eriksson (U. Tartu) for comments on this manuscript. I am also grateful for the Herchel Smith Trust (University of Cambridge) for supporting this work.

References

1. Pagani L *et al.* 2016 Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**, 238–242. (doi:10.1038/nature19792)
2. Robinson MR *et al.* 2017 Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* **1**, 1–13. (doi:10.1038/s41562-016-0016)
3. Yang DY, Eng B, Wayne JS, Dudar JC, Saunders SR. 1998 Improved DNA extraction from ancient bones using silica-based spin columns. *Am. J. Phys. Anthropol.* **105**, 539–543. (doi:10.1002/(SICI)1096-8644(199804)105:4<539::AID-AJPA10>3.0.CO;2-1)
4. Rohland N, Hofreiter M. 2007 Ancient DNA extraction from bones and teeth. *Nat. Protoc.* **2**, 1756–1762. (doi:10.1038/nprot.2007.247)
5. Dabney J *et al.* 2013 Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci.* **110**, 15758–15763. (doi:10.1073/pnas.1314445110)
6. Skoglund P, Storå J, Götherström A, Jakobsson M. 2013 Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40**, 4477–4482. (doi:10.1016/j.jas.2013.07.004)
7. Kuhn JMM, Jakobsson M, Günther T. 2018 Estimating genetic kin relationships in prehistoric populations. *PLOS ONE* **13**, e0195491. (doi:10.1371/journal.pone.0195491)
8. Cann RL, Stoneking M, Wilson AC. 1987 Mitochondrial DNA and human evolution. *Nature* **325**, 31–36. (doi:10.1038/325031a0)
9. Underhill PA, Passarino G, Lin AA, Shen P, Lahr MM, Foley RA, Oefner PJ, Cavalli-Sforza LL. 2001 The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43–62. (doi:10.1046/j.1469-1809.2001.6510043.x)
10. Chan EKF *et al.* 2019 Human origins in a southern African palaeo-wetland and first migrations. *Nature* **575**, 185–189. (doi:10.1038/s41586-019-1714-1)
11. Gerbault P *et al.* 2014 Storytelling and story testing in domestication. *Proc. Natl. Acad. Sci.* **111**, 6159–6164. (doi:10.1073/pnas.1400425111)

12. Rosenberg NA, Nordborg M. 2002 GENEALOGICAL TREES, COALESCENT THEORY AND THE ANALYSIS OF GENETIC POLYMORPHISMS. *Nat. Rev. Genet.* **3**, 380–390. (doi:10.1038/nrg795)
13. Ballard JWO, Whitlock MC. 2004 The incomplete natural history of mitochondria. *Mol. Ecol.* **13**, 729–744.
14. Nielsen R, Beaumont MA. 2009 Statistical inferences in phylogeography. *Mol. Ecol.* **18**, 1034–1047. (doi:10.1111/j.1365-294X.2008.04059.x)
15. Posth C *et al.* 2016 Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe. *Curr. Biol.* **26**, 827–833. (doi:10.1016/j.cub.2016.01.037)
16. Loog L *et al.* 2019 Ancient DNA suggests modern wolves trace their origin to a Late Pleistocene expansion from Beringia. *Mol. Ecol.* **n/a**. (doi:10.1111/mec.15329)
17. Larson G *et al.* 2007 Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc. Natl. Acad. Sci.* **104**, 15276–15281. (doi:10.1073/pnas.0703411104)
18. Valdiosera CE *et al.* 2007 Staying out in the cold: glacial refugia and mitochondrial DNA phylogeography in ancient European brown bears. *Mol. Ecol.* **16**, 5140–5148. (doi:10.1111/j.1365-294X.2007.03590.x)
19. Thalmann O *et al.* 2013 Complete Mitochondrial Genomes of Ancient Canids Suggest a European Origin of Domestic Dogs. *Science* **342**, 871–874. (doi:10.1126/science.1243650)
20. Li JZ *et al.* 2008 Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319**, 1100–1104. (doi:10.1126/science.1153717)
21. Novembre J *et al.* 2008 Genes mirror geography within Europe. *Nature* **456**, 98–101. (doi:10.1038/nature07331)
22. McVean G. 2009 A Genealogical Interpretation of Principal Components Analysis. *PLoS Genet.* **5**, e1000686. (doi:10.1371/journal.pgen.1000686)
23. Skoglund P, Sjödin P, Skoglund T, Lascoux M, Jakobsson M. 2014 Investigating Population History using Temporal Genetic Differentiation. *Mol. Biol. Evol.* , msu192. (doi:10.1093/molbev/msu192)
24. Pritchard JK, Stephens M, Donnelly P. 2000 Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945–959.
25. Alexander DH, Novembre J, Lange K. 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664. (doi:10.1101/gr.094052.109)
26. Lawson DJ, Hellenthal G, Myers S, Falush D. 2012 Inference of Population Structure using Dense Haplotype Data. *PLOS Genet.* **8**, e1002453. (doi:10.1371/journal.pgen.1002453)
27. Wang S *et al.* 2008 Geographic Patterns of Genome Admixture in Latin American Mestizos. *PLOS Genet.* **4**, e1000037. (doi:10.1371/journal.pgen.1000037)
28. Lawson DJ, Dorp L van, Falush D. 2018 A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* **9**, 1–11. (doi:10.1038/s41467-018-05257-7)
29. Dorp L van *et al.* 2015 Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLOS Genet.* **11**, e1005397. (doi:10.1371/journal.pgen.1005397)
30. Wright S. 1949 The Genetical Structure of Populations. *Ann. Eugen.* **15**, 323–354. (doi:10.1111/j.1469-1809.1949.tb02451.x)
31. Wright S. 1931 Evolution in Mendelian Populations. *Genetics* **16**, 97–159.

32. McRae BH, Beier P. 2007 Circuit theory predicts gene flow in plant and animal populations. *Proc. Natl. Acad. Sci.* **104**, 19885–19890. (doi:10.1073/pnas.0706568104)
33. McRae BH, Dickson BG, Keitt TH, Shah VB. 2008 Using Circuit Theory to Model Connectivity in Ecology, Evolution, and Conservation. *Ecology* **89**, 2712–2724. (doi:10.1890/07-1861.1)
34. Malaspinas A-S. 2016 Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. *Mol. Ecol.* **25**, 24–41. (doi:10.1111/mec.13492)
35. Petkova D, Novembre J, Stephens M. 2016 Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100. (doi:10.1038/ng.3464)
36. Mathieson I *et al.* 2018 The genomic history of southeastern Europe. *Nature* **555**, 197–203. (doi:10.1038/nature25778)
37. Loog L, Lahr MM, Kovacevic M, Manica A, Eriksson A, Thomas MG. 2017 Estimating mobility using sparse data: Application to human genetic variation. *Proc. Natl. Acad. Sci.* , 201703642. (doi:10.1073/pnas.1703642114)
38. Hey J, Nielsen R. 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci.* **104**, 2785–2790. (doi:10.1073/pnas.0611164104)
39. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013 Robust Demographic Inference from Genomic and SNP Data. *PLOS Genet.* **9**, e1003905. (doi:10.1371/journal.pgen.1003905)
40. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011 Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034. (doi:10.1038/ng.937)
41. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genet.* **5**, e1000695. (doi:10.1371/journal.pgen.1000695)
42. Freedman AH *et al.* 2014 Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLOS Genet* **10**, e1004016. (doi:10.1371/journal.pgen.1004016)
43. Li H, Durbin R. 2011 Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496. (doi:10.1038/nature10231)
44. Sheehan S, Harris K, Song YS. 2013 Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics* **194**, 647–662. (doi:10.1534/genetics.112.149096)
45. Schiffels S, Durbin R. 2014 Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925. (doi:10.1038/ng.3015)
46. Mazet O, Rodríguez W, Grusea S, Boitard S, Chikhi L. 2016 On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity* **116**, 362–371. (doi:10.1038/hdy.2015.104)
47. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009 Reconstructing Indian population history. *Nature* **461**, 489–494. (doi:10.1038/nature08365)
48. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012 Ancient Admixture in Human History. *Genetics* **192**, 1065–1093. (doi:10.1534/genetics.112.145037)
49. Rasmussen M *et al.* 2014 The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229. (doi:10.1038/nature13025)
50. Green RE *et al.* 2010 A Draft Sequence of the Neandertal Genome. *Science* **328**, 710–722. (doi:10.1126/science.1188021)

51. Castelo R, Roverato A. 2012 Inference of regulatory networks from microarray data with R and the bioconductor package qgraph. *Methods Mol. Biol. Clifton NJ* **802**, 215–233. (doi:10.1007/978-1-61779-400-1_14)
52. Lazaridis I *et al.* 2014 Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413. (doi:10.1038/nature13673)
53. Eriksson A, Manica A. 2012 Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl. Acad. Sci.* **109**, 13956–13960. (doi:10.1073/pnas.1200567109)
54. Malmström H *et al.* 2009 Ancient DNA Reveals Lack of Continuity between Neolithic Hunter-Gatherers and Contemporary Scandinavians. *Curr. Biol.* **19**, 1758–1762. (doi:10.1016/j.cub.2009.09.017)
55. Bramanti B *et al.* 2009 Genetic Discontinuity Between Local Hunter-Gatherers and Central Europe's First Farmers. *Science* **326**, 137–140. (doi:10.1126/science.1176869)
56. Hofmanová Z *et al.* 2016 Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci.* **113**, 6886–6891. (doi:10.1073/pnas.1523951113)
57. Saag L *et al.* 2019 The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Curr. Biol.* **29**, 1701-1711.e16. (doi:10.1016/j.cub.2019.04.026)
58. Silva NM, Rio J, Currat M. 2017 Investigating population continuity with ancient DNA under a spatially explicit simulation framework. *BMC Genet.* **18**, 114. (doi:10.1186/s12863-017-0575-6)
59. Shennan S, Downey SS, Timpson A, Edinborough K, Colledge S, Kerig T, Manning K, Thomas MG. 2013 Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nat. Commun.* **4**, 2486. (doi:10.1038/ncomms3486)
60. Bevan A, Colledge S, Fuller D, Fyfe R, Shennan S, Stevens C. 2017 Holocene fluctuations in human population demonstrate repeated links to food production and climate. *Proc. Natl. Acad. Sci.* **114**, E10524–E10531. (doi:10.1073/pnas.1709190114)
61. Eriksson A, Betti L, Friend AD, Lycett SJ, Singarayer JS, Cramon-Taubadel N von, Valdes PJ, Balloux F, Manica A. 2012 Late Pleistocene climate change and the global expansion of anatomically modern humans. *Proc. Natl. Acad. Sci.* **109**, 16089–16094. (doi:10.1073/pnas.1209494109)
62. Timmermann A, Friedrich T. 2016 Late Pleistocene climate drivers of early human migration. *Nature* **538**, 92–95. (doi:10.1038/nature19365)
63. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15942–15947. (doi:10.1073/pnas.0507611102)
64. Liu H, Prugnolle F, Manica A, Balloux F. 2006 A Geographically Explicit Genetic Model of Worldwide Human-Settlement History. *Am. J. Hum. Genet.* **79**, 230–237.
65. Warmuth V *et al.* 2012 Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 8202–8206.
66. Ray N, Currat M, Foll M, Excoffier L. 2010 SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics* **26**, 2993–2994. (doi:10.1093/bioinformatics/btq579)

67. Currat M, Excoffier L. 2011 Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proc. Natl. Acad. Sci.* **108**, 15129–15134. (doi:10.1073/pnas.1107450108)
68. Raghavan M *et al.* 2015 Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* , aab3884. (doi:10.1126/science.aab3884)
69. Kelleher J, Etheridge AM, McVean G. 2016 Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Comput. Biol.* **12**, e1004842. (doi:10.1371/journal.pcbi.1004842)
70. Haller BC, Messer PW. In press. SLiM: An Evolutionary Simulation Framework. , 578.
71. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. 2013 Approximate Bayesian Computation. *PLoS Comput Biol* **9**, e1002803. (doi:10.1371/journal.pcbi.1002803)
72. Racimo F, Woodbridge J, Fyfe RM, Sikora M, Sjögren K-G, Kristiansen K, Linden MV. 2019 A geostatistical approach to modelling human Holocene migrations in Europe using ancient DNA. *bioRxiv* , 826149. (doi:10.1101/826149)
73. Mathieson I *et al.* 2015 Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503. (doi:10.1038/nature16152)