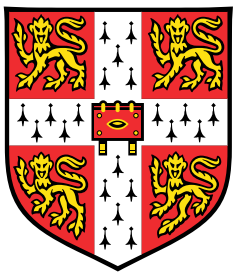
Ancestral Paths: Redefining local genetic ancestry and its inference with application to Europeans



Alice Pearson

Department of Genetics

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Downing College

September 2022

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Alice Pearson

September 2022

Ancestral Paths: Redefining local genetic ancestry and its inference with application to Europeans

Alice Pearson

Recently, two new approaches have transformed our understanding of human population history. Firstly, the sequencing of ancient genomes which gives us a snapshot of past genetic variation. We can therefore make inferences from observed genetic signatures present before historical events such as population bottlenecks and natural selection have obscured them from the modern gene pool. Ancient DNA has thus revealed what cannot be determined from modern genomes alone. Secondly, the development of methods that aim to reconstruct population genealogies from genetic variation data. Together with an understanding of how evolutionary processes alter genealogies, this has allowed inference of historical and ongoing processes in real world populations. The latest updates in these approaches now allow us to combine the two and infer genealogies involving both present-day and ancient individuals.

In this thesis I present a new method to infer local ancestry along sample chromosomes. The method applies machine learning to tree sequences built from ancient and present-day genomes and is based on a deterministic model of population structure, within which I introduce the concept of 'path ancestry'. I show with extensive simulation that the method is robust to a variety of demographic scenarios and generalises over model misspecification. Subsequent downstream analyses include estimating past effective population size, timing of population specific selection and the time since admixture for individuals. I apply the method to a large ancient DNA dataset covering Europe and West Eurasia to paint all sample chromosomes. I show that the inferred admixture ages are a better metric than sample ages alone for understanding movements of people across Europe in the past.

Acknowledgements

I would like to start by acknowledging Richard. I have been fortunate to have such a thoughtful and patient supervisor, who has always provided insightful comments and suggestions in our meetings.

I would also like to thank my parents. Visiting my home in Dorset has helped me get through the most difficult periods and reminds me who I am.

Finally I want to acknowledge Derek, my husband. During my PhD he has lived a transatlantic existence, putting his life in the USA on pause. I will always be grateful to him for his support and sacrifice.

Abstract

Recently, two new approaches have transformed our understanding of human population history. Firstly, the sequencing of ancient genomes which gives us a snapshot of past genetic variation. We can therefore make inferences from observed genetic signatures present before historical events such as population bottlenecks and natural selection have obscured them from the modern gene pool. Ancient DNA has thus revealed what cannot be determined from modern genomes alone. Secondly, the development of methods that aim to reconstruct population genealogies from genetic variation data. Together with an understanding of how evolutionary processes alter genealogies, this has allowed inference of historical and ongoing processes in real world populations. The latest updates in these approaches now allow us to combine the two and infer genealogies involving both present-day and ancient individuals.

In this thesis I present a new method to infer local ancestry along sample chromosomes. The method applies machine learning to tree sequences built from ancient and present-day genomes and is based on a deterministic model of population structure, within which I introduce the concept of 'path ancestry'. I show with extensive simulation that the method is robust to a variety of demographic scenarios and generalises over model misspecification. Subsequent downstream analyses include estimating past effective population size, timing of population specific selection and the time since admixture for individuals. I apply the method to a large ancient DNA dataset covering Europe and West Eurasia to paint all sample chromosomes. I show that the inferred admixture ages are a better metric than sample ages alone for understanding movements of people across Europe in the past.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Contents

Genome sequences are a product of their evolutionary history and therefore knowledge of how genomes have evolved is important to understanding how they work and how they influence our lives today. Population genetics analysis underpins the pursuit of this knowledge and arguably provides the deepest comprehension of evolutionary change of any branch of biology.

Mutations are the foundational unit of study for all population genetic analysis. Mutations are changes in the DNA sequence which take place either before or during the DNA copying process and are then passed on in future DNA replications. The process of mutation produces variant forms of genes termed 'alleles'. Evolution can be thought of as changes in the frequency of alleles in a given population over time. Mutation therefore creates the heritable variation in a population required for evolution. After many generations evolution can create marked differences in allele frequencies between populations. The evolutionary processes which drive changes in allele frequency will be described in Section 1.1.

A fundamental practice within population genetics is ancestry inference [90], as it is often necessary in many population genetics applications to assign a set of individuals to different population groups [20, 119, 11]. Ancestry inference techniques exploit the allele frequency differences to delineate populations or 'ancestries' and/or assign individuals to populations. However, the concept of ancestry is not well established [76]. Its definition is often context dependent which can confuse both researchers and the public, thereby making its deeper meaning in some applications questionable.

There are two tools that are central to analysis of genetic data. The first is DNA sequencing which allows observation of the mutations present in individuals. The advent of next-generation sequencing technologies that came with advances in computer science and molecular biology created an exponentially growing pool of data from modern genomes, capturing millions of genetic variants. Single nucleotide polymorphisms are the most commonly used type of genetic variant. These are substitutions of a single nucleotide for another which are present in at least 1% of the population. Dense SNP data is now available from ancient genomes which allows tracking of mutations and allele frequencies over space and time (Section 1.2).

The second tool is mathematical modelling of how genes evolve in populations over time (Section 1.3). The evolutionary processes that drive evolution are well understood

(Section 1.1) and DNA replication imposes strong constraints so mathematical models of how genomes change over time are easy to construct. While simplified models will fail to capture the full complexity of the natural world, many simple models behave in similar ways to the corresponding complex system and so can be used to make rich inferences. It is only with these models that inferences about the evolutionary history of samples can be gleaned from the mutations observed with DNA sequencing. A relevant example is genealogical inference tools which can infer the genealogical relationships between large sets of sample DNA sequences using SNP data [117, 53].

My thesis combines a collection of aforementioned tools and principles to advance ancestry inference and from this our understanding of the evolutionary history of present day Europeans. I first redefine the meaning of ancestry to be more flexible by considering time as a dimension. With this new definition, I then develop a new method for inferring local genetic ancestry that uses the central tools of population genetics; dense SNP data from modern and ancient Eurasian genomes, and a recent genealogical inference program, RELATE [117, 116]. Finally, I show how to carry out various analyses using these.

1.1 Evolutionary processes and population structure

There are four evolutionary processes that act at the level of populations that affect the number of mutations present in a population, referred to as the allele frequency. The first is mutation. As described above, this process introduces new variation into the population. The next is gene flow, where alleles are introduced into a population by migrants from another population. Gene flow either changes the frequency of alleles already present in the receiving population or introduces new alleles.

Genetic drift and selection, the third and fourth processes, affect the representation of alleles in the next generation and so alter the allele frequency between generations. Genetic drift is the random sampling of alleles from the pool of parental alleles to form the next

generation. Sampling starts at the gamete level where only a small sample of the possible gametes that are produced by an organism are fertilised. Drift then continues at all life stages from fertilisation to reproduction where random events (not related to allele fitness) remove individuals from the populations. Over many generations this results in stochastically fluctuating allele frequencies over time. At a single biallellic locus drift can be modelled using the binomial distribution where p is the allele frequency of allele A, $1-p$ is the frequency of allele B and N, the constant haploid population size, is the number of trials. The expected frequency of allele A in the next generation p' is p and its variance is

$$\mathbf{Var}(p') = \frac{p(1-p)}{2N}.$$

When under drift alone, an allele will either be completely lost from the population or become present in all members of the population (fixation) given sufficient time. The probability of fixation of an allele is equal to its initial frequency. Given that the variance is inversely proportional to population size, the allele frequencies will fluctuate more rapidly in small populations and therefore alleles will be fixed (or lost depending on their initial frequency) more rapidly, i.e drift has a stronger effect in smaller populations. This model of binomial sampling works under the assumption of a highly idealised population. Real world populations never fit all assumptions which almost always results in more drift being observed in these populations than an idealised population of the same size. The effective size (N_e) of a real-world population is the size of an idealised population that would experience the same amount of genetic drift. N_e is a fundamental concept in population genetics as it allows the above model of binomial sampling to be applied to real world populations.

Selection is a bias in the sampling of alleles from the population, where some alleles tend to be oversampled or undersampled in the next generation. For example, if an allele confers resistance to a disease, its fitness (tendency to perpetuate into future generations) will increase relative to the non-resistant alleles. The expected frequency of allele A in the next

generation p' is no longer p, but is weighted by the fitness of the allele W_a,

$$\mathbf{E}[p'] = p\frac{W_a}{\bar{W}},$$

where $\bar{W}$ is the average fitness of all alleles in the population. It is therefore the relative fitness of an allele that biasses the sampling of that allele.

Selection acts only when there is a difference in fitness between alleles and causes an allele to increase or decrease in frequency over time, with a proclivity towards one direction, unlike genetic drift which is random. Genetic drift is always acting so alleles under selection will also fluctuate in frequency around the expected value. This means that when selection and drift act concurrently, a positively selected allele can drift to loss by chance and likewise a deleterious allele can drift to fixation. Given that the variance in allele frequency under drift is inversely proportional to N_e, it is therefore the effective population size that determines the efficacy of selection.

Population structure is often defined as a difference in allele frequencies between groups or populations. This occurs when individuals cannot randomly mate and the assumption of panmixia is violated. This can be because groups are totally separated by an uncrossable barrier such as a mountain range, or this can be more continuous where different parts of a large population have different allele frequencies because the whole population is not panmictic. Over time, evolutionary processes acting separately in the divided groups will create different allele frequencies in the two groups.

Genetic drift is the most influential process that gives rise to divergent allele frequencies in structured populations because not only is it always acting, but it is always acting on all alleles. Furthermore it acts entirely without a directional bias as to whether alleles increase or decrease in frequency at each generation. The probability that an allele takes the same frequency trajectory in separate populations becomes vanishingly small after multiple generations. When the N_e of different groups is small the rate of divergence in frequencies

is greater. If the groups experience different environments, selection on different alleles will also cause different allele frequencies. Lastly, mutations that appear after separation contribute to differing frequencies.

Gene flow counteracts drift by homogenising allele frequencies. For example, consider a simple model of one infinitely sized mainland and one small island population, with migration from mainland to the island but not from island to mainland. This example captures the effect of gene flow acting against drift. The allele frequency of an allele on the mainland population is p_x and p_y on the island population. The mainland population is infinitely large and so the frequencies of alleles are constant. In every generation, fraction m of the island individuals are migrants from the mainland and fraction $1-m$ were already on the island. The expected allele frequency in the island individuals at time $t+1$ is given by,

$$p_{y,t+1} = (1-m)p_{y,t} + mp_x.$$

Over time, the allele frequency in the island population will approach the frequency of the mainland. The time it takes for the frequencies to converge depends on the difference between the starting frequency in the island and the mainland $p_{y,t=0} - p_x$ and the migration rate m. This demonstrates how gene flow erodes population structure by reducing the difference in allele frequencies between groups. Admixture is when gene flow occurs between groups that were previously isolated, or structured.

Detecting population structure fundamentally involves detecting differences in allele frequencies. One way to do this is to measure heterozygosity. Wright's F_{st} statistic measures relative mean heterozygosity between biallelic loci in subdivided populations [126]. When populations split, alleles are more likely to reach fixation by drift and therefore the mean heterozygosity within populations stays the same or decreases. When this happens the mean total heterozygosity, imagining all individuals belong to the same randomly mating

population, increases. F_{st} is calculated from these heterozygosity measures as,

$$F_{st} = \frac{H_T - H_S}{H_T} = 1 - \frac{H_S}{H_T},$$

where H_S is the average of the heterozygosities within subpopulations, $\frac{1}{K}\sum_{i=1}^{K} 2p_i q_i$ with K subpopulations, and H_T is the total heterozygosity from the average of allele frequencies between subpopulations, $2\bar{p}\bar{q}$.

Another commonly used method for detecting population structure is dimensionality reduction, in particular Principal Component Analysis (PCA) [81]. The genotypes of individuals can be coded as integers by the number of non-reference alleles present, making them well-suited for PCA. These integers are normalised using the allele frequency and fill a matrix of N x S entries, where N is the number of diploid samples and S is the number of sites. PCA transforms the data to maximise variation and produce principal components (PCs), new variables that are linear combinations of the initial variables. The PCs are uncorrelated and are ordered in decreasing amount of variation they explain. When samples are plotted together by their first two PCs, distinct clusters form which are attributed to discrete populations.

I make use of both F_{st} statistics and PCA when assessing population structure in Chapter 2. In Chapter 5 I highlight some more approaches for detecting population structure given variation data.

1.2 Ancient DNA

Despite a wealth of data, patterns in modern genomes alone are difficult to interpret as they are an indirect measure of past events. Moreover, much of the genetic variation which was present in past populations does not exist in the modern gene pool due to ancient demographic events that exacerbate drift, such as bottlenecks and isolation. Recent technological advances

have enabled the sequencing of DNA extracted from ancient specimens, so called ancient DNA (aDNA). Ancient DNA gives us a genetic snapshot of a time before confounding processes have taken place, meaning we regain some of the lost information. We gain a lower bound on when genetic mutations first appeared and insight into what genetic changes may have taken place over time, helping us to infer what events and forces have acted to produce the patterns in modern genomes. The use of aDNA has transformed our understanding of human origins and evolution in recent years ([113]).

Despite its great utility, ancient DNA is notoriously difficult to sequence and analyse. After death of an organism, DNA repair mechanisms no longer function and over time post-mortem damage accumulates. Cytosine bases deaminate to uracil on 5' and 3' terminal overhangs and appear as C to T changes on the forward strand and G to A changes on the reverse strand when sequenced, which can mask informative mutations [13]. Also, double stranded breaks accumulate through oxidation and enzyme activity, creating fragments that are too short to align. In addition, ancient DNA samples are often contaminated with DNA fragments from microbes and modern human DNA. A number of processing steps are needed to identify and extract contaminating DNA but it still can create issues with data authenticity. Overall, these challenges result in low coverage of the endogenous genome in ancient samples.

Two approaches exist that aim to boost the data available from ancient DNA in order to perform many important analyses. The first is to enrich the dataset for known genotypes of interest by designing specific probes that capture fragments containing target loci, which can then be amplified and sequenced, so called capture sequencing. A target set of approximately 1.2 million SNPs is commonly used in assays for human genomes and the 1240K dataset is free to download and consists of thousands of ancient and present-day individuals genotyped at these positions. This gives access to genome-wide data from ancient samples with low amounts of endogenous human DNA present and can increase efficiency by targeting sites

that will actually be analysed. The standardised set of SNPs also makes it easy to analyse new sample data with previously published samples. The major drawbacks with capture are that analysis is limited to previously ascertained SNPs and potentially important information is lost at sites in between. Additionally, the non-random selection of SNPs to include in the assay leads to biasses in downstream analysis when compared to whole genome sequence data [58]. Plus, systematic biases can be introduced at the molecular level, where the likelihood of capture with a probe is not independent of the variant present.

The second strategy is to infer the unobserved genotypes that are missing between those that are genotyped, called genotype imputation. Imputation is often applied to modern genomes to boost power and lower the cost in GWAS studies [70]. When applied to sparse ancient genotype data it greatly increases the information content and the range of analyses that can be performed, including haplotype-based approaches as well as allele frequency methods. Plus it can be applied to shotgun sequenced data with genotype calls, which mitigates biases incurred from capture sequencing.

Imputation is based on linkage disequilibrium (LD) of haplotypes. LD refers to the non-random association between alleles at different loci [112]. The extent of LD and the distance between loci that are associated is influenced by many factors including selection, mutation, drift, population structure, admixture and genetic linkage [112]. Recombination is the exchange of genetic material between homologous chromosomes during meiosis which breaks associations between loci on either side of a recombination break point, acting to decay LD. When many loci are considered, haplotype blocks containing multiple SNPs in LD appear, separated by recombination hotspots [28].

The correlation between SNPs means that the presence of unobserved alleles can be predicted from those that are present at loci in LD. In order to build models of the haplotype structure from which to impute SNPs at unobserved loci a reference set of high coverage phased genomes is required. The basic principle of many imputation tools is to learn the

associations between SNPs from a reference panel and then, using various approaches, predict the presence of unobserved SNPs given the observed SNPs in query sequences. Predictions can either be made from hard genotype calls in query sequences or using the genotype likelihood scores with a probabilistic framework [14, 46, 103].

The major drawback of imputation of ancient samples is that it is impossible to recover alleles that have been lost between the sample age and the present-day due to drift. Similarly rare alleles of approximately <0.1% minor allele frequency are difficult to impute. Using large global reference panels to include more diversity with the addition of high coverage ancient samples, it is possible to achieve high imputation accuracy [6]. Importantly, imputation has the potential to remove systematic bias from capture data by removing genotype errors, plus it has been shown to reduce the affect of ascertainment bias [32]. This means that data from capture and shotgun sequencing can be merged through imputation, increasing sample sizes for a wide range of analyses.

In Chapter 2, I introduce a large dataset of ancient genomes that has been imputed to whole-genome sequence data and later I will discuss the possible effects imputation is having on my analyses.

1.3 Properties of the coalescent

Originally introduced by Kingman [55], the coalescent is the name given to the process that underlies statistical inference of genealogies. It traces lineages backwards in time, modelling how they coalesce, until one lineage remains, the most recent common ancestor (MRCA).

The coalescent generates a probability distribution over trees, and allows a probability to be given for any tree. It is applied to a sample of individuals from a population with the assumption that the history of the sample is a smaller genealogy embedded within the whole

population genealogy and so demographic parameters of the population are reflected in the sample genealogy.

For the discrete time coalescent, in a population with a diploid effective population size N_e, the probability that two lineages share a common ancestor (coalesce) in the previous generation is $\frac{1}{2N_e}$. Consequently, the probability that they do not coalesce is $1 - \frac{1}{2N_e}$. Hence, the probability of two lineages coalescing an arbitrary number of generations in the past t is given by

$$P(\text{Coal = t}) = \left(1 - \frac{1}{2N_e}\right)^{(t-1)} \left(\frac{1}{2N_e}\right).$$

This is a geometric series with parameter $\frac{1}{2N_e}$, so the expected time to coalescence for two sequences is $2N_e$.

Above is the coalescent for two sequences. For a sample of more than two sequences, the n-coalescent, we also consider the number of ways of choosing two sequences from the sample $\binom{n}{2} = \frac{n(n-1)}{2}$, where n is the sample size. The probability that any two sequences coalesce in the previous generation is therefore:

$$P(Coal) = \frac{n(n-1)}{2} \left(\frac{1}{2N_e}\right).$$

The same logic as above follows to calculate the probability of two lineages coalescing at t generations,

$$P(\text{Coal = t}) = \left(1 - \frac{n(n-1)}{2} \left(\frac{1}{2N_e}\right)\right)^{(t-1)} \frac{n(n-1)}{2} \left(\frac{1}{2N_e}\right),$$

and the expected time to a coalescence event is:

$$\mathbf{E}[t] = \frac{4N_e}{n(n-1)}.$$

Often it is conceptually and computationally advantageous to scale the discrete coalescent in continuous time, so that one unit of time equates to the expected time for two sequences to find a common ancestor, $2N_e$. This makes the continuous-time coalescent independent of population size. The waiting time t^c for a coalescence event between n sequences is modelled by the exponential distribution with rate $\binom{n}{2}$,

$$P(\text{Coal} \leq t^c) = 1 - e^{\binom{n}{2}t^c} \tag{1.1}$$

t^c can be converted back to time in generations t by multiplying by the population size $t = 2N_e t^c$. The continuous-time coalescent is broadly applied and is referred to as the basic coalescent.

The coalescent describes tree topologies onto which mutations can be added to model sequences at the leaf nodes. The number of neutral mutations on a branch is sampled from a Poisson distribution with intensity $\frac{t\theta}{2}$, where t is the branch length and $\theta = 4\mu N_e$ is the scaled mutation rate, where N_e is the population size and μ is the per sequence per generation mutation rate. The timing of mutations along each branch is random and all branches are independent. With a known tree (topology and branch lengths, T), one can calculate the probability of a tree with mutations as

$$P(T|Mutations) = P(T)P(Mutations|T), \tag{1.2}$$

where $P(T)$ is the basic coalescent prior (equation 1.1) and $P(Mutations|T)$ is a product of Poisson distributions across all branches.

However, when the tree is unknown, to calculate the probability of a set of sequences, one must integrate over all possible trees,

$$P(Mutations) = \int_T P(T)P(Mutations|T)\, dT.$$

The number of possible tree topologies and branch lengths become prohibitively large as the number of sample sequences increases. MCMC, importance sampling and other techniques have been employed in the past in order approximate the integral over all possible genealogies [57, 44, 37].

The basic coalescent is founded on assumptions of random mating, constant population size and no selection. The structure of genealogies will change when these assumptions are not met, therefore the basic coalescent is often extended to model these deviations. Throughout my analysis I rely on data produced by a coalescent simulator, msprime [52], and a genealogy inference tool that uses coalescent priors, RELATE [117, 116]. Additionally, I use the basic coalescent in my methods to infer demographic parameters in Chapter 5.

Another assumption of the basic coalescent is the absence of recombination. Recombination means that a set of sample sequences cannot be related to each other by a single coalescent tree but a graph. A recombination event splits genetic material backwards in time into two ancestors while coalescent events join genetic material from two sequences in one ancestor. Hudson [47] was the first to model the coalescent with recombination with a 'breadth first' approach, backwards in time. Sequences coalesce and recombine with waiting times sampled from two competing exponential distributions. The result is a complicated graph that is very difficult to infer, which I elaborate on in the next sections. I introduce later models of the coalescent with recombination in Chapter 5.

1.4 Ancestral Recombination Graph estimates as tree sequences

During meiosis, genetic material is exchanged between homologous chromosomes in the process of recombination. A recombination event will cause a single chromosome to be inherited from two parents in the previous generation which, looking backwards in time,

appears to split a lineage into two lineages. Conversely, when two chromosomes share a common ancestor in the past the two lineages will join, backwards in time, in a coalescence event. The full collection of coalescences and recombination events for a set of sample chromosomes is called the Ancestral Recombination Graph (ARG) and it entirely describes the history of relationships of those sequences. In the absence of recombination, a set of sequences can be related to each other by a single tree containing all their coalescences as the internal nodes. The tree topology can be inferred by the pattern of shared derived mutations between the sequences given sufficient mutations. However, if there is a recombination event in any of the ancestors, each piece either side of the breakpoint has a different pattern of coalescences and therefore we observe two different trees relating the sample sequences. Multiple ancestral recombination events produce a sequence of changing trees as you move from one end of the sample chromosomes to the other, each encoding the genealogy of a chunk of DNA and each tree change reflecting one or more recombination events. Trees located closer together on the chromosome will be correlated, meaning they share more edges as fewer recombination events have occurred between them. In this way a tree sequence represents the outcome of recombination, in contrast to an ARG which represents the recombination events themselves. ARGs contain many nodes that do not alter the genealogy. It is inefficient to store all of these events so only those that change the tree topology are stored in the tree sequence [101].

Knowledge of the full genealogical history of many whole genome sample sequences would greatly improve our ability to answer questions about their evolutionary history. A lot of effort has therefore gone into devising methods to as accurately as possible infer ARGs from genomic sequences. ARGs however are notoriously hard to infer, especially for larger sample sizes. This is mainly because the number of possible ARGs relating samples is very large and increases rapidly with sample size, plus the information in genome sequences is often insufficient to choose a specific ARG above all others. Sampling from the posterior

distribution of ARGs given the observed genetic data and a set of model assumptions, as is the approach of ARGweaver [102], is computationally very expensive. Two methods that were introduced in 2019 use heuristic approaches to infer tree sequences between many DNA sequences, namely RELATE [117] and tsinfer [53]. The latest release of both these programmes enables sample sequences to be placed in the past.

1.5 RELATE inference of tree sequences

In this section I describe the RELATE method of inferring tree sequences because Chapter 2 is a direct application of RELATE.

The RELATE software aims to infer the true underlying tree sequence from phased genotype data. Unlike ARGweaver, RELATE is scalable to thousands of sample sequences and offers the ability to infer population size changes through time under a panmictic model. The method can be divided in several stages: Firstly, RELATE constructs a distance matrix at every site that stores the probability of each haplotype i copying from all other haplotypes j. Because recombination will change these probabilities, RELATE calculates them using SNP information flanking each focal SNP with a modified Li and Stephens hidden Markov model which considers ancestral and derived states.

From the distance matrices RELATE uses hierarchical clustering to build a tree topology which finds clusters of haplotypes that coalesce with each other before other clusters. This can be done for every distance matrix at every SNP but is very inefficient and slow, especially for large sequences. To improve efficiency, RELATE makes use of the fact that trees at neighbouring sites are likely to be very similar if not identical, given no recombination has happened between them. Starting from one end of a chromosome RELATE constructs a tree topology and maps mutations to the branches, where all descendants below the branch with a mutation on it are carriers. Only when a mutation at the next focal SNP cannot be uniquely mapped to the topology of the previously constructed tree does RELATE proceed

to estimate a new topology at that SNP. To be robust to sequencing errors the constraint of 'uniquely mapping' is relaxed so that not all descendants of a branch must be derived for the mutation. This has the effect of increasing the speed of the algorithm, but also results in an underestimation of the number of trees.

After the first stages, RELATE has constructed rooted binary trees that adapt to changes in local genetic ancestry due to recombination. Next RELATE infers branch lengths, first for constant population size, using an iterative Markov Chain Monte Carlo (MCMC) algorithm. Initially the order of coalescence events is chosen randomly but within topological constraint. The likelihood of the tree $P(\mathbf{t}|\mathbf{m})$, with branch lengths $\mathbf{t} = (t_b)_{b=0,..,2N-2}$ conditional on the mutations $\mathbf{m} = (m_b)_{b=0,..,2N-2}$, using equation 1.2, is given by

$$P(\mathbf{t}|\mathbf{m}) = P(\mathbf{t})P(\mathbf{m}|\mathbf{t}) = P(\mathbf{t}) \prod_{b=0}^{2N-2} P(m_b|t_b). \tag{1.3}$$

The algorithm proceeds by proposing a change in the order of coalescence (again within topological constraint) or a change in the time while k lineages exist. The likelihood of the new proposed configuration is compared to that of the current configuration. Only the branches that have lengths altered by the change need to be included in the likelihood calculation which improves efficiency. The change is accepted if the likelihood of the new configuration is greater than the existing configuration or if the ratio is greater than a uniformly sampled acceptance threshold less than one. After enough proposals the stationary distribution, $P(\mathbf{t}|\mathbf{m})$, representing a maximum likelihood set of branch lengths given the mapping mutations, is reached.

The initially inferred branch lengths can then be used in an add on module that re-estimates branch lengths, this time under a model of variable population size through time. Maximum likelihood, population-wide coalescence rates are calculated for time epochs for the current tree configuration. These are then incorporated into the above MCMC process when calculating the likelihood of proposed changes, therefore re-estimating the branch

lengths but for a variable population size. This is iterated until convergence. After the final re-estimation, changes in population size between the time epochs can be inferred given that population size is inversely proportional to the final calculated maximum likelihood coalescence rates.

A recent update to RELATE now means that ancient genomes can be built into the tree sequences [116]. Sample ages passed to RELATE are used to constrain the coalescences of lineages from ancient samples to only times and only with lineages that precede the date of the sample. Additionally branch lengths must be measured from the ancient sample age to the next coalescence during likelihood calculations.

1.6 A review of European population structure

During the last glacial maximum, Palaeolithic populations of Europe and West Asia were isolated in climatic refugia. Genetic drift from population bottlenecks, lack of gene flow and selection pressures resulted in distinct genetic populations during this time [27]. When the ice sheet started retreating between 16,000-13,000 years ago [19], Europe began to be repopulated with people moving out of refugia and subsequently admixing. The Western European Hunter-Gatherers (WHG) spanned across western and southern Europe [35, 74, 49, 82] while Eastern European hunter-gatherers (EHG) occupied the west of present-day Russia, Finland, Latvia and down to the Pontic-Caspian steppe. These groups were related but EHG samples have a component of Ancestral North Eurasian (ANE) ancestry characterised by the Mal'ta specimen [39, 100]. Many Hunter-Gather samples found between the ranges of WHG and EHG appear as a mix of the two ancestries including Scandinavian hunter-gatherers (SHG) and Ukrainian specimens [62, 50]. The population of Mesolithic Europe therefore genetically appears as a cline of hunter-gatherer ancestry from Eastern to Western extremes.

Caucasus Hunter-Gatherers (CHG) were also present during the Mesolithic in the Near East and show genetic continuity with early Iranian farmers and affinity to present-day Armenians [39]. Although they are highly differentiated, the CHG/Iranians form a clade with the Anatolian farmers distinct from the WHG/EHG [49]. The timing of divergence between the four distinct groups appears to correspond to the onset of last glacial maximum [49, 61], during which time these population became isolated from each other in refugia.

At the start of the Neolithic around 9000BP, farmers from Anatolia moved into Europe, reaching Britain by 6000BP [12]. These incoming Anatolian farmers admixed to a small extent with the local WHG to form the Neolithic farmer population [74, 45]. Some sites produce genomes that look like Anatolian farmers with small amounts WHG ancestry [74, 75]. Other sites in the Balkans and Hungary exhibit individuals with little to no Anatolian farmer ancestry but are associated with a farming culture [50, 29]. In general, there is an increase in Hunter-Gatherer ancestry across Europe in Neolithic farmers during the Middle to Late Neolithic that likely involved persistent local WHG populations rather than an expansion from an isolated region [66]. This resurgence appears to be particularly large in present day France [15]. Overall, the movement of Anatolian farmers into Europe and the subsequent admixture with WHGs appears to have been slow with variable levels of admixture over time and geography, likely involving admixture between already admixed individuals [18]. I will later explore this heterogeneity in individual Neolithic farmer samples.

During the late Neolithic, steppe populations, characterised by the Yamnaya culture, appear as a mix of EHG and CHG/Iranian ancestries [39, 49, 61]. At the start of the Bronze Age after 5000 years ago, migrants with this "Yamnaya" ancestry moved west into Europe and had a profound impact on the genetic landscape [39], with steppe ancestry appearing first in individuals from central-eastern Europe [74] and spreading rapidly into central and northern Europe [4]. By 4,500 years ago, steppe ancestry appears in the British Isles and Ireland, brought by a population associated with the Corded Ware people and who replaced

approximately 90% of Britain's gene pool within a few hundred years [89, 16]. In central and northern Europe Y-DNA haplotypes common in Neolithic farmers almost disappear with the arrival of migrants and are replaced by steppe haplotypes [39, 74], suggesting that admixture was a male dominated [34], although this conclusion has been contested [63]. The impact of steppe migrants in Iberia was not so profound [73] where there appears to have been a more subtle genetic influx in contrast to the genetic turnovers in northern and central European populations. Modern populations of Iberia have the least Steppe ancestry and Scandinavian and Northern European populations the most.

It is believed that the steppe migration brought Indo-European languages to Europe and it is probable this is the case for central and northern Europe [39]. However, a few samples from Bronze Age Anatolia contain no steppe ancestry but show evidence of Indo-European languages, suggesting the ultimate source of Indo-European languages may not be in the steppe populations but earlier in Anatolia or the Caucasus. Likewise, contact of Anatolian populations with the greek Mycenaens and Minoans questioned whether it was the Yamnaya steppe migrants who were the vector for Indo-Europeans languages to the southeastern Europe [60]. A recent paper provides more clarity. Steppe ancestry was documented in ancient genomes from Bronze Age Balkans, Armenians and Myceneans, showing links from all ancient and present-day Indo-European language branches to Bronze Age steppe migrants. The origin of both proto-Anatolian languages and the sister family proto-Indo-European languages is in West Asian from which Caucasus populations expanded north, bringing ancestry north to both steppe populations and south to Anatolian neolithic populations which lead to the divergence of the two language families. The connection between Anatolians and steppe migrants is therefore through West Asian in a southern arc [59].

Present-day Europeans are often described as a three-way mix of WHG, Anatolian Farmer and steppe Yamnaya [62]. Figure 1.1 depicts the movement of these three ancestral populations into Europe. Given that the Yamnaya ancestry itself resolves into EHG and

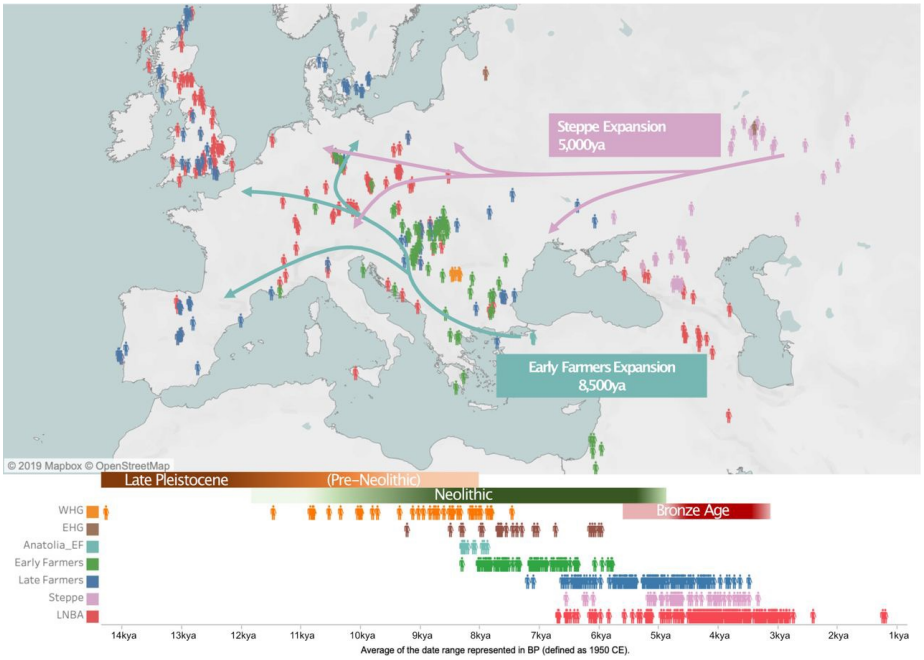


Fig. 1.1 Figure from a recent paper by Souilmi et al. describing the major movements of people into Europe and the geographic and temporal distribution of some ancient Eurasian samples. Each human symbol represents a sample and the colours indicate different groups classified into populations according to archaeological records. The green lines depict the generalised migration route of Anatolian farmers into Europe 8.5kya, where they admixed with Western Hunter-Gatherers to create the Early European Farmers (EF). Similarly, the pink arrows represent the generalised movement of the Yamnaya which resulted in admixture with Late European Farmers (LF) 5kya, giving rise to Bronze Age (LNBA) societies [115].

CHG/Iranian ancestry [49, 59], I propose a model of four ancestral streams that lead to present-day Europeans (Chapter 2).

1.7 Concept of local genetic ancestry and approaches to estimate it

As described in Section 1.1, when individuals can no longer mate randomly within a population and allele frequencies begin to diverge between groups, the population is said to be structured. Over time, gene flow between two subgroups within the larger population can cease, resulting in two new populations. Admixture events occur when there is migration

between two such divergent populations and they interbreed. Chromosomes of the resulting admixed individuals will originate in one of the two ancestral populations. With each generation there is recombination, meaning that over time the chromosomes in the admixed population will be a mosaic of chunks originating in the two ancestral populations. These chunks are inherited together on chromosomes creating admixture linkage disequilibrium (LD). As the number of generations since the admixture event increases, the length of these ancestral chunks will decrease. Local ancestry inference (LAI) is the process of decomposing admixed chromosomes into these ancestral chunks and assigning each chunk an ancestral label. Many tools are available that perform LAI due to its importance for understanding population structure, migration history and disease risks [72, 5] .

Early LAI methods such as STRUCTURE [24] and ANCESTRYMAP [92] use unlinked markers that are characteristic of populations, so called ancestry informative markers (AIMs). The hidden Markov model (HMM) structure employed by these methods models admixture LD as Markov chains, with the hidden states as the ancestry labels and the observations as genotype data. The memoryless nature of Markov chains means these methods assume that markers are independent and so do not model background LD, LD within the ancestral populations due to their population histories (drift). Additionally, parameters such as time since admixture and ancestry proportions have to be input, which are not always known.

With the decrease in cost of genotyping in recent years, the ability to deconvolve local ancestry with greater resolution using denser data became possible. Dense SNPs violate the assumption of independence between SNPs because of background LD, an independence which earlier HMM approaches assumed. Methods that leverage this denser data emerged such as LAMP [106] that uses a clustering algorithm within short overlapping windows and then combines the results by majority vote [33]. LAMP employs a pruning step which makes sure SNPs are unlinked in the ancestral populations. An extension method, WINPOP

[91] allows for one recombination within each window and dynamically alters window size making it more flexible to different times since admixture, especially those that are older.

The advantage of LAMP and WINPOP is that they can infer local ancestry without the need for a reference panel of data from ancestral populations. However, not explicitly modelling background LD within ancestral populations prevents the use of many informative SNPs that are linked. Haplotype frequencies vary more between populations than SNP frequencies meaning haplotypes can potentially provide greater ancestral resolution, especially when the admixing populations are more closely related. Many methods now both utilise denser SNP data and model background LD together with admixture LD by using extensions of the basic HMM approach. SABER [121] uses a Markov-hidden Markov model to account for background LD at consecutive markers. HAPMIX however, uses HMMs to model LD at two levels, allowing small-scale transitions between haplotypes within a reference population and large-scale transitions between reference populations [98, 111]. LAMP-LD was created as an extension to LAMP, using the same sliding window approach but with an added HMM to assign ancestry [7].

Recently machine learning has been utilised for local ancestry inference and has shown to be computational efficiency and accurate. Furthermore, the ever increasing availability of genomic data provides training data for supervised methods. RFMix uses a conditional random field, parameterised by Random Forests trained on reference panels to infer ancestry in windows across admixed chromosomes [69]. RFMix has been shown to be superior to previous methods, especially when the time since admixture is short and the ancestral populations are closely related. Subsequently, GNOMix has been shown to outperform RFMix and other LAI methods. GNOMix has a two module approach to inference: The base module is a classifier trained on an ancestral reference panel that outputs an initial ancestry estimate for segments in admixed chromosomes. The chromosomes are split into windows and each is assigned ancestry by a separately trained classifier. The classifier is modular

and multiple types were tested including Random Forests, Linear Regression and Support Vector Machines. The smoother module is then employed to refine ancestry estimates. The smoother uses the estimates in surrounding windows as input to predict the final ancestry assignment in the focal window [43]. In Chapter 2, I compare my method of local ancestry inference to that of GNOMix.

However, despite all these advances, most of the existing local ancestry methods require sequences from a discrete set of ancestral populations to form a reference panel. Choosing the reference panel is based on several factors such as availability of genomes, ADMIXTURE components [1] or questions that local ancestry inference is aiming to answer. In reality, haplotypes are formed by mutation in an ancestor and are inherited from generation to generation through history, likely through many populations and admixture events. An ancestral population may itself be an admixed population from an event earlier in history (Figure 1.2). In other words, populations are more like a braided river than a sequence of well-defined discrete population identities. So assignment of a haplotype in an admixed chromosome to that ancestral population does not inform us of ancestry further back in time.

Many LAI methods allow multiple admixture events [69, 43, 106, 121] and multiple ancestral populations in the reference panel. One solution might be to include multiple reference populations that represent populations involved in the deeper history of the focal admixed samples i.e use 'grand-ancestral' populations instead of the immediate ancestral populations in the reference panel. Yet this precludes the use of the ancestral populations to use the 'grand-ancestral' populations in their place, when in fact a haplotype is from both. Taking the example from Figure 1.2, population C is an ancestral population to population E. A haplotype in E can be assigned to both C and A; both are true at the same time. Existing LAI methods could include both C and A as discrete ancestral reference populations, producing confusing results or choose either C or A for the reference panel and only use half the available data. In other words, all LAI methods treat ancestral populations as discrete

entities with no genealogical relationships with each other, requiring them to effectively draw a line at some in the past time and take the populations existing at that time as 'pure' ancestral populations.

My aim is to develop a local ancestry inference method that takes time and genealogical relationships of ancestral populations into account. For this I redefine 'ancestry' as no longer a discrete population identity, but a complete path back in time through the population history. The path that a haplotype takes backwards in time from a focal individual is fully informative about its local ancestry: I am asking, in what populations has a haplotype been carried by inheritance through a structured population history? By determining this path, its relationship to all relevant historical and admixing populations is established.

I explain this in more detail using Figure 1.2 which shows a diagrammatic example of a population structure with two consecutive admixture events and two population split events. Population E is the focal population within which I want to perform LAI. I have representative samples from populations A, B, C, and D from within this structure which are termed 'ancestral populations'. Between these four ancestral populations, there are four paths that haplotypes could have taken backwards in time from population E, through multiple populations:

$$E \to C \to A$$
$$E \to D \to A$$
$$E \to C \to B$$
$$E \to D \to B$$

By using paths as local ancestry labels instead of discrete population identities, I am able to use all available data from all four ancestral populations and assign four different labels that convey meaningful information about the history of a haplotype.

My idea is to embed tree sequences inferred by RELATE into models such as Figure 1.2 of population structure. Population structure will alter the shape of coalescent trees as not

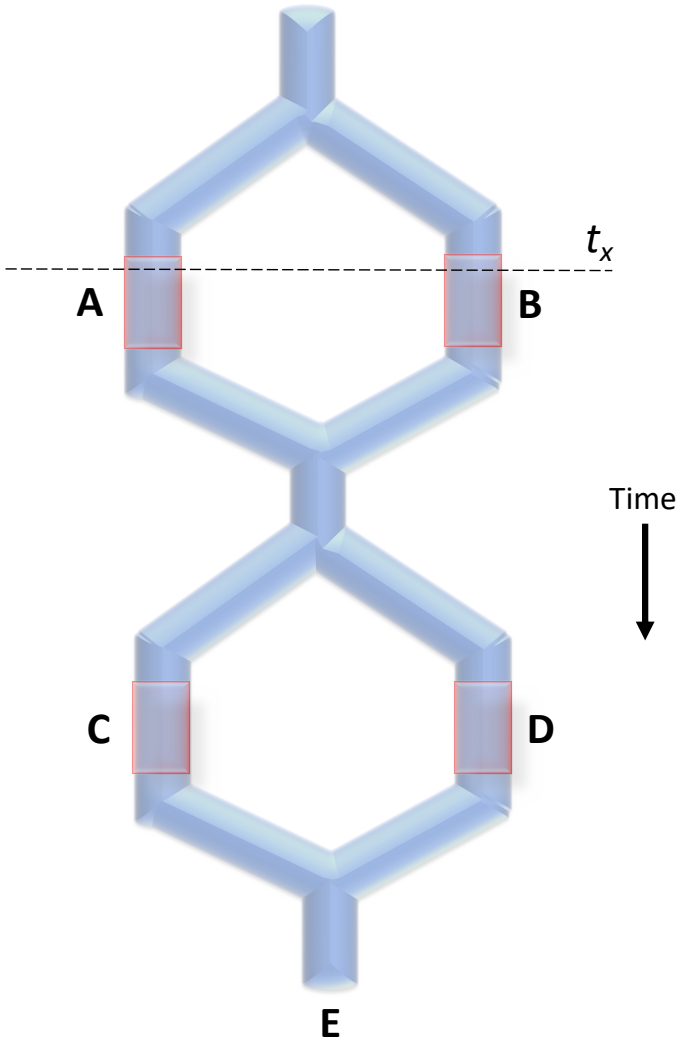


Fig. 1.2 Schematic of a structured meta-population that goes through two population split events and two admixture events. Populations marked A, B, C and D are 'ancestral populations' and population E is the admixed population whose local ancestry is of interest. The time slice t_x is marked where the population sizes of ancestral populations A and B ($N_A(t_x)$ and $N_B(t_x)$) can be calculated from the coalescences occurring between the lineages passing through those parts of the structure at that time.

all lineages have the same probability of coalescing with each other. By assuming therefore that the topologies and branch lengths of trees inferred by RELATE reflect the underlying population structure, I devised a way to determine the path from a feature I extract from trees covering each haplotype (Chapter 2). Lineages passing through one path can randomly mate

and so I can apply the basic coalescent to infer demographic parameters in different parts of the structured population at different points in time (Chapter 3). For example, at time t_x in Figure 1.2, I can find $N_A(t_x)$ the population size in ancestral population A using all the lineages that take paths $E \to C \to A$ and $E \to D \to A$.

Ancient samples within this framework are a huge advantage. The further back in time a population existed, the more difficult it is to find a proxy population that exists today and from whom genomic data is available. Ancient samples are likely to be less diverged from the true ancestral populations and so act as better representatives that existed before one or many admixture and split events. Moreover, with present day samples alone the number of lineages in trees falls going back in time as coalescences remove lineages. At deeper timescales the number of lineages becomes so low that there is little power for inference using the basic coalescent. Ancient samples increase power for inference deeper in time by injecting lineages into trees at these older times.

1.8 Thesis overview

In this thesis I propose a new method to paint individual chromosomes with their local genetic ancestries in terms of their recent evolutionary past. Instead of thinking of genetic ancestry as belonging to a certain static ancestral population, I redefine ancestry as a path going back in time through a structured population history as each haplotype is inherited through multiple populations that split and admix. The method involves building tree sequences with RELATE, then using a neural network to classify the ancestry path for each sample haplotype in a genomic region given the local tree. I train the neural network using data simulated from a model that represents the major ancestry flows contributing to modern European genomes over the last 50,000 years. I compare the results of testing my method on simulated data to results from GNOMix [43]. I also test how a range of simulated

demographic scenarios affects classifier performance and how robust the method is to model misspecification.

Using the local ancestry painting I describe techniques to 1. infer selection and changes in population size in a structured population based on the assignment of coalescences in the trees to paths and 2. infer the time since admixture using the rate of switching between ancestral painted segments.

I apply the method to the MesoNeo genomes, a large dataset of newly published and previously published ancient genomes that have been shotgun sequenced and imputed. From my results I draw conclusions about the history of Europeans.

I divide this thesis into the following chapters. Chapter 2 describes the construction of a model of European population structure based on the MesoNeo dataset. Furthermore, I detail the method of inferring path local ancestry from tree sequences. Chapter 3 covers how to use the local ancestry painting to estimate time since admixture, followed by Chapter 4 which describes the application of this method to painted MesoNeo genomes. Chapter 5 explores methods to calculate effective population size within ancestral populations and infer path-specific selection, with some results from the MesoNeo genomes. Chapter 6 is a discussion of my work, its applications and future uses.

Chapter 2

Building a model of European population structure

Contents

2.1 Population genetic analysis of MesoNeo genomes

My research was funded from a Wellcome Trust collaborative grant within which the aim was to sequence a large dataset of five to ten thousand ancient whole genomes from people that lived mostly in the last 10,000 years from Eurasia (Figure 2.1). Ancient DNA was extracted from the petrous bone and dental cementum of 317 specimens and shotgun sequenced. After merging with >1300 previously published shotgun sequenced genomes, the genomes were imputed using a new imputation method GLIMPSE [103], which is optimised for low coverage samples. A 1000 Genomes reference panel was used with high coverage ancient samples included in the panel. Imputation was tested with 42 high coverage ancient genomes which were downsampled, imputed and compared to their full coverage version. Genomes down to 0.1X coverage displayed good imputation accuracy [3]. The final dataset comprises 1,490 genomes, once filtered for coverage (>0.1X), low imputation quality and close relatives [3], at 3.7 million SNPs filtered at >0.5 imputation INFO score.

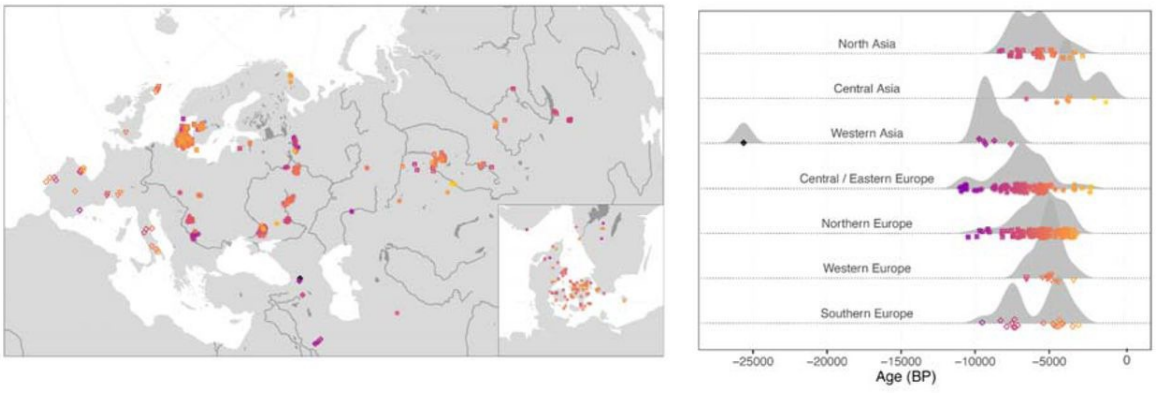


Fig. 2.1 Geographic and temporal distribution of the 317 newly reported MesoNeo genomes [3]. Age of the sample is indicated by colour and geographic region by symbol.

The dataset covers a critical time period during which radical changes in lifestyle occurred with the transition from a hunter gatherer existence to farming and urbanisation. It has been suggested that many diseases, especially chronic, autoimmune and inflammatory diseases, are aggravated by a mismatch in the lifestyle of our ancestors and our current lifestyle [67, 87, 26]; variants that were beneficial and selected for in the past may have a deleterious effect in a modern environment and be associated with disease, and reciprocally previously neutral variants may now be adaptive. Understanding what parts of chromosomes are inherited from which ancestral groups and the evolutionary processes that acted during these past transition may help inform our understanding of disease susceptibility and how we treat and prevent diseases.

For my initial investigation I subset the full MesoNeo dataset to those that had archaeological locations in Europe and West Asia and performed Principal Component Analysis to assess the population structure and assign samples to groups. The imputed genomes were filtered to SNPs in the Affymetrix Human Origins Panel [94]. I used EIGENSOFT smartpca [97] with the outlier removal option disabled to perform PCA directly on the imputed ancient samples together with 1000 Genomes GBR samples, with no projection. The samples were plotted by their first two principal components (Figure 2.2).

Broadly, principal component 1 (PC1) separates north-south geographically and PC2 separates east-west. Western hunter-gatherers (WHG) and Eastern hunter-gatherers (EHG) are the extreme ends of a cline along PC2 across the right side of the plot, suggesting they are extensions of the same continuous population that spans a large geographical range across Europe. Scandinavian hunter gatherers lie in between these two groups.

The Anatolian farmers (Ana) form a cluster in the left centre of the plot defining genetic ancestry from the southeast. This cluster consists of slightly earlier samples from Turkey with some later samples from Europe that have little to no WHG admixture.

Along PC1, the Neolithic farmers (Neo) lie in a smear between the Ana and WHG, characterising the expansion of Ana out of Anatolia and into Europe, admixing with WHG at varying levels.

The Caucasus hunter-gatherers (CHG) including early Iranian farmers, form a distinctive cluster in the top left corner of the plot, separate from the other hunter-gatherer groups and the Ana, despite the Southern Caucasus and Anatolia being geographically close.

The Bronze Age Anatolian group (BAA), above the Ana cluster, represent a later Anatolian population with genetic influence from the CHG.

The Yamnaya (Yam) group from the Pontic Steppe, consistent with their formation by admixture between the EHG and CHG, lie between these groups in PCA space.

Finally, European Bronze Age samples sit in the centre of the plot between the Yam and Neo, a result of the Yamnaya movement west into Europe and admixture with the Neolithic farmers. The 1000 Genomes GBR samples fall within the Bronze Age cluster.

To take forward into my next analyses, I further subset the ancient genomes from this subset of the MesoNeo dataset to samples that were tightly clustered in the groups relevant to the genetic history of Europeans. The subset was chosen using the PCA analysis described above, where individuals that fall in between the ancient groups in the PCA were removed so that only samples diagnostic of each population were kept. This totalled 476 diploids (952 haploids) including 91 GBR 1000 Genomes samples. The number of samples in each population, shown in Table 1, was used in the following sections when building a model of European population structure.

2.2 Demography construction and simulation in msprime

Genome sequences are a product of their evolutionary history, including large-scale migrations, admixture and population divergence. Using the extensive amount of previous

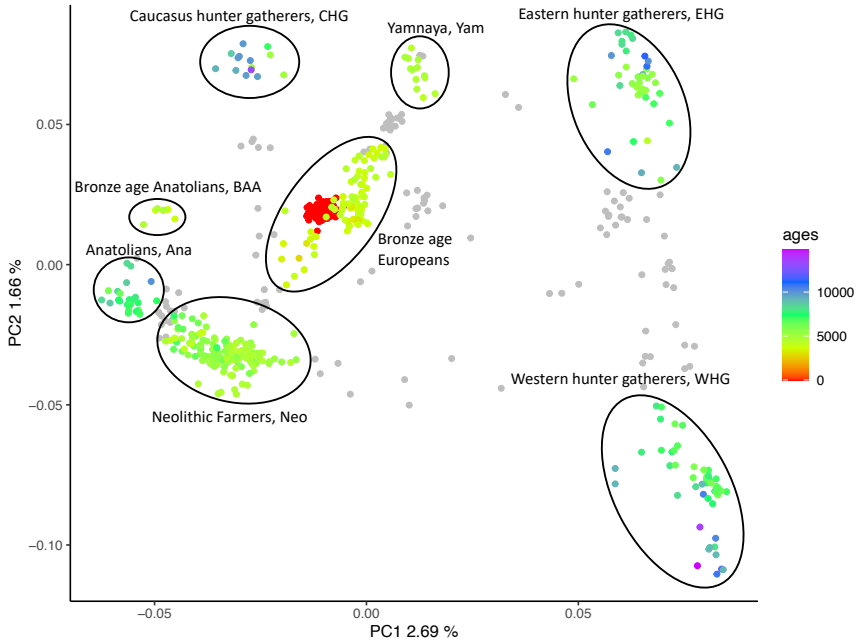


Fig. 2.2 Subset of MesoNeo genomes, plotted by their first two principal components. Samples that are diagnostic of each ancient European group are coloured on a continuous scale by their radiocarbon age in years ago. The samples that fall in between the circled, coloured samples and were removed from further analysis, are coloured in grey. The percentage of variance explained by each component is displayed in the axes labels.

Table 2.1 Table of the number of haploid genomes from each ancient European population in the subset of the MesoNeo dataset chosen from PCA analysis. Total of 952 haploid sequences.

Population	Number of haploid samples
GBR	182
Bronze Age	162
Western hunter gatherers	102
Eastern hunter gatherers	86
Anatolian farmers	50
Neolithic farmers	302
Yamnaya	26
Caucasus hunter gatherers	28
Bronze Age Anatolian	14

work elucidating the genetic history of Europeans, as described in section 1.6 and the PCA (Figure 2.2), I put together a standardised model of European population structure.

Figure 2.3 shows a schematic of this demographic model that describes the population structure in Europe during the last 50k years. Shortly after the expansion of anatomically modern humans into Eurasia, I model a population split ∼45kya (1500 generations ago) between the Northern Europeans (NE), who continued travelling northwest into Europe, and West Asians (WA) who stayed more locally in the Levant and South Caucasus area. The WA population then splits to form the CHG/Iranians and the Ana ∼24kya (800 generations ago). Within the NE, ∼18kya (600 generations ago), the WHGs and EHGs diverge. At that point the four separate populations that make up present day European ancestry are distinct in the model. The model subsequently describes how admixture between these populations leads to the modern European gene pool. Firstly, the formation of the Neolithic farmers (Neo) from admixture between WHGs and the Ana ∼7.8kya (259 generations ago), secondly the formation of Yam from admixture of EHGs and CHGs during the early Bronze Age ∼5.3kya (177 generations ago) and finally the formation of the Bronze Age gene pool ∼4.9kya (166 generations ago) as an equal mix between Yam and Neo. The formation of the BAA from admixture of CHGs with Ana happens ∼5.1kya (170 generations ago).

Six different paths that haplotypes can take from any sample in the subset are shown in different colours in Figure 2.3. Path 1 = red, starts from the present day Europeans, going back through Neolithic farmers, Anatolian farmers, West Asians to the root. Path 2 = purple, starts at present day Europeans, going back through the Yamnaya, Caucasus hunter gatherers, then West Asians to the root. Path 3 = black, starts with present day Europeans, going back through Neolithic farmers to Western hunter gatherers and then through Northern Europeans to the root. Path 4 = orange, starts at present day Europeans, going back through the Yamnaya to Eastern hunter gatherers and then through Northern Europeans to the root. Paths 5 = blue, starting in the Bronze Age Anatolians and joins path 1 part. Path 6 = cyan, starts in Bronze

Age Anatolians and joins path 2. When paths overlap, lineages from all overlapping paths can coalesce.

In order to better understand the impact of population structure on local ancestry inference and explore techniques to achieve this inference in real genomes, I simulated genotype and tree sequence data from this model. I constructed my standardised model of European population history in msprime [52], a coalescent simulator. The population sizes and admixture fractions are shown in the schematic in Figure 2.3 (See section 2.3 for parameter choice). Samples were taken from populations and times to match each real sample (Table 2.1), using their radiocarbon or context dates, to mimic my MesoNeo subset.

Upon simulation, data for sample individuals is produced, in the form of both VCF files and tree sequence files, that is consistent with the constructed demographic history. Typically I simulate chromosomes of length 200 Mbp with a recombination rate of 1e-8 per bp per generation and neutral mutations dropped onto the branches at a rate of 1.25e-8 per bp per generation. It is also possible to simulate using a real human chromosome recombination map in which case the sequence length is given by the file and a variable recombination rate across the sequence in applied. To manipulate and examine trees in tree sequence files I use the tskit python API (https://tskit.dev/tskit/docs/stable/introduction.html).

My model of European population structure, constructed in msprime, has been submitted to the stdpopsim catalogue (https://stdpopsim.readthedocs.io/en/latest/introduction.html) and is publicly available for users to simulate data from.

2.3 Comparison of simulated and real variant data

2.3.1 PCA Analysis

I performed Principal Component Analysis on the simulated genotype data to compare the structure to that of the real data. I used the same procedure as in section 2.2 on nine

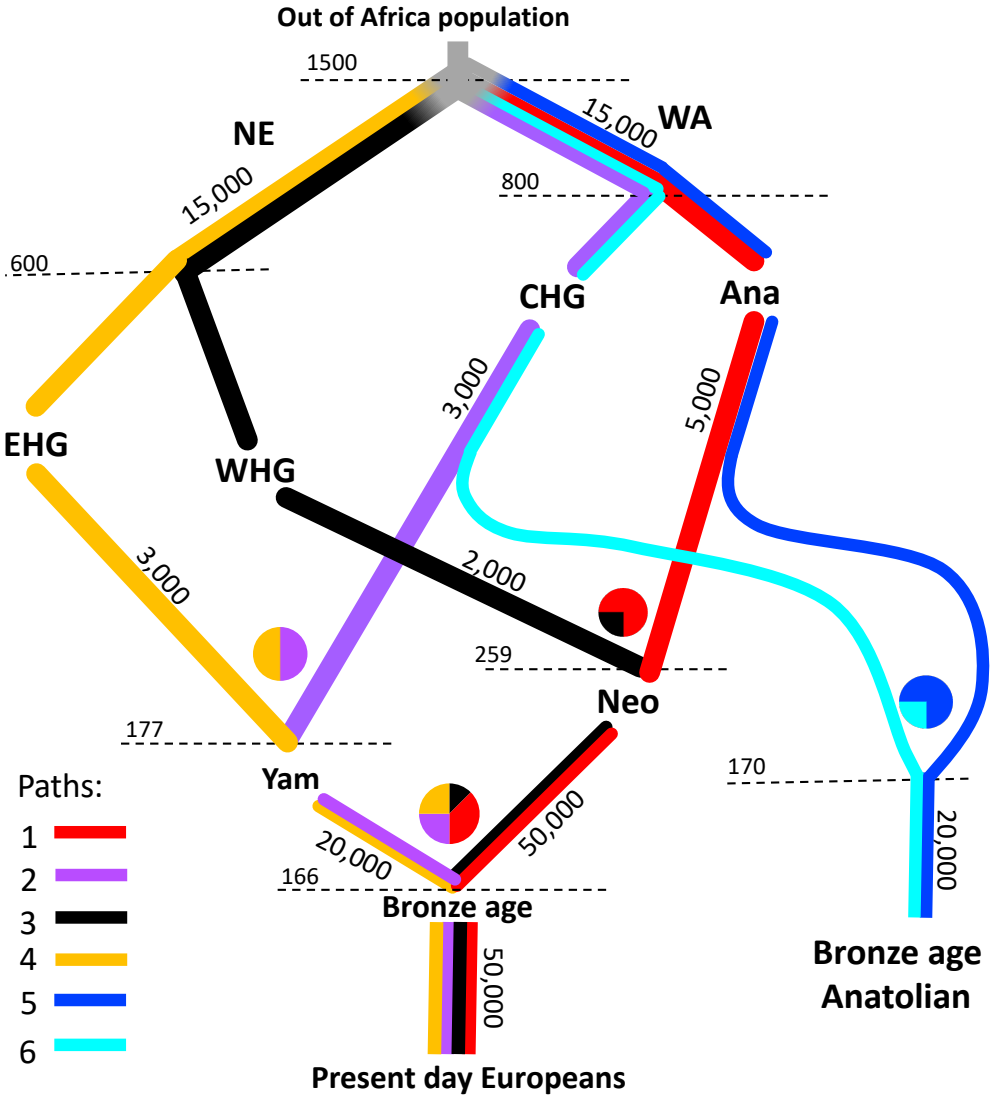


Fig. 2.3 Schematic diagram of the model of European population structure, implemented in msprime. Samples are taken from points throughout the model corresponding to the radiocarbon ages of the real MesoNeo samples in generation ago. Population labels are abbreviated. Bronze Age = Bronze Age European population, Yam = Yamnaya steppe, Neo = Neolithic farmers, WHG = Western hunter gatherers, CHG = Caucasus hunter gatherers, EHG = Eastern hunter gatherers, Ana = Anatolian farmers, NE = Ancient Northern Europeans, WA = Ancient West Asians. Effective population sizes are shown along edges and admixture fractions are displayed as pie charts. The different paths are coloured. The timing of population splits and admixture events are shown at the dotted lines in units of generations ago.

simulated chromosomes, filtered for minor allele frequency of 5%. In Figure 2.4 the cluster for each ancient population falls in the same vicinity of PCA space as the real MesoNeo samples and the variation explained by PC1 and PC2 is comparable. Overall, the similarity of the PCA suggests that a lot of the underlying structure that determines how these groups relate to each other is captured by the demographic model.

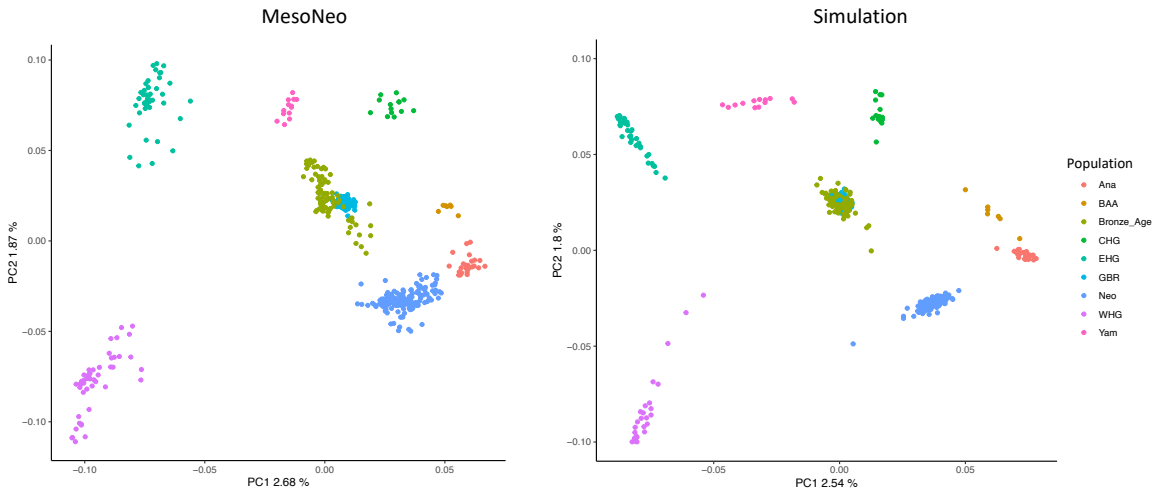


Fig. 2.4 Principal Component Analysis on the simulated data compared to the real data, plotted by the first two principal components.

2.3.2 F_{st} Statistics

Weir and Cockerham weighted F_{st} statistics were calculated over all chromosomes using VCFtools between all pairwise populations in the MesoNeo subset dataset. The same was done for nine simulated sequences of length 200Mbp. Higher F_{st} values suggest greater divergence between a pair of populations so lower values are expected between populations that diverged more recently from each other and between ancestral and admixed populations.

The results shown in Figure 2.5 display the F_{st} pairwise values in the simulated data plotted against those in the real data. The correlation of F_{st} values between the real and simulated data is appreciable at 0.96. This means the demographic model produces data with relative population divergences that are very similar to that in the real data, suggesting the model represents the real population structure well.

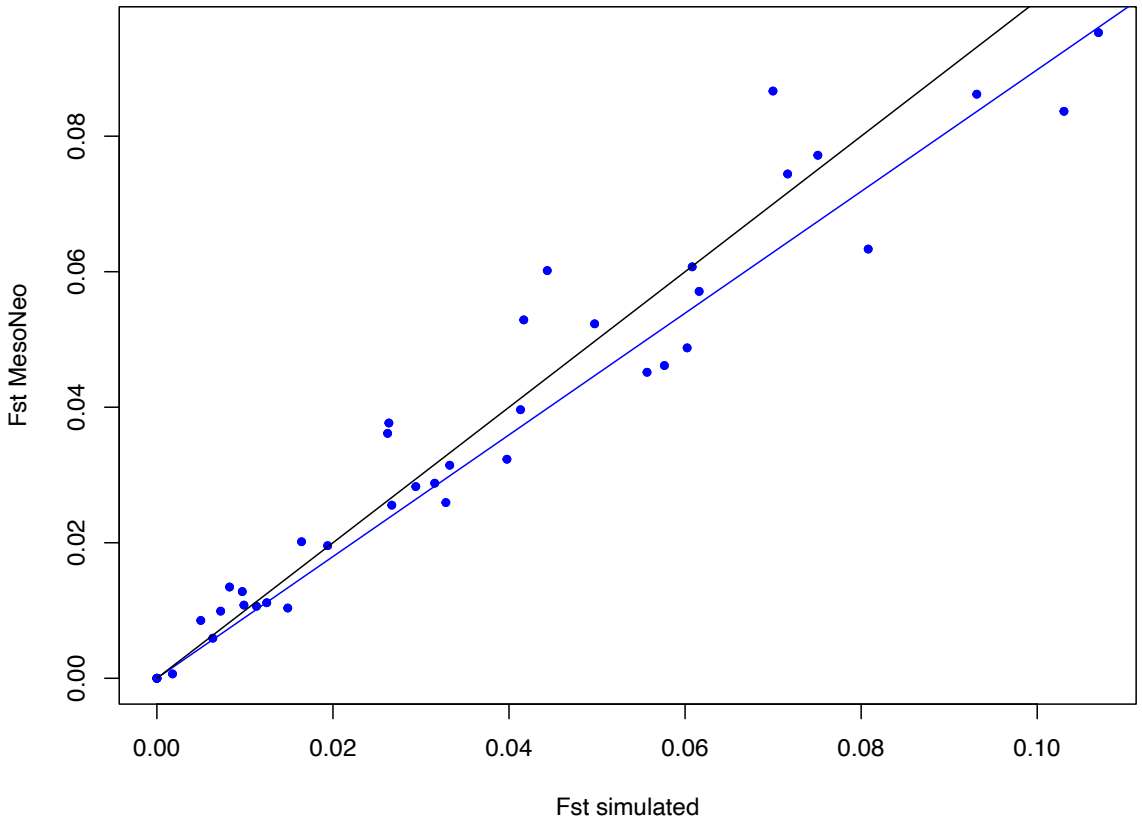


Fig. 2.5 Plot of the simulated F_{st} values against the real MesoNeo F_{st} pairwise values. The solid black line represents the 1:1 mapping. The solid blue line is a linear regression on the points, showing a 0.89 coefficient.

The relationships of these ancient groups and present day Europeans to each other and the split/admixture times are well established and needed no adjustments in the model. However, to produce PCA and F_st results in simulated that matched the real data as shown, the population sizes needed some tuning. Both the F_{st} analysis and PCA final results suggest

that, while the model is undoubtedly not entirely correct, it produces simulated data that looks close enough to the true data, by these measurements, that I can continue with my analysis on the basis that it forms a good enough approximation.

2.4 A method for estimating local genetic ancestry

From the perspective of the local genealogical trees, many existing programs aim to describe the local ancestry of individual haplotypes by identifying their closest relative haplotype(s) from a set of reference sequences. The population identity of the closest relative(s) indicates the local ancestry of the focal haplotype. In other words, these programs only examine samples leaves under the first coalescence node above the focal haplotype.

However, with ancient samples and admixture events, the first coalescence alone is insufficient to understand the full ancestry of a given haplotype (Section 1.7). Firstly, the first coalescence event may occur at a time younger than the age of some sampled groups, in which case the older sampled individuals could not be found as the closest relatives and the full local ancestry of haplotypes is not correctly established. Secondly, with some sampled populations formed via admixture of other sampled populations, the closest relatives to a haplotype may by chance be from an admixed population even if the age of first coalescence is old enough to capture older sampled groups.

In order to capture the full ancestral history of a sample at a site, I redefine ancestry as a path as described in Section 1.7. My method aims to infer the path that haplotype chunks, that are covered by a single tree in a tree sequence, have taken through a population history. To determine this 'path' ancestry, I leverage information in nodes above the first coalescence.

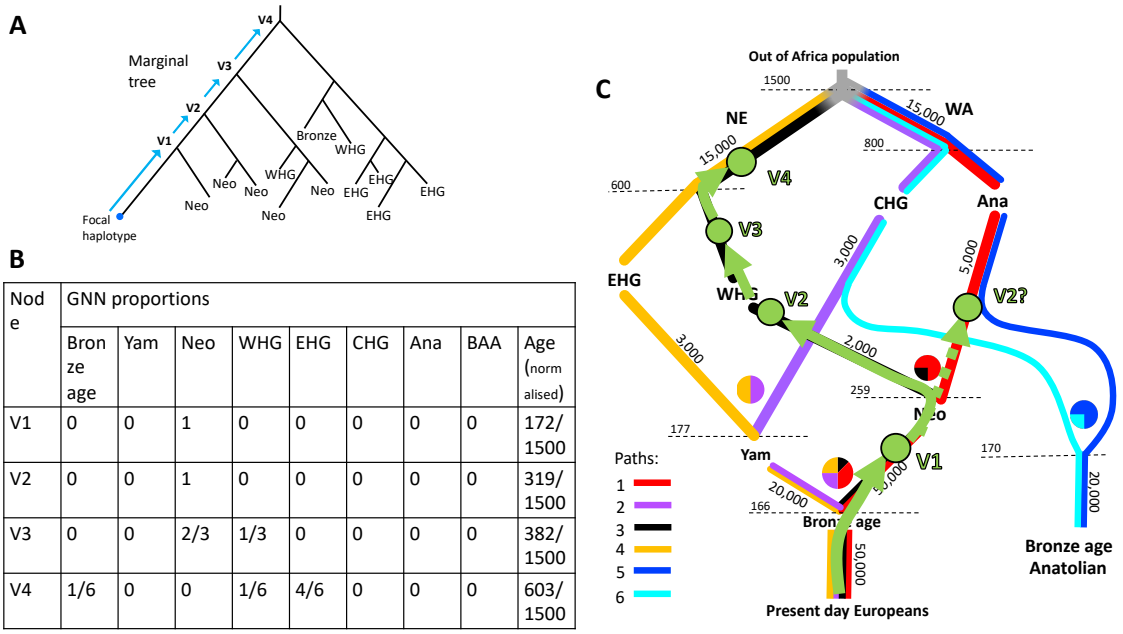
I first simulate variant and tree sequence data from a demographic model of the history of the populations under analysis as constructed in the previous Section. Next I infer RELATE tree sequences from the simulated genotype data. I then extract a feature vector of the trees, that is related to tskit's Genealogical Nearest Neighbours [53], and train a neural network to

predict the path from this feature vector. The true path label for training is known from the corresponding simulated tree sequences. The neural network can then be applied to RELATE tree sequences inferred from the real data to classify a path for every sample in every tree.

2.4.1 Genealogical Nearest Neighbour distributions

As an illustrative example, using Figure 2.6, I take a European present-day haplotype for which I want to infer the path going back through the model of European population history. I extract and analyse its marginal tree from the tree sequence describing its genealogical relationship to all other sampled haplotypes. From that haplotype I traverse up the tree, jumping to successive parent nodes towards the root i.e parent, grandparent, great-grandparent etc. Ideally, I would like to identify what population each of these internal nodes, or ancestors, is from. The population label of internal nodes is recorded in simulated tree sequences but in RELATE inferred tree sequences this is not possible.

I therefore adapted the concept of Genealogical Nearest Neighbours (GNNs) from Kelleher et al. [53]. This involves recording the proportion of each ancestral group that makes up the sample leaves below each node, not including leaves seen at previously analysed nodes further down the tree. I refer to the distribution of leaf ancestry proportions at a node as GNNx , where x is the xth node examined towards the root. The ordered collection of all x GNN distributions of all x nodes examined during a tree traversal reflects the path that the focal haplotype has taken to the root (Figure 2.6). The key is to use coalescences above the first and to consider the GNNs together, not independently. Figure 2.6 demonstrates how by looking at these ordered GNNs one can determine the path. The method also works when traversing up to the root and finding the path for ancient sample haplotypes. Therefore, I can assign local ancestry to Bronze Age, Yam, Neo and BAA samples, while WHG, EHG, CHG and Ana sample haplotypes have paths given by their population identity alone.



Node	GNN proportions								
	Bronze age	Yam	Neo	WHG	EHG	CHG	Ana	BAA	Age (normalised)
V1	0	0	1	0	0	0	0	0	172/1500
V2	0	0	1	0	0	0	0	0	319/1500
V3	0	0	2/3	1/3	0	0	0	0	382/1500
V4	1/6	0	0	1/6	4/6	0	0	0	603/1500

Fig. 2.6 An overview of GNN extraction. **A** shows an example marginal tree that relates the focal haplotype to all other haplotypes in the dataset. Samples are shown by their population labels. **B** shows the GNN matrix determined from the marginal tree, traversing up four nodes towards the root V1-V4. **C** shows how the path can be determined through the model of population history given the GNN matrix by mapping the nodes to the paths.

I need to be able to assign paths to millions of sample haplotypes so this cannot be a manual process. Instead, I implemented a supervised machine learning method, specifically a convolutional neural network (CNN) using the Python Keras package with a TensorFlow backend. The network was trained using a categorical cross-entropy function, Adam optimisation and a batch size of 30. The final layer is a softmax transformer. A class is determined as that with the highest softmax value.

The input to the network is the set of GNN distributions for the first five informative nodes traversed towards the root, configured as a 5 x 9 matrix, one row per node examined. Columns 1-8 contain the proportions, between 0 and 1, of leaves belonging to each of the 8 ancient sampled groups (Bronze Age, Bronze Age Anatolian, Yamnaya, Neolithic, WHG, EHG, Anatolian, CHG) and column 9 contains the normalised age. Informative nodes are those that have at least one leaf from the set of ancient sampled groups. If the root of the tree is reached in less than five informative nodes, then the remaining rows of the matrix are filled with -15 as 'padding'. The output of the network is a numerical label 1-6 characterising the path (Figure 2.3).

2.4.2 Training and testing a neural network

A large amount of training data and corresponding true labels can be generated by simulation of tree sequences from the model. I simulated three 200Mbp tree sequences, using different random seeds. RELATE tree sequences were inferred from the corresponding simulated VCF files. Default parameters for RELATE were used; 1.25e-8 mutation rate and starting population size estimate of haplotypes of 30,000. Sample ages were passed to RELATE which are used to constrain topologies and coalescence times.

RELATE underestimates the number of trees by over a factor of ten (Table 2.2). This is likely because of the RELATE's relaxation of 'uniquely mapping' when placing mutations on branches where not all descendants of a branch must be derived for the mutation for RELATE

Table 2.2 The number of trees in the true tree sequences compared to the number of trees inferred by RELATE from the corresponding simulated VCF files.

Number of trees	
Simulated	RELATE inferred
991409	90044
987636	89468
989834	89867
991693	90139
992928	90018
987797	89177
991613	89679
992180	90356
988802	90002

to estimate a new tree topology. The purpose of the relaxation is so that tree building is robust to sequencing errors, which in the case of simulated data, is not applicable. As an observation, when this relaxation is abolished the number of trees estimated by RELATE doubles which is still an underestimation by over a factor of 5.

Accuracy and precision

I extracted 10,000 training pairs of GNNs and true labels from each of the five admixed sample populations (GBR, Bronze Age, Yamnaya, Neolithic farmers and Bronze Age anatolians), 50,000 pairs in total. GNNs were taken from the trees covering evenly spaced sites across three tree sequences output from RELATE, avoiding most of the correlation between trees. True path labels were taken from the trees covering the same sites in the corresponding simulated tree sequences. I trained the classifier described above to predict the path labels from the GNNs.

For testing, I simulated five more tree sequences and tested the classifier on 10,000 GNNs from each. Across the tree sequences I obtained a mean accuracy of 93.12% with a standard deviation of 0.29%. To test the precision in each class, I pooled the testing GNNs from all five sequences and applied the classifier. Figure 2.7 is a confusion matrix comparing the classed

labels to the true labels for all testing data from all populations and classes, normalised by the sum of the rows to show the precision of the classifier in each class i.e of the labels assigned a class, how many are true positives. The network displayed a high precision for every class. Path 5 and 6 have the lowest precision, being confused for each other or path 1 and 2 respectively. This is understandable as path 5 and 6 lead from the BAA and join paths 1 and 2. Paths 5 and 6 likely have overlapping GNN features with each other and paths 1 and 2.

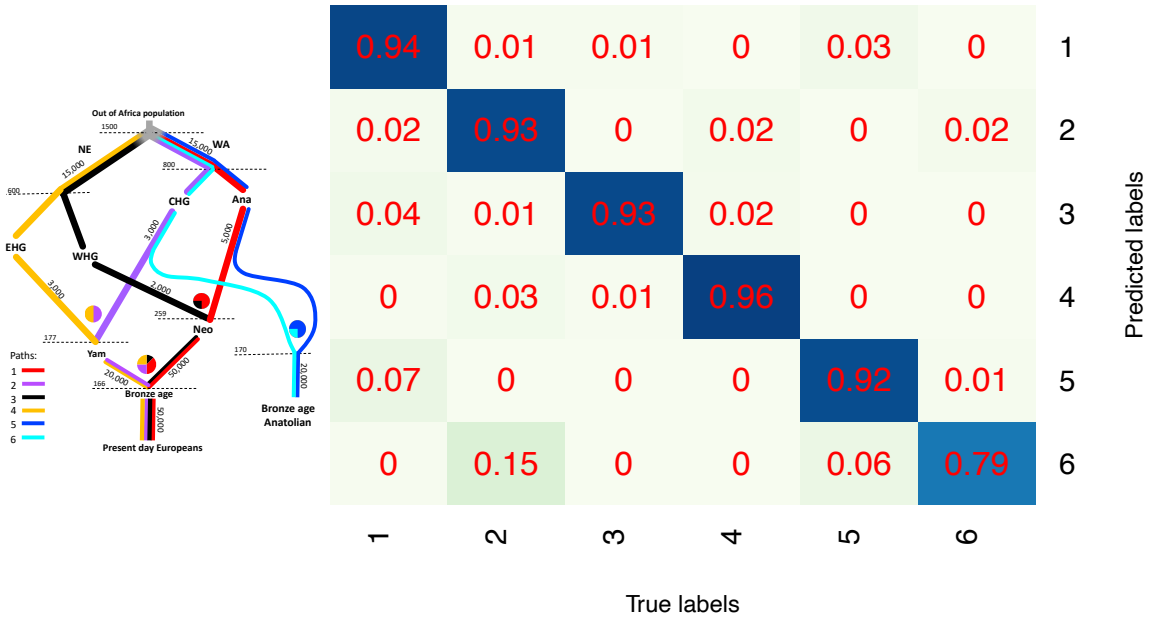


Fig. 2.7 Confusion matrix normalised by the sum of row to show the precision values when testing a classifier trained on the model of European population structure

Separating the confusion matrix into one for each population, Figure 2.8, shows that the GBR obtains the lowest accuracy. GBR samples also have the youngest sampling time. More generations since the admixture events means more recombination events to break down admixture LD, resulting in shorter tracts of ancestry. RELATE uses flanking SNPs to calculate the distance matrix and allows some relaxation on the mapping of SNPs when constructing new trees. These properties make it more difficult to determine the switch of ancestry at the edges of tracts as there is some inertia in the changing of tree topologies

compared to the true tree sequences. Shorter tracts produces more edges and therefore less accuracy in the GNN assignment. However, high accuracy is still maintained for all populations, even the GBR.

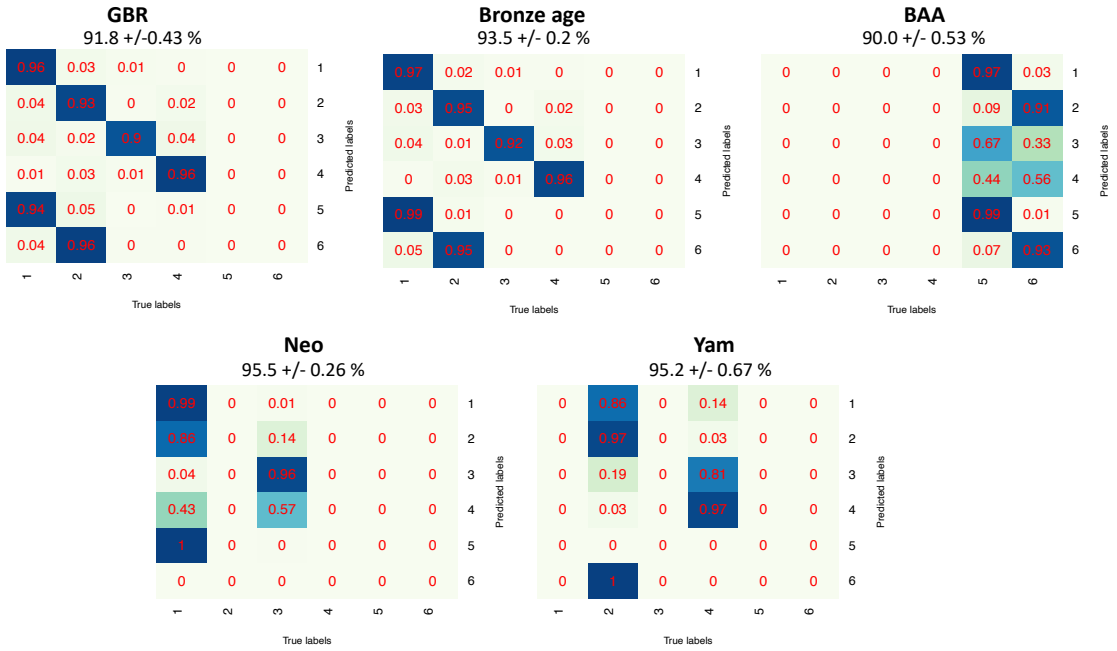


Fig. 2.8 Confusion matrices per population, normalised by the sum of row to show the precision values, when testing a classifier trained on the model of European population structure.

To visually see the performance of the classifier, Figure 2.9 shows a classified painted chromosome alongside the true simulated chromosome. Noise appears as short tracts of correlated trees, within larger chunks of ancestry covering many sites in admixture LD.

Comparison to an advanced LAI tool

To see how my method compares to advanced existing methods, I tested my method against GNOMix, a recent LAI tool that has been shown to outperform previous methods on whole genome data [43]. GNOMix, like other LAI methods, views each reference population as a discrete ancestral population with no awareness of relationship to other reference populations. Therefore, for the GNOMix reference panel I use samples from the

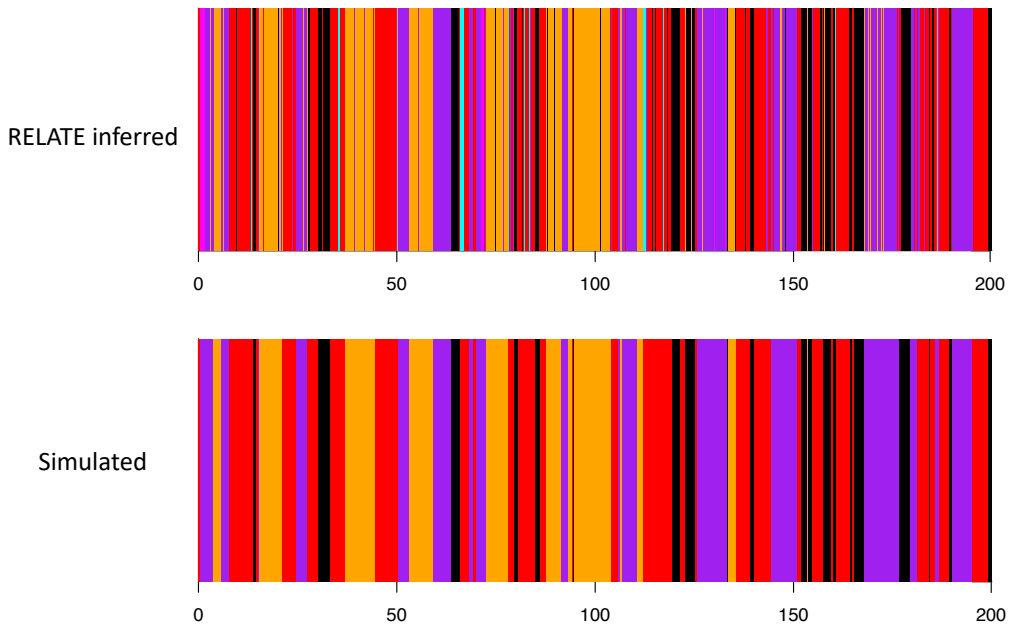


Fig. 2.9 Example of a simulated painted haploid chromosome from a Bronze Age individual and the corresponding RELATE inferred chromosome, painted by classifiation.

four ‘path’ populations (EHG, WHG, CHG and Ana) to represent paths 1-4 from the model. These are the populations that lie on one path only (Figure 2.3). For simplicity I only test the four admixed populations, GBR, Bronze Age, Neo and Yam, as query sequences to GNOMix and do not test the BAA (although this could be done separately using CHG and Ana as ancestral populations).

Inference with GNOMix was done using the default logistic regression base and xgboost smoother modules. All other parameters were default including a window size of 0.2cM for Bronze Age, Neo and Yamnaya samples. Because the GBR are 166 generations since admixture I decreased the window size to 0.02cM for them to account for the smaller admixture tracts.

For testing, I simulated five 200Mbp length sequences from the model of European population structure. I extracted ancestry predictions, produced by my method and by GNOMix, from evenly spaced sites across the five sequences. The mean and standard deviation for each admixed population, across the five sequences, obtained by each method

Table 2.3 Table of mean accuracy values of LAI for four admixed populations tested with either GNOMix or ancestral paths method. The mean was taken across five testing simulated sequences by taking evenly space sites from each population.

Population	GNOMix accuracy	Ancestral Paths accuracy	T-test p-value
GBR	74.7	91.8	5.963e-12
Bronze Age	87.1	93.5	1.171e-07
Neolithic farmers	94.4	95.5	1.491e-05
Yamnaya	98.6	95.2	1.593e-05

were used to perform two-sample T-tests to test for a significant difference in accuracy (Table 2.3).

For the GBR, Bronze and Neo the ancestral paths method is significantly more accurate. GNOMix was significantly more accurate for the Yam population. While GNOMix has high accuracies for populations closer to the time of admixture, classification of the present-day GBR samples is much less accurate than my method. Despite reducing the window size for GBR classification, it appears GNOMix struggles with smaller sized ancestry tracts. My method displays high accuracy for all populations, demonstrating its versatility compared to GNOMix.

Excluding Bronze Age Anatolians

My model of European population structure includes the Bronze Age Anatolians (BAA) from which paths 5 and 6 lead. Given these paths are not directly relevant to the history of present day Europeans, I tested how the accuracy of classification is altered when the BAA are removed from the GNN distributions and paths 5 and 6 are removed as labels. This leaves a four path model and a reduced GNN matrix size over which to train a neural network. I tested the accuracy of this classifier in each of the 4 path classes, averaged over 5 testing tree sequences, in the GBR population.

The overall accuracy across the four paths in the model with no BAA is 94.3% +/- 0.27%. This is significantly greater than the overall accuracy in the model containing BAA when

taken over all six paths in that model (p-value = 1.473e-05). However when averaging over just the four non-BAA paths in the model containing BAA I obtain an accuracy of 94.6% + 0.46%, which is not significantly different to the classifier trained on the model excluding BAA (p-value = 0.0918).

Fig. 2.10 Confusion matrices showing the precision in each path tested over GBR samples in 5 testing tree sequences. The classifier trained on the model including BAA can predict one of six paths, while the classifier trained on the model excluding BAA can predict one of four paths.

This result can be seen in Figure 2.10 where the precision in each of the four non-baa paths in the model including BAA is very similar to those in the model excluding BAA.

This result demonstrates that accessory paths such as those leading from the BAA in my model of European population structure can be added with no detriment to precision of other paths. Although the overall accuracy across all paths is lower when the BAA are included, the accuracy is still over 90%.

2.4.3 Conditions affecting classifier performance

The "path ancestry" inference approach outlined above is applicable to many populations of humans and other species. To help decide what types of populations and time resolutions

this method would be appropriate for I assessed its performance under a variety of demographic scenarios. I simulated data from a range of demography models, systematically varying features and parameters and I measured the overall classifier accuracy and precision in each class when tested on simulated data.

Unless otherwise specified, I simulated tree sequences of 20 Mbp in length with 1.25e-8 mutation rate and constant recombination rate of 1e-8. All admixture events involve equal proportions contributed by each participating population. The ordering of population split and admixture events is determined randomly by sampling from the active populations. The timing of split events is three generations after the previous split event while the timing of admixture events is fifteen generations following the previous admixture event. All populations have an effective population size of 10,000. Twenty five diploid samples were taken from all the 'path' populations that are present after all split events have occurred and also from all admixed populations. The sampling time for each population was sampled uniformly from within the time each population was active and all samples per population are taken at the same time. An example is shown in Figure 2.11.

Training GNNs and label pairs were extracted from all admixed populations. To gather the same number of pairs for each path in each admixed population from sufficiently spaced sites, I simulated five training tree sequences and five testing sequences. The largest source of variation is across tree sequences so testing was performed on each sequence separately and a mean and standard deviation across sequences was calculated. Two sample t-tests were then performed to assess if there was a significant difference in classifier accuracy when a parameter was changed. I then pooled the testing GNNs and applied the classifiers to produce confusion matrices and calculate precision values in each class.

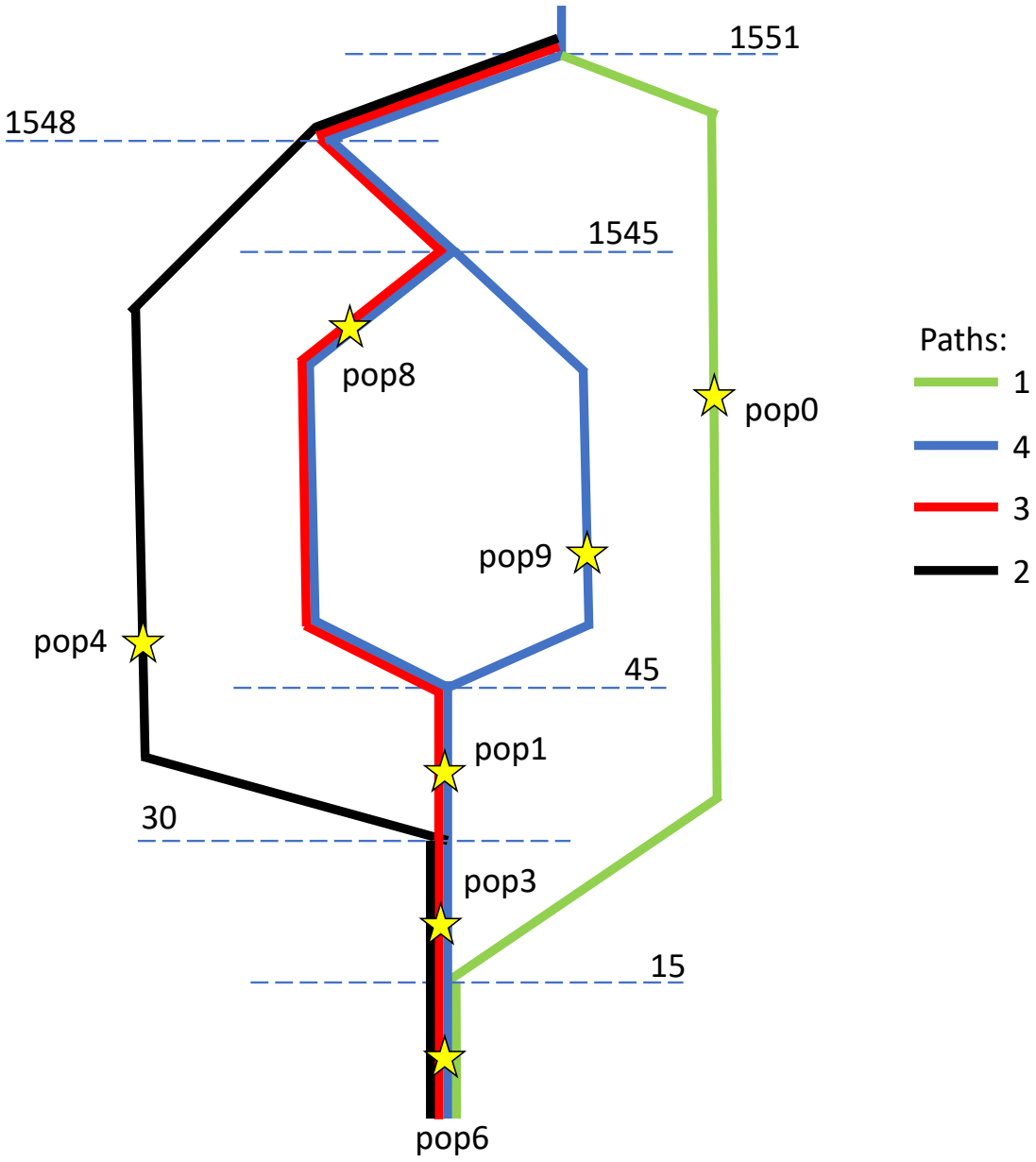


Fig. 2.11 An example of a four path demography that was simulated for testing. Stars indicate approximately where in time 25 diploid samples are taken. The populations and paths are arbitrary numbered. Population split and admixture times are shown on the dotted lines in units of generations ago. This example has a 'path' divergence time of 1500 generations, the time period between the last population split and the first admixture event.

Inference with varying path number and divergence time

I tested how the number of paths in the model and how much differentiation between the 'path' populations affected the classifier's ability. This was to explore 1. how complex the demographic history could be for the classifier to tease apart separate paths and 2. how applicable the method is to more finely structured populations with recent admixture compared to populations with deep structure.

I simulated demographic models that contained two, four, six and eight paths. All demographic models started with a single trunk population that through binary population splits, divides into several populations corresponding to the number of paths. These populations remain separate for around a specified number of generations (10, 50, 100, 500, 1500 or 3500) before admixing successively with each other until one population remains (Figure 2.11). The same random seed was used to simulate models of the same path number, so all models with the same path number but different divergence times had the same ordering of population splits and admixtures.

Figure 2.12 shows the accuracy and standard deviation for classifiers trained on all demographic combinations of path number and separation times. The accuracy decreases as the number of paths in the model increases and as the number of generations that all the paths are diverged decreases. All path numbers show a rapid increase in accuracy from 100 to 500 generations of divergence. The more generations that the 'path' populations are separate, the larger the allele frequency differences become between paths for pre-existing variants due to drift acting for a longer period of time. Additionally more mutations accumulate that differentiate paths while paths are separated. More mutations and greater allele frequency differences means that RELATE is better able to resolve the correct topologies, which results in the GNNs looking more consistent within each class.

More paths increases the number of classes that the classifier must differentiate between and therefore the opportunity to confuse between those classes. Models with more paths

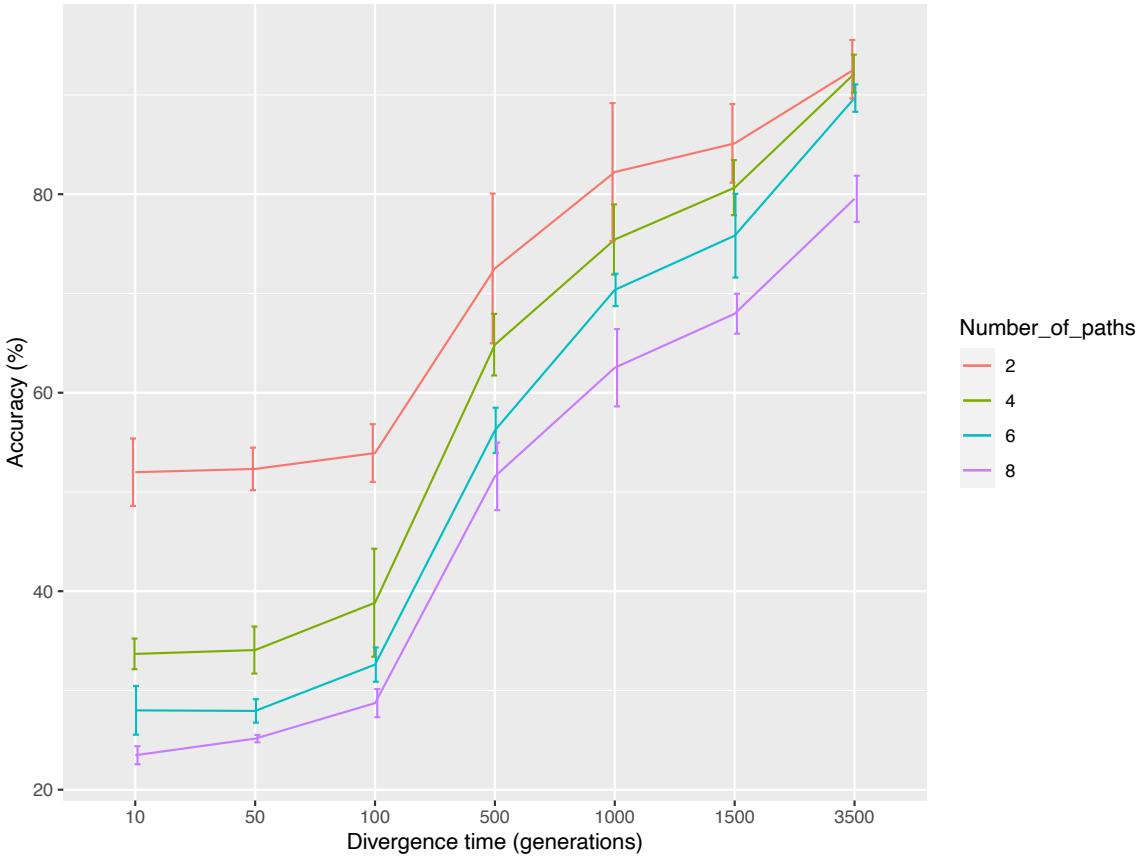


Fig. 2.12 Plot showing the mean and standard deviation of accuracy of classifiers trained on models with different path numbers and 'path' population divergence times.

will require more time of divergence to achieve the same accuracies as models with fewer paths. Models with only two paths maintain accuracies of above 50% with as few as ten generations of separation time. By 3500 generations (∼100,000 years) of divergence, models containing two, four and six paths have accuracies above 90% and overlapping error bars, showing that with enough time for divergence the decrease in accuracy due to path number can be mitigated.

Overall, models with smaller path numbers are better suited when there is very fine scale, recent structure involving closely related populations. Likewise, when the populations are very diverged and deep structure is present, models containing more paths are viable.

Inference with imbalanced population size

Next, I tested how imbalance in population size on paths affects classification. I used a demographic model of four paths and 1500 generations of separation between 'path' populations. One 'path' population was chosen to have a population size of 50,000 from its emergence after a split to its disappearance after an admixture. All other population sizes were 10,000. A control set of tree sequences were simulated with the same population split and admixture events but with all population sizes set at 10,000.

Compared to the control demographic with an accuracy of 80.66% +/- 2.48%, the classifier trained on the imbalanced population size demographic demonstrates a significantly lower accuracy of 73.45% +/- 0.71% (two sample t-test p=0.0033). Comparing the confusion matrices (Figure 2.13), path 3, containing the differently sized population, exhibits the drop in precision and is mostly confused with path 4. This makes sense given paths 3 and 4 are sister paths, descending from a common population that split (Figure 2.11). A larger population size will reduce the number of coalescence events occurring around the time of higher population size compared to the other paths, pushing coalescences into older time periods before paths 3 and 4 separated. Many GNNs for path 3 will not have a coalescence event falling in the period of higher population size, making them look like path 4 or other GNNs and so are misclassified.

Inference with imbalanced sampling

It is characteristic of ancient DNA datasets that some populations may only be represented by a few samples, while others many. Rather than subsetting some groups to match the sample size of the groups with the fewest samples, I tested if an imbalanced number of samples taken from each group alters the classification of certain paths. I simulated from three demographics with divergence time of 1500 and four paths and trained a classifier to each. In the control demographic model, the sample size was 25 diploids for all 'path' and

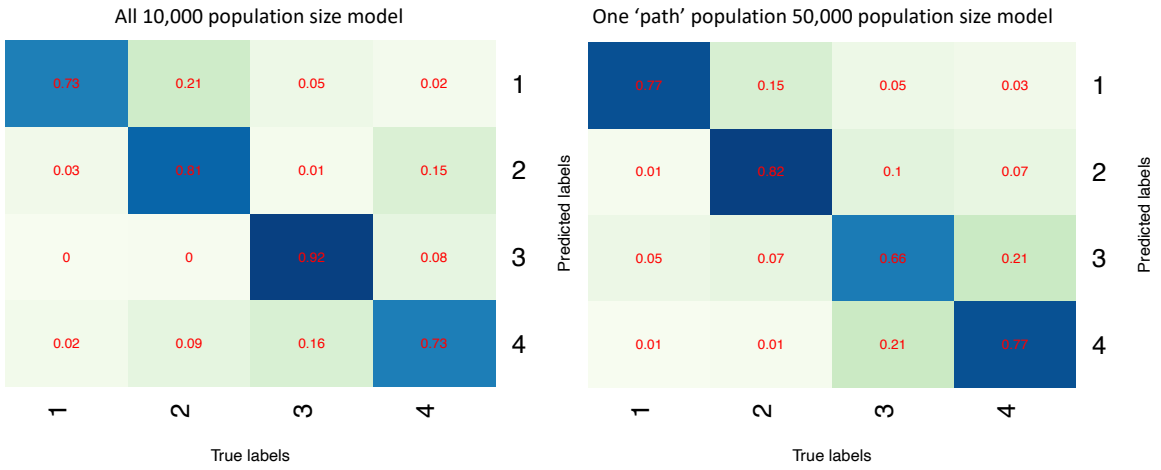


Fig. 2.13 Confusion matrices of the model with all populations with size 10,000 compared to the model where one population has size 50,000.

Table 2.4 Table displaying the mean and standard deviation of accuracy of classifiers trained on models with variable number of samples taken from a population along path 1.

Number of samples from path 1 population	Mean accuracy (%)	Accuracy standard deviation (%)
5	80.22	4.48
25	80.66	2.77
50	82.49	2.65

admixed populations. In the other demographic model, one 'path' population sample size was either reduced to 5 diploids or increased to 50 diploids. The same order and timing of split and admixture times were used, and all population sizes were 10,000. Path 1 contained the population with variable samples.

There was no significant difference in overall accuracy of the classifiers when tested pairwise with two sample t-tests (p-values = 0.318, 0.856, 0.364). Table 2.4 shows the accuracies for each model and the standard deviations. The model with 5 sampled diploids on path 1 has the highest standard deviation. This is likely because there are fewer examples

of path 1 labels from those samples in the training data and so the training has not captured as much of the variance of path 1 GNNs as other classes when training the classifier. Testing that classifier, this translates to greater variance in the accuracy.

In the confusion matrices (Figure 2.14) path 1 has greater precision in the 50 diploid model than the other two models, which have comparable precision values in path 1. However, the overall mean accuracies are not significantly different suggesting there is no systematic bias due to imbalanced sample sizes, rather an increase in precision due to greater sample size on one path.

Inference and overall sample size

Lastly, I tested the effect of overall sample size on classification ability. I simulated from seven different demographic models with a divergence time of 1500 and four paths and trained a classifier to each. Each had a different number of diploids sampled evenly from the admixed and 'path' populations: 5, 10, 15, 20, 25, 30 or 50 diploids. The same order and timing of split and admixture times were used, and all population sizes were 10,000. The same number of training GNNs were used to train classifiers for each demography so as not to confound results with different amounts of training data.

Figure 2.15 shows how the mean accuracy increases as the number of diploids taken from all sampled populations increases. There is a rapid increase in accuracy between 5 and 20 diploids, after which the increase in accuracy with more samples is less. At 50 diploids the accuracy is around 90% and the standard deviation is low. Even with only 10 samples, with a 1500 generation diverged four path demography, the accuracies can be over 70%.

More samples means more variation in the training GNNs, given that I used the same number of training GNNs from each demography. This prevents overfitting of the neural network allowing for greater flexibility to novel testing GNNs. This results in greater accuracy and smaller standard deviations upon testing.

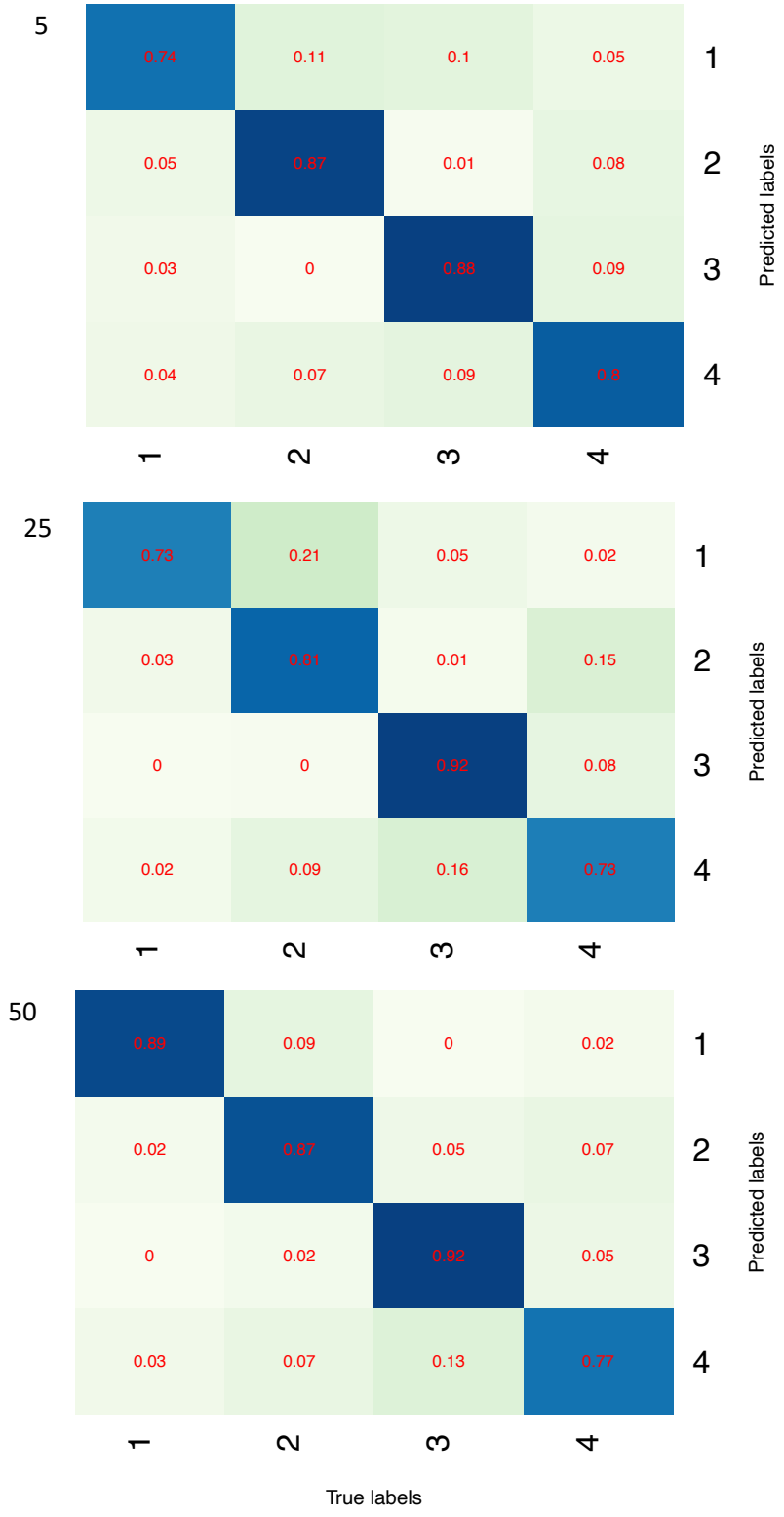


Fig. 2.14 Confusion matrices of models with different number of diploids sampled from a path1 population.

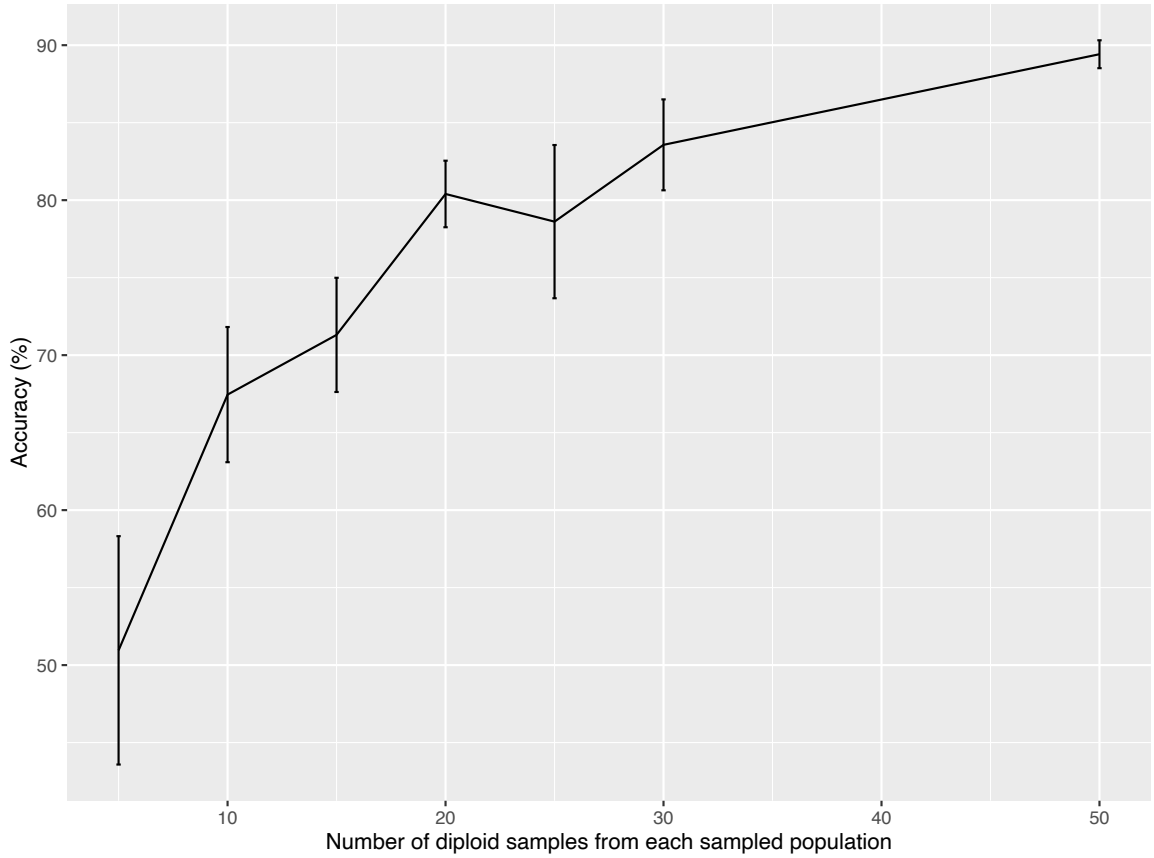


Fig. 2.15 Change in accuracy as the number of diploids sampled from all admixed populations increases. Errors bars show the standard deviation over five testing tree sequences.

2.4.4 Testing model misspecification

In Section 2.4.1 I showed that the model of population history in Europe produced data that, from F_{st} statistics and PCA, looked similar to the real data. My neural network for determining European local ancestry is trained on simulated data that I believe to match the real data well enough. However, the true history of any population is never known exactly and the following section explores how a classifier copes when it is trained on data that does not match the testing data in various ways. For all investigations I used demographies with four paths and 1500 generations of path separation time. Except where specified, simulations were carried out in the same way as described in Subsection 2.4.3.

For each parameter under scrutiny, I simulate two datasets with only the parameter under scrutiny differing between the two and train a classifier to each. Testing within the datasets was performed as in Subsection 2.4.3. When testing between datasets, the classifier of one was applied to each of the five testing tree sequences from the other dataset. This results in a mean accuracy and standard deviation for all four combinations of classifiers and testing data, to which I can then apply two sample t-tests to determine whether there is a significant difference in classifier accuracy when applied to testing data that does not match the training data. The two classifiers are then applied to pooled testing data in all four combinations to produce confusion matrices.

Inference with different population size

Using the same simulations and classifiers as those used for testing an imbalance in population size along paths, I tested the two classifiers trained on a 10k model and an imbalanced 50k model on GNNs extracted from the alternative demography. A difference in population size in a 'path' population will change the GNN structure, where a smaller population size will produce more coalescence events in the path population and a larger population size will conversely produce fewer coalescences. A difference in either direction

in the GNN structure between training and testing data will result in the paths being less easy to recognise by the classifier. Predictably there is a significant decrease in accuracy when test data is classed by the classifier trained on different training data to when classed by the classifier trained on the corresponding training data (Table 2.5). Two sample t-tests produce p-values of 0.007 and 0.0067.

There is no significant decrease in accuracy when the 50k classifier is applied to 10k data compared to 50k data (p-value = 0.418). However there is a significant decrease in accuracy (p-value = 9.661e-05) when the 10k classifier is applied to 50k data compared to 10k data. An unexpected deficit of coalescences along a path is more confusing to a classifier than a surplus i.e the absence of a defining GNN is more detrimental than the presence of an extra GNN.

Table 2.5 Table displaying the mean and standard deviation of accuracy of classifiers trained on models with the same or different population size than the testing data, along path 3.

Classifier \ Testing data	10k all	50k imbalanced
10k all	80.66. +/- 2.77	68.95 +/- 2.12
50k imbalanced	74.52 +/- 2.61	73.45 +/- 0.71

Inference with different admixture times

To explore how the classifier generalises to data with different admixture times, I simulated from one demographic model with admixture events occurring every 15 generations back in time from the present day and another every 30 generations. So the latter did not result in a smaller time while the 'path' populations were diverged, which would confound the results, I increased the time of the most recent population split in that demographic model.

Table 2.6 shows that the accuracy for the model with 30 generations separating admixture events is not significantly different than the accuracy of the model with 15 generations, when tested on their corresponding testing data and trained classifier (p-value = 0.976). When I

swap the classifiers, to test the 30 generation separated testing data with a 15 generation separated classifier and vice versa, there is no significant change in accuracy in either case compared to testing corresponding data and classifiers (p-values = 0.977, 0.716). The classifiers are able to generalise and compensate for the difference in admixture times. This indicates that, despite providing the node ages in the GNNs, the classifiers are largely using the topology of trees and not the coalescence times.

Table 2.6 Table displaying the mean and standard deviation of accuracy of classifiers trained on models with the same or different admixture times than the testing data.

Classifier \ Testing data	30 gen. admixture	15 gen. admixture
30 gen. admixture	80.58 +/- 4.81	81.27 +/- 2.33
15 gen. admixture	80.49 +/- 4.65	80.66 +/- 2.77

Inference with different admixture fractions

I next investigated how different admixture fractions would change the classification accuracy. We simulated a dataset with all admixture fractions 50/50 and another with all 25/75. There is no significant change in accuracy when the 50/50 classifier is applied to 25/75 testing data compared to testing the corresponding 25/75 data and classifier (p-value = 0.393). Neither is there a significant change in accuracy when the 25/75 classifier is applied to 50/50 data compared to testing the corresponding 50/50 data and classifier (p-value = 0.589). This suggests that classifiers are able to generalise when the admixture fractions are misspecified.

Table 2.7 Table displaying the mean and standard deviation of accuracy of classifiers trained on models with the same or different admixture fractions than the testing data.

Classifier \ Testing data	50/50	25/75
50/50	80.66 +/- 2.77	85.53 +/- 1.95
25/75	79.58 +/- 3.27	86.82 +/- 2.51

Inference when samples are drifted from ancestral populations

For all models we have been simulating samples that are taken directly from the simulated populations. The ancient samples we have in the MesoNeo dataset are unlikely to be individuals from the ancestral populations involved in the admixture events themselves, but instead more or less closely related to them. To investigate the effect of drift between the true ancestral admixing populations and sampled populations, we simulated population splits so that the ancient samples were taken from 'hanging branches', slightly diverged from the admixing lineages. A separation time of ten generations from the true ancestral populations simulates approximately 300 years of drift. We trained classifiers using a demographic model where the ancient samples are taken directly from the ancestral admixing populations and tested its performance on GNNs from the drifted model.

Table 2.8 Table displaying the mean and standard deviation of accuracy of classifiers trained on models with ancient samples that were drifted or not from the true ancestral populations.

Classifier \ Testing data	Drifted	Not drifted
Drifted	83.28 +/- 1.44	84.60 +/- 0.78
Not drifted	81.51 +/- 3.32	82.63 +/- 3.79

Table 2.8 shows the accuracy results when classifiers are testing on different testing data. There is no significant change in accuracy between any two pairs of testing data and classifier combinations demonstrating that classifiers are able to generalise between drifted and directly sampled models (p-values = 0.106, 0.314, 0.634, 0.320, 0.733, 0.122).

Inference for additional samples

The number of ancient samples that are available continues to increase and so more samples relevant to population histories can be incorporated into datasets. Likewise, samples may be removed from datasets after more stringent filtering. To test whether classifiers can generalise to more or fewer samples, I applied a classifier trained on a simulation with 20

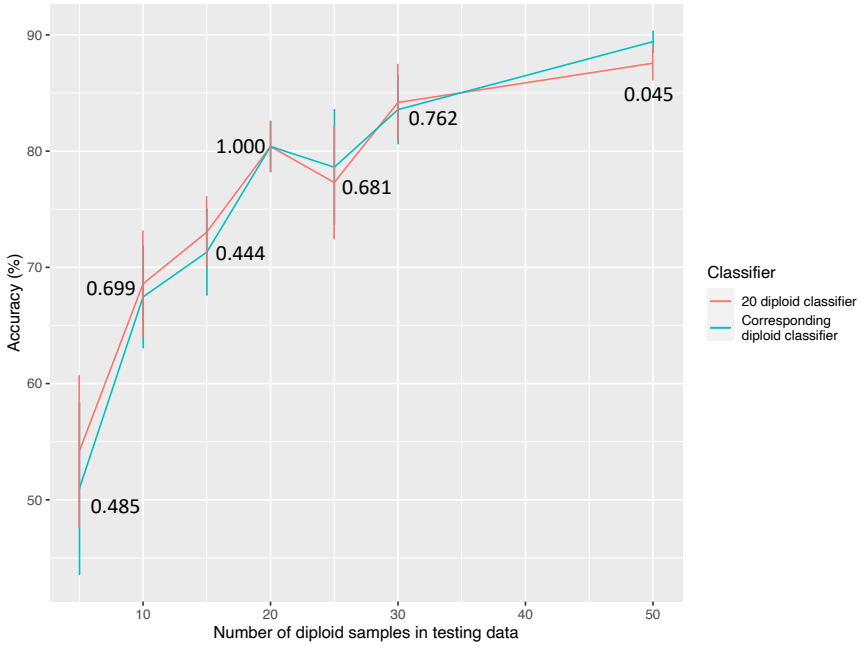


Fig. 2.16 Change in accuracy as the number of diploids sampled from all sampled populations increases using a classifier trained on 20 diploids vs the corresponding classifier trained on the same number of diploids as the testing data. Errors bars show the standard deviation over five testing tree sequences. Numbers show the two sample T-test p-values comparing the two classifiers.

diploids taken from all sampled populations to testing data from a simulation with a different number of diploids per sampled population (5, 10, 15, 25, 30, 50). GNNs were extracted using all samples in the demography, as if testing a corresponding classifier.

Figure 2.16 shows that there is no significant difference when using the classifier trained on 20 diploids compared to the corresponding classifier for testing data containing 5, 10, 15, 25 ad 30 diploids (p-value>0.05). For data containing 50 diploids there is a marginally significant decrease in accuracy using the 20 diploid classifier.

Classifiers are able to generalise to fewer samples than in the training data. When the number of samples is much larger than simulated in the training data, the accuracy begins to decrease. This decrease is only to a small extent even for more than twice the number of samples demonstrating that classifiers are flexible to differences in sample size.

Inference of ghost lineages

Genetic evidence has revealed ghost populations in many species, including humans. This is when a population is inferred to have existed but is not sampled with DNA or in the fossil record. To test whether my method could be used in a case involving a ghost population, I applied a classifier trained on a simulation where one 'path population' of the four paths present was not sampled.

I tested two scenarios from Figure 2.11 : One where the samples from population 0 were removed meaning path 1 contained the ghost lineage; and one where samples from population 9 were removed meaning path 4 contained the ghost lineage.

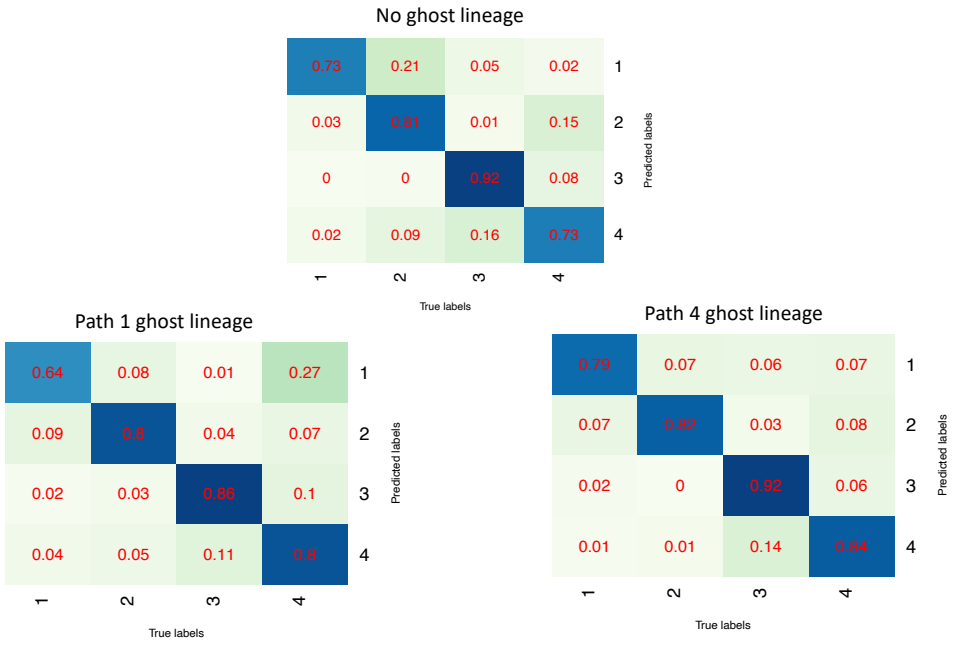


Fig. 2.17 Confusion matrices showing the precision of classifiers in each class when various populations where removed from training and testing data to mimic ghost lineages.

Figure 2.17 shows that the precision in all paths is not decreased when path 4 contains the ghost lineage and the precision of the classifier to predict path 4 actually increases. It appears that the classifier receives enough information from other populations present along path 4 to continue to identify path 4 with high precision.

When path 1 contains the ghost lineages, the precision of the classifier is decreased in path 1. This is likely because there are no other populations along path 1 that are included in the GNNs, meaning that when population 0 is removed it becomes harder to identify the path.

Overall, the accuracy remains high for identifying ghost lineages. The precision maintained depends on whether there are other sampled populations involved that can inform the classifier of path identity. The ability to identify ghost lineages, even with little to no other populations present along the lineages highlights the advantage of a path structure compared to a frame work of single population identity that is used by other local ancestry inference tools.

2.5 Conclusion

I have developed a method to infer local path ancestry using a neural network trained to recognise path labels from GNNs extracted from trees in a tree sequence. I have shown that the method is accurate for classifying European haplotypes into one of 6 paths and that it performs as well or better than an advanced local ancestry inference tool, GNOMix, especially in populations at greater time since admixture.

I have shown what sort of populations the method would be appropriate for and I have demonstrated that the method is robust to a variety of misspecifications in the underlying population model. My findings suggest that the choice of how complex to make the model will depend on what accuracy the user is willing to accept:

- For demographies with fewer paths, the populations can be more closely related to each other, or have experienced less time of separation.
- Demographic histories with large differences in effective population sizes may be less accurate.

- More samples means more accuracy, so I would recommend using all available samples even if there is an imbalance across paths. By including the same number of GNNs examples of each path and each different admixed population in the training data, one should be correcting for an imbalance.

- Fewer samples overall or on one path will be more vulnerable to overfitting and so it is tempting to simulate many samples from all paths to improve classifier accuracy. However, simulating the same number of samples as the real data will produce tree sequences that resemble those inferred from the real data more and so will produce classifiers that are more applicable to the real data. Simulating more independent training tree sequences to balance smaller sample sizes should mitigate the effect of overfitting.

I have demonstrated that classification with a path structure allows to identification of ghost lineages. This is a major advantage offered by a path framework over other local ancestry inference tools, that use single population identities as ancestries within which the presence of ghost lineages in chromosomes cannot be inferred.

The range back in time from the age of focal samples over which paths can be drawn is limited by the recombination rate and the availability of relevant population proxy genomes. The recombination rate will define how far back in time a haplotype can originate before the length seen in focal samples is too small to detect or differentiate from noise tracts, recombination decays linkage disequilibrium over time (Section 1.2). When there are no proxy samples available for populations to differentiate paths, paths will collapse together. This is especially harder deeper in time, highlighting the advantage of ancient samples.

Rather than being a 'black box', we can trace back through the GNNs and demographic structure why the overall accuracy and class precision is altered when various parameters are changed. When applied to different populations, this intricate understanding of how the

classifier is working helps when drawing deeper conclusions and making decisions on how to build a demographic model.

Chapter 3

Estimating time since admixture and admixture fractions

Contents

Admixture events are pervasive throughout all human population histories. Some events happened recently in the past such as the admixture of Europeans and Africans in America, an example which has been widely used in genetic studies of admixture events. Travel across the world became easier during the last century, so admixture between people with relatively diverged ancestries became more widespread. With the ability to sequence large numbers of present-day people, these recent events can be studied.

Although more distant in time, ancient admixture events also shaped modern population genetics and therefore affect the lives of modern people today. As demonstrated in the model of European population structure, Figure 2.3, multiple large-scale migrations and admixtures are key features of European history. Importantly, it was through these events that lifestyle changes were precipitated. Knowing when admixture occurred across the continent and in what proportions informs discussion on the speed of migration, under what conditions admixture may have been favourable and also helps with investigating natural selection and heritable diseases.

Signs of ancient admixture events are harder to see in modern genomes since the signal decays exponentially with time and can be obscured by subsequent demographic processes. Ancient genomes provide us with genetic material that was created closer to the time of admixture, making studying these ancient events easier.

3.1 Existing approaches

Existing approaches that aim to date admixture events from genomic data can be split into two categories: Those that use linkage disequilibrium (LD-based) and those that use ancestry tract lengths (haplotype-based). Both are founded on the understanding that admixture brings genetic material from two (or more) diverged populations together and that over time recombination breaks up the segments of the chromosome that originate in each ancestral population into smaller and smaller chunks. The result of this process is two-fold: admixture LD decreases exponentially with genetic distance; and the distribution of tract lengths can be modeled by an exponential decay function. The amount of exponential decay of both measures depends on the time since admixture under the assumption that recombination in each generation occurs according to a Poisson process with rate 1 along the chromosome when using genetic distance units. It is therefore possible to extract the time since admixture

by charactering decreasing LD with genetic distance [84, 68, 40, 130, 95] or the distribution of tract lengths [65, 96, 36].

Haplotype-based methods first perform local ancestry inference and then characterise the distribution of tract lengths to obtain an estimate of the time since admixture. These methods tend to be harder to perform since they require phased data and accurate local ancestry inference, which can be difficult without good reference ancestral populations. Errors in local ancestry inference break up ancestry tracts making blocks appear smaller than in truth and thus results in overestimated time since admixture. These errors are more detrimental to large ancestry tracts and so haplotype-based methods are less accurate for predicting recent admixture events.

LD-based methods are more flexible as they do not require phased data and progress has been made in their ability to distinguish between background LD and admixture LD beyond simply using a minimum starting distance [68]. Unlike haplotype-based methods, small errors in local ancestry assignment or admixture LD can be smoothed over by using multiple chromosomes and thus these methods are more robust to noisy data [18]. Additionally some LD-based methods aim to assess whether data can be modelled better by a mixture of exponential decay curves, indicating that multiple admixture events or continuous admixture has occurred [40, 130, 95].

At very large times since admixture (>200 generations), the ancestry blocks are eroded by recombination to sizes that are too small to be distinguished from background LD, thus both these methods have an upper bound on the time since admixture that can be inferred.

3.2 A method to estimate time since admixture

Given the robustness to noisy ancestry inference, I devised an LD-based method to infer the time since admixture of samples from my painted chromosomes that is similar to that used in GLOBETROTTER [40]. Segments are painted by path, so admixture corresponds to

when two paths join in a population history. This means that tracts made up of two or more paths combined can be used to date multiple admixture events, where one event that joins multiple paths follows other events that join two or more paths. For example, the joining of paths 1, 2, 3 and 4 in the Bronze Age admixture event in Figure 2.3 is preceded by paths 1 and 3 joining in the Neolithic farmers and paths 2 and 4 joining in the Yamnaya. This opens the possibility of being able to date older admixture events of the parent populations from within the later admixed samples, for example, dating the admixture of EHG and CHG from within bronze age samples using Yamnaya chunks. Likewise, the admixture fraction refers to the fraction of the genomes taking each path.

With empirical sampling I plot the probability of being in the same path as a function of genetic distance and fit an exponential decay curve to the distribution. The parameters of the exponential decay correspond to the values of interest, time since admixture and admixture fraction. This can be done for each sample chromosome individually and the results across the whole genome for each sample combined to give an admixture time estimate and admixture fraction for each individual.

The sampling process is as follows:

1. Sample a starting position uniformly from between 0 and 0.5cM from one end of the chromosome.
2. Record the starting path.
3. Move 1cM towards the end of the chromosome.
4. Record the path 1cM away.
5. Now move to a new starting position 1cM away from the previous starting position.
6. Repeat steps 2-5 until the end of the chromosome.

Repeat the above steps for testing distances 1-50cM in step 3 and 4. Over all autosomes, I can then calculate the probability of being in path y given starting in path x, d cM away, for all path combinations of x and y, including x=y. The shortest distance is 1cM, as in ROLLOFF, to avoid the effect of background LD.

The probability when x=y decays exponentially with genetic distance and the rate of decay depends on the time since admixture. Parameters are determined by nonlinear least squares regression of the data to the formula

$$P(\text{same path}) = \alpha e^{(-\beta d)} + \theta \tag{3.1}$$

The probability will asymptote to the probability of being in path x across the whole chromosome, which intuitively is the admixture fraction of path x, therefore the admixture fraction is given by θ. The β parameter represents the time since admixture in generations (Figure 3.1).

For samples which are the product of admixture between populations that are themselves admixed, I combine the paths that make up each parent population, creating larger admixture chunks. This opens the possibility of being able to date older admixture events of the parent populations from within the later admixed samples, for example, dating the admixture of EHG and CHG from within bronze age samples using Yamnaya chunks.

3.3 Performance on simulated data

To test the performance of this method, I simulated data from the model of European population structure described in Section 2.2. I executed the analysis on the admixed populations to infer time since admixture and admixture fractions of paths.

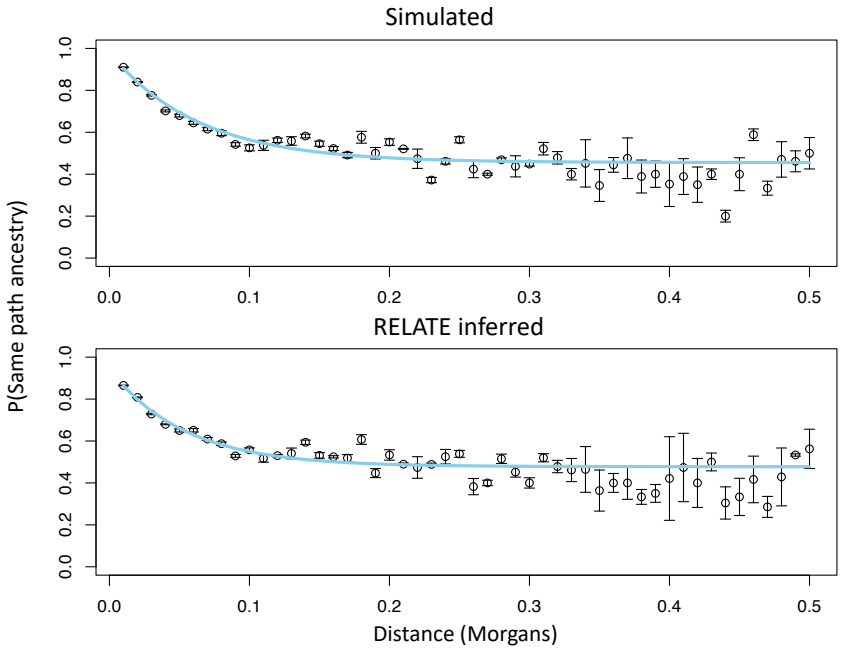


Fig. 3.1 Exponential decay curves fit to the decreasing ancestry correlation at increasing distances between two positions along the chromosome. This example is from fitting curves to the Neolithic farmer chunks (paths 1 and 3) to a Bronze Age sample simulated at 19 generations since admixture. The β estimate for the simulated data is 18.05 generations and for the RELATE inferred data the estimate is 18.64 generations.

3.3.1 Time since admixture analysis

I dated the admixture times of all individuals by extracting the β parameter from fitting exponential decay curves as described above. Counts for all distances were taken across all nine tree sequences and the two haplotypes from each individual were treated independently. The probability of remaining in the same path was therefore calculated using the total counts from across 18 independent sequences per individual. For admixture events between two populations, each individual has two estimates of the admixture age, one calculated per ancestral path. I combine the estimates of the two in a way that minimises the standard error to give a weighted average value of time since admixture for each individual. The time of admixture in generations ago from present-day can therefore be calculated as the sum of the sample age in generations ago and the estimated time since admixture for that sample. The method was applied on both the simulated painted tree sequences and on tree sequences that were inferred by RELATE from the simulated data and painted using a classifier. The results are shown in Figure 3.2 and Tables 3.1 and 3.2.

Table 3.1 Table of results of admixture time analysis performed on simulated painted tree sequences. For three admixed populations the mean time of admixture across all simulated samples and the standard deviation of estimates around the mean is shown. All values are in units of generations ago.

Population	Simulated time of admixture	Mean time of admixture	Standard deviation (+/-)
Bronze Age	166	166.03	3.86
Neolithic farmers	259	258.54	12.96
Yamnaya	177	177.74	1.66

In Figure 3.2, as the sample ages become more recent and further from the admixture time, the variance of estimates around the true value increases for both simulated and RELATE inferred results. Likewise, Figure 3.3 shows how the standard error of estimates increases as the sample ages get further from the time of admixture. More generations since admixture

Table 3.2 Table of results of admixture time analysis performed on tree sequences that were inferred by RELATE from the simulated data and painted using a classifier. For three admixed populations, the mean time of admixture estimate across all simulated samples and the standard deviation of estimates around the mean is shown. All value are in units of generations ago.

Population	Simulated time of admixture	Mean time of admixture	Standard deviation (+/-)
Bronze Age	166	166.45	4.56
Neolithic farmers	259	257.28	16.67
Yamnaya	177	177.24	2.00

mean that more recombination events have occurred resulting in shorter ancestry tracts. Shorter tracts make it harder to differentiate between background LD inherited from the parent populations from the relevant admixture LD. Therefore the standard error for the β parameters when fitting exponential decay curves is greater and the variance in β parameters estimates is greater. This is the case for both simulated data and RELATE inferred data showing there is a limit to the length of admixture tracts, even in data with no noise, for inferring time since admixture accurately.

Noise appears as short tracts of misclassified correlated trees in RELATE inferred data, which starts to become difficult to differentiate from admixture LD when the admixture tracts become comparable in length. The standard error values for the analysis of tree sequences inferred by RELATE up to 100 generations since admixture are very similar to those from the analysis of the simulated data (Figure 3.3). This suggests that up to approximately 100 generations since admixture the method is robust to noise introduced by classification error in the painted chromosomes. Similar results can be seen in Figure 3.2, as the sample ages become more recent and further from the admixture time, the deviation of estimates from the true value increases in both sets of data but to a greater extent in RELATE inferred data. Likewise, the overall standard deviation of the mean time since admixture estimate across all

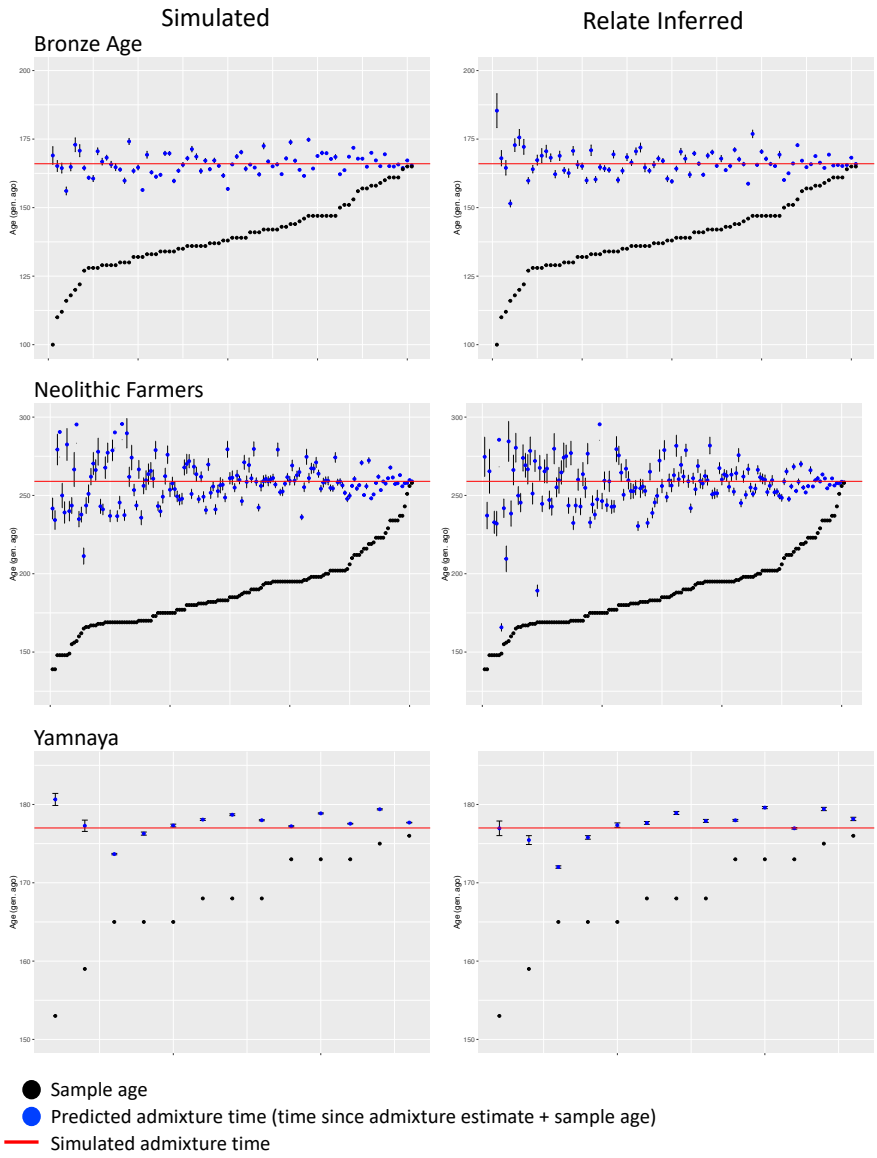


Fig. 3.2 Plots showing the predicted admixture time with standard error bars and the sample age for three simulated admixed populations. The time since admixture from each painted individual was calculated from the β parameter described in Section 5.2 and the time of admixture plotted from the sum of the time since admixture and sample age in generations ago.

samples is slightly larger in the RELATE inferred data compared to the simulated data for all populations (Tables 3.1 and 3.2).

In results of analysis of the simulated data, the Bronze Age mean admixture time predicted across all samples matches the true value of 166 (Table 3.1) and a small standard deviation

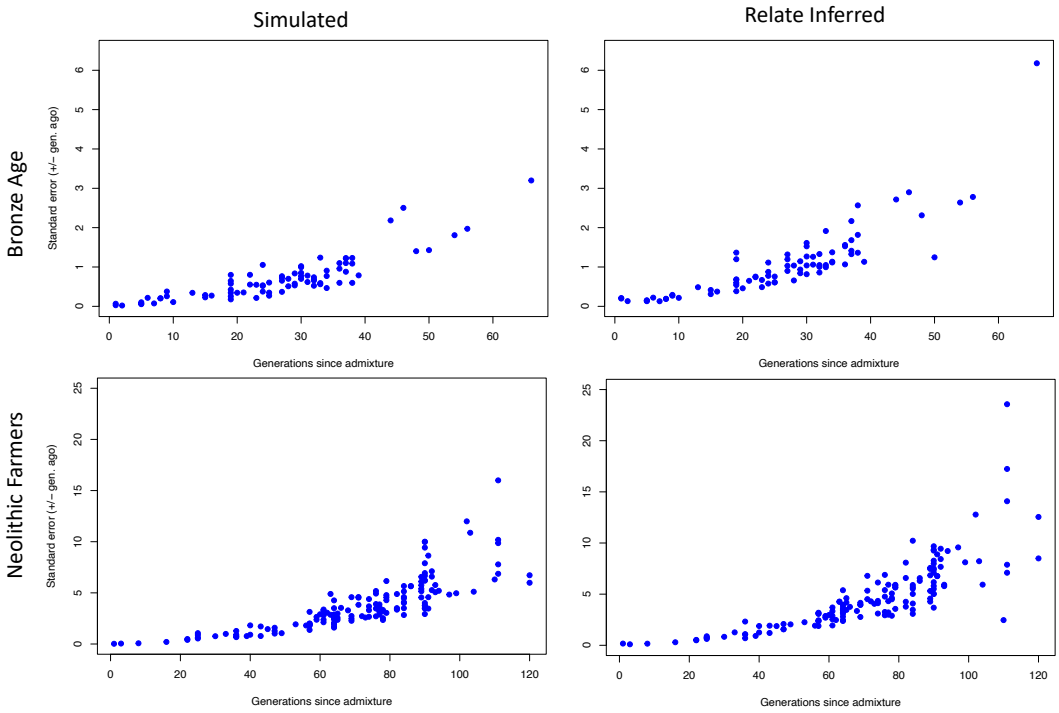


Fig. 3.3 Plots showing the decrease in standard error of admixture time estimates as sample age increases and gets closer to the simulated time of admixture for Bronze Age and Neolithic farmer populations. The left column shows the results for admixture time analysis on the simulated tree sequences and the right shows the results for tree sequences inferred by RELATE from the same simulated data and painted using a classifier.

of estimates of +/- 3.86 generations. Neolithic Farmers display a mean admixture time slightly under the true value of 259 and a larger standard deviation of estimates of 12.96 +/- generations. The slightly poorer ability to estimate Neolithic farmer admixture time is because some samples have ages that are further from admixture compared to the Bronze Age population, resulting in the mean being slightly under the true admixture time and a greater standard deviation. The error may also be increased by the ancestry proportions of 75/25 Anatolian and WHG respectively, as the exponential decays are harder to fit when the difference between the starting value, near 1, and the asymptotic value of 0.75 is not as large. Yamnaya samples are a maximum of 17 generations from the true admixture time of 177. The mean predicted admixture date matches that simulated and there is a small standard deviation of estimates of +/- 1.66 generations.

Results of analysis of RELATE inferred data were very similar to that of the simulated data (Table 3.2). The Neolithic farmer's mean β estimate is slightly further from the truth in the RELATE inferred data analysis than in the simulated data analysis due the same reasons explained above; more samples with a greater time since admixture, above 100 generations, meaning noise introduced by classification has more of an effect of decreasing the accuracy of the β estimates.

For the present-day samples that are all 166 generations since admixture, in both simulated and RELATE inferred tree sequences, the admixture tracts have become too short to produce a reliable estimate of time.

3.3.2 Admixture fraction analysis

With painted chromosomes, there are two ways to calculate the admixture fractions in individuals. The first method is to calculate the proportion of the genome in base pairs that is painted with each path ancestry by summing the distance covered by each path, divided by the total length of the simulated genome. I use admixture fractions for individuals calculated in this way directly from simulated data painted with the true paths as the true admixture fractions to compare the results from both methods to.

Figure 3.4 shows the fraction of the genome taking each path calculated over a simulated genome of nine tree sequences that were inferred from simulated data using RELATE and painted with a classifier. The four 'path' populations, WHG, EHG, Anatolians and CHG, have genomes painted by almost entirely one path. All samples contain a small proportion of the genome from paths that were not a component of those populations in the European model, the result of error in classification of trees.

To find the admixture fractions for admixed samples whose ancestral populations are themselves admixed and made up of two paths, I sum the proportions of the paths that make up each for the two ancestral populations. For example in Bronze Age individuals I sum path

2 and 4 that make up the Yamnaya and paths 1 and 3 that make up the Neolithic farmers (Figure 3.5).

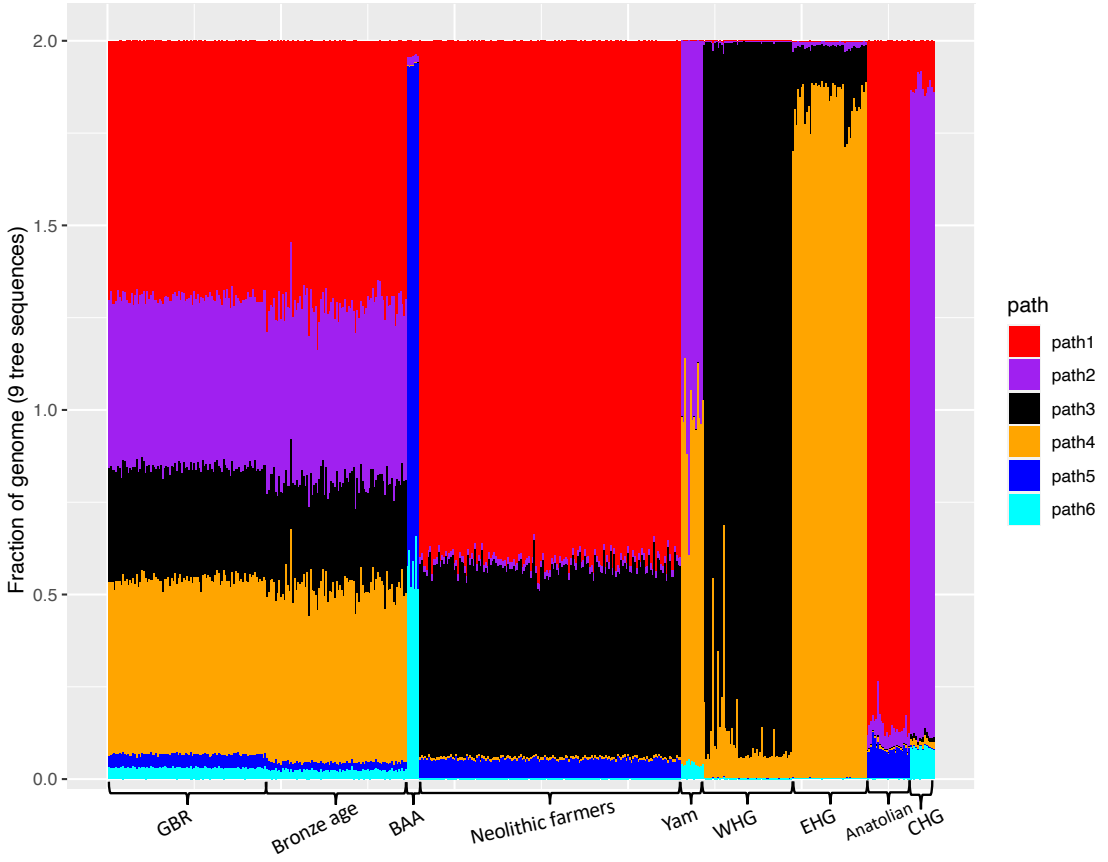


Fig. 3.4 Bar plot showing for each simulated individual, the proportion of the diploid genome taking each path. Analysis was done over nine RELATE tree sequences of 200Mbp in length each, that were inferred from data simulated from the European model of population structure. Colours correspond to paths matching those used in 2.3

Samples with older ages, that are closer to the time of admixture, tend to have proportions of component paths further from the fractions input to the simulation. The reason for this effect is that admixture events in msprime appear, forwards in time, as a new admixed population formed in a single generation of individuals from each ancestral population in the specified proportions. This means in the first few generations following an admixture event, large chunks of sequence that take the same path remain in the population as there have been few recombination events between sequences taking different paths. In addition, msprime

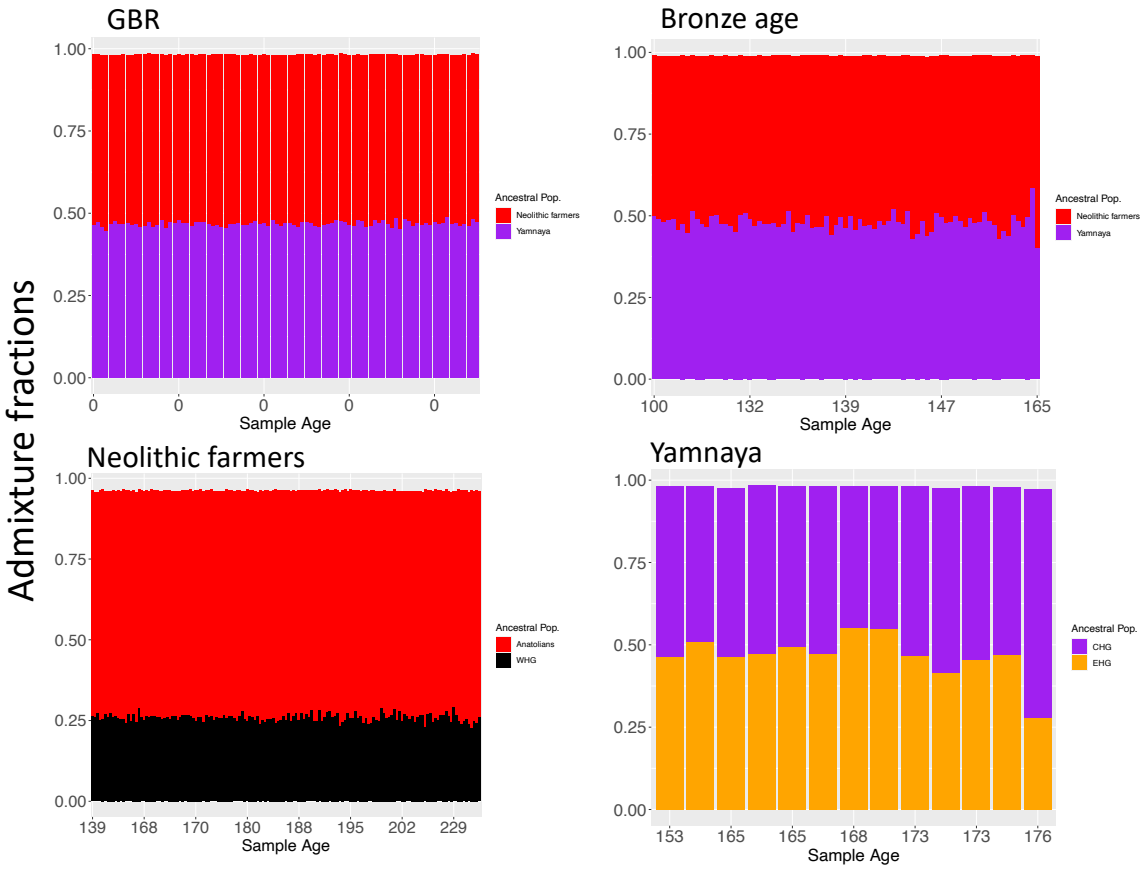


Fig. 3.5 Barplot showing for each admixed group the admixture fractions calculated by adding together the fractions of component paths if necessary from Figure 3.4. Samples are ordered along the x axis by increasing sample age, decreasing the time since simulated admixture.

simulation haploid sequences are arbitrarily paired up within the populations to form diploids. After only a few recombination events, by chance some arbitrarily paired diploid samples may have greater or lesser proportions of some paths than the true simulated fraction of individuals who admixed. After many more generations and recombination events between sequences taking different paths, without any further admixture events, the proportions of each path in haploid sequences are homogenised and closely match the simulated admixture fractions. However, sample age does not have any effect on the error in admixture fraction estimates and all errors are less that 0.1 (Figure 3.6).

The second method to calculate admixture fractions is by extracting the θ parameter from equation 3.1. When exponential decay functions are fit to decreasing ancestry correlation as genetic distance increases as described in Section 3.2, θ is the asymptotic value of the exponential decay curves. It represents the overall probability of being in a particular path across a sequence, which I assume is equivalent to the admixture fraction of that path. Noise in the data is in the form of short sections of correlated trees which are too short to fit exponential decay curves. This means, along with the curves being an approximate fit, the admixture fractions do not necessarily sum to 1.

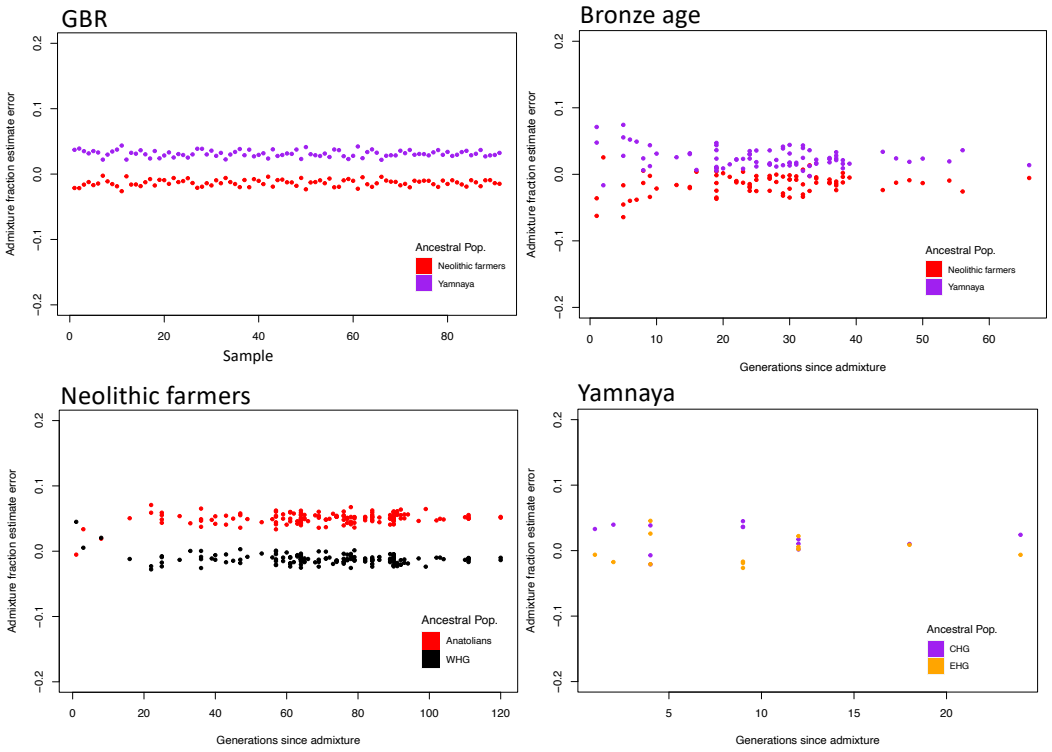


Fig. 3.6 The difference between admixture fractions in simulated data compared to θ value calculated from RELATE inferred, classified data. Each individual has two points, one for each ancestral population. Individuals are ordered by time since admixture.

Unlike the time to admixture analysis, samples with ages closer to the time of admixture exhibit θ values that tend to be further from the true simulated fractions (Figure 3.7). These samples have larger chunks inherited from ancestral populations which reduce power when

fitting the exponential decay curve to find the asymptotic value. Generally, the error is low for samples with at least 10 generations since admixture but those closer can have errors of up to 0.2 in fraction. Some ancestral populations display greater error likely because some paths are more often confused for another path by the classifier and so contain more noise chunks.

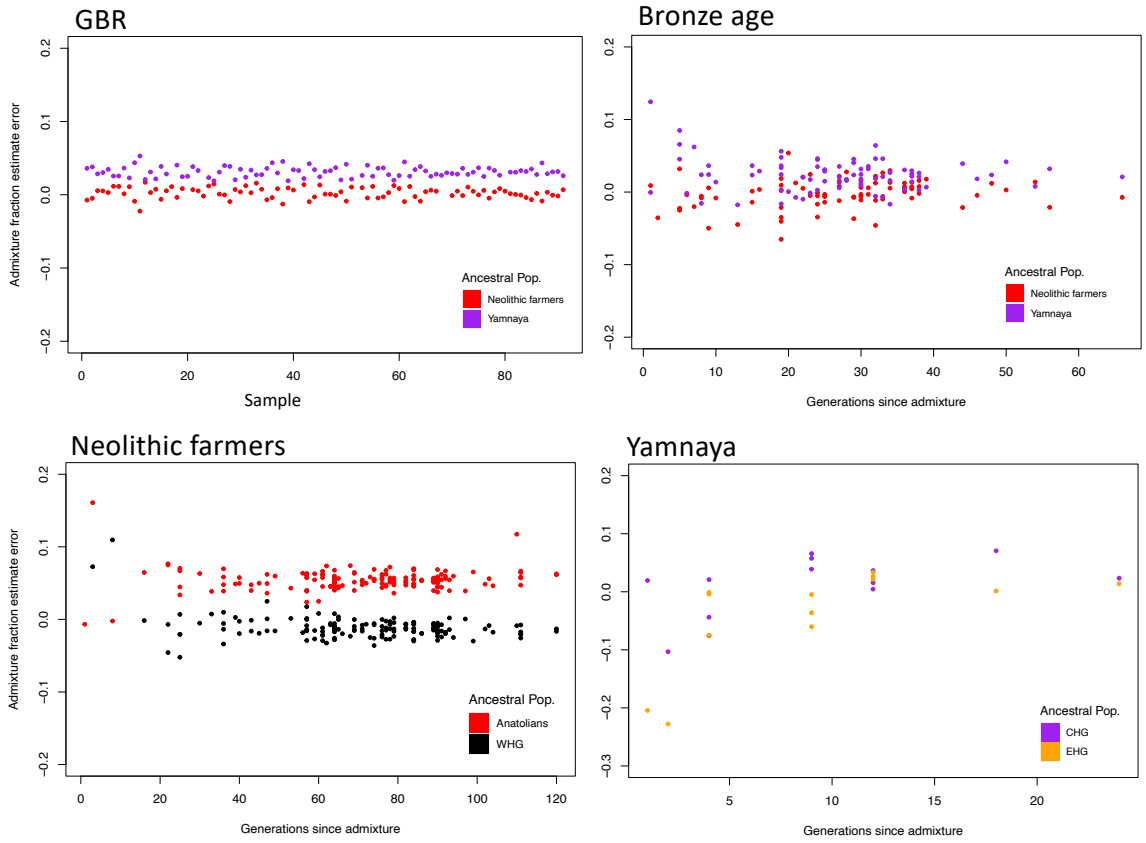


Fig. 3.7 The error in admixture fraction estimates, calculated as the difference between the true simulated fractions for each individual and θ from equation 3.1. Each individual has two points, one for each ancestral population. Individuals are ordered by time since admixture.

Overall, the first method for estimating admixture fractions by summing over base pairs produces results with very low errors for all individuals regardless of age. The second method of extracting θ has comparable error values but only for samples with ages more than 10 generations since admixture. For more reliable estimates of admixture fractions, I recommend the first method. However, θ is easy to obtain if a time since admixture analysis

is carried out anyway and so this method can be a convenient use of time and resources, bearing in mind the sample age limitation.

3.4 Pulse vs. continuous admixture

In the above Sections describing a method to infer admixture time, I assumed that admixture happens as a single event followed by panmixia in the admixed population. In reality admixture is unlikely to happen in a single pulse and is more likely to happen over a number of generations with continuous influx and admixture of migrants. A challenge faced by all methods that attempt to estimate the time since admixture is when admixture has been ongoing and more continuous in nature than a single hybridisation event. In the case of fitting exponential decay curves to tract length or LD data, continuous admixture will create a curve that is a mixture of multiple exponential decay curves. I can potentially tease apart the component decay curves in order to infer continuous admixture has occurred.

I explored one approach to identify more continuous admixture by fitting curves of up to decreasing distances i.e 1-50cM 1-40cM, 1-30cM etc. It is possible to fit different exponential decay curves with increasing rate parameter values, biassing towards older admixture events. This technique of fitting several exponential decay curves with different rate values to different parts of the curve is a promising approach to inferring multiple or continuous admixture events. However, there are identifiability problems as the estimates are very subject to noise so there is probably a limit to what can be done when making estimates of continuous admixture.

Chapter 4

Application of methods to MesoNeo genomes

Contents

In this Chapter I apply the method developed in Chapters 2 and 3 to the MesoNeo genomes and discuss the implications for the history of Europeans.

4.1 Inferring tree sequences and evaluating model fit

I inferred a RELATE tree sequence for each autosome of our subset of 476 MesoNeo genomes, using the fine scale recombination map for each chromosome, a mutation rate of 1.25e-8, and starting population size estimate of 30,000. All chromosomes for all samples were painted by a classifier trained on the model of European population structure as described in Chapter 2.

As a measure of how well the model fits the real data, I explored the extent to which coalescences are within paths, as expected according to the model, or between them in violation of the model. As a comparison, I simulated tree sequences with panmictic demography. All else remained equal to the structured model except there were no split or admixture events and only a single population. The population size was the sum of those on all paths of the structured model at any given time point. Samples were taken at the same times and arbitrarily assigned to the populations. RELATE tree sequences of this unstructured simulated data were inferred and painted using the classifier trained on the structured model using the samples with randomly assigned population labels.

For the MesoNeo dataset, the average number of coalescences within paths across all trees is almost always higher than those between paths (Figure 4.1). The values per path are comparable to those seen in the tree sequences inferred from data simulated from the model of European structure. In contrast, the simulated unstructured genealogies have similar within and between path coalescence numbers, with only those within pseudo-path 3 greater than all between path levels.

Some between path coalescences are permitted within the model framework during certain time periods. 'Illegal' coalescences are those that are not permitted in the model. For example, a coalescence event between a path 1 lineage and a path 2 lineage between 166 and 800 generations ago would count as illegal given the model. The legal/illegal coalescence ratio in Figure 4.2 is very high between 150 and 350 generations for the MesoNeo genealogies,

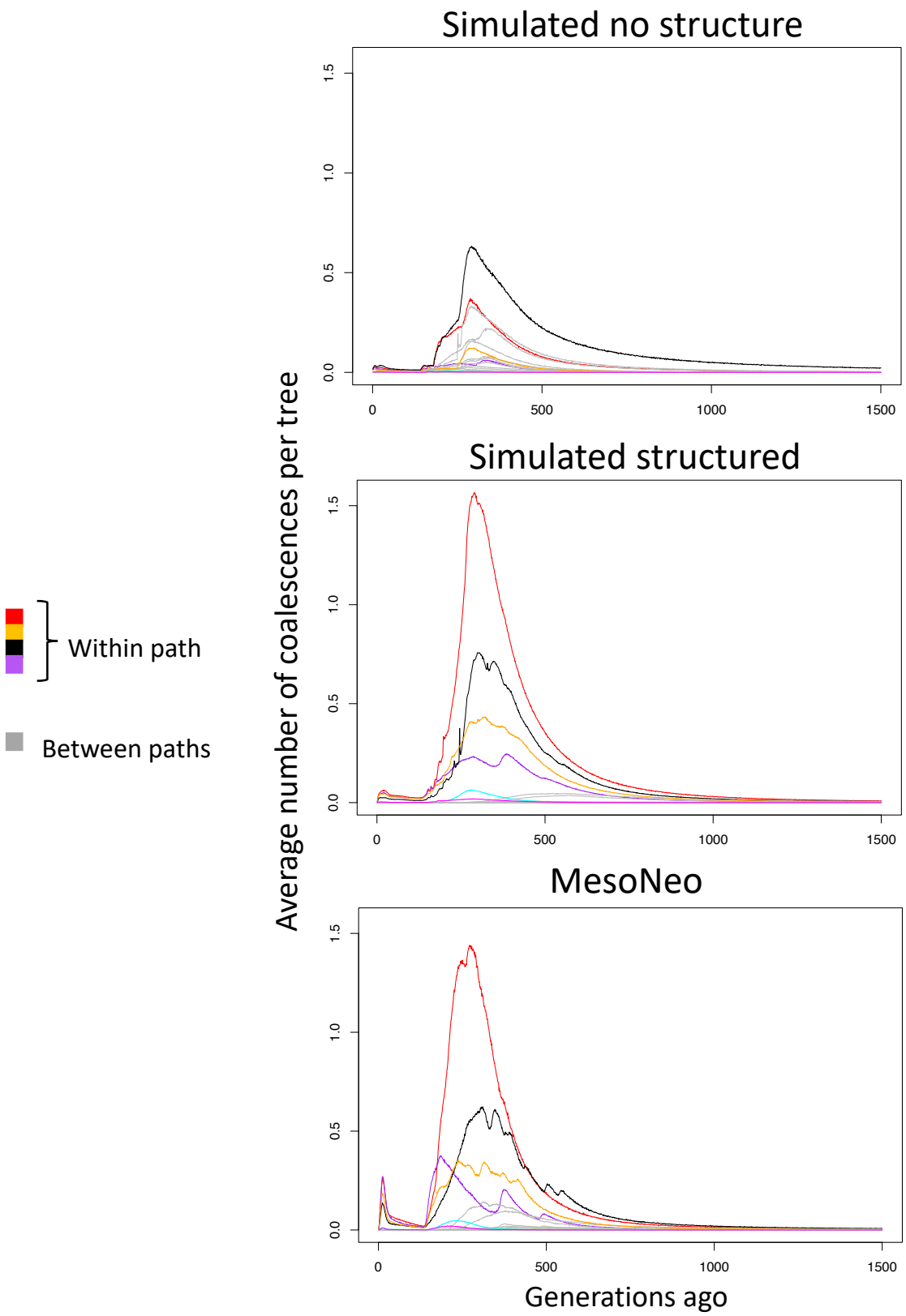


Fig. 4.1 Comparison of the counts of within and between path coalescences for tree sequences inferred from simulated structured data, simulated unstructured data and MesoNeo genomes, all painted using the same classifier.

whereas the unstructured genealogies have a low ratio throughout never reaching above 2. While the ratio is lower in the MesoNeo genealogies than the ratio in the tree sequences inferred from simulated structured data, it is always above that of the unstructured genealogies

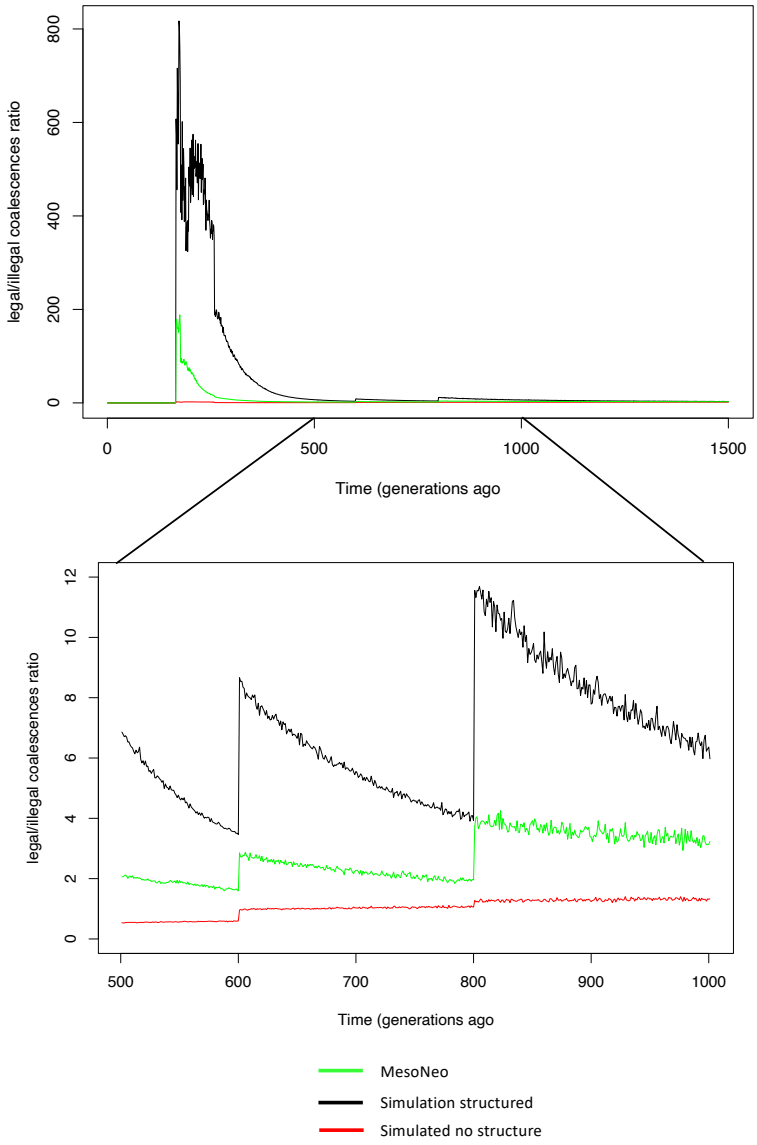


Fig. 4.2 Ratio of 'legal' to 'illegal' coalescences for tree sequences inferred from structured simulated data, unstructured simulated data and MesoNeo genomes. Legality is determined by which lineages assigned to paths are permitted to coalesce at given time-points within the model of European population structure. The top plot shows the ratio for the whole model time period. The bottom plot shows a 500 generation time section in order to see the difference in ratio better.

and only approaches 1 around 600 generations ago, at which point there is a population split in the model.

These results suggest that both the RELATE inference and the path classification have captured structure within the MesoNeo genealogies. The similarity of the counts of coalescences within and between paths through time of the MesoNeo genealogies and those simulated from the structured model support the PCA and F_{st} evidence from Chapter 2, that the model is a good representation of the real data.

4.2 Estimation of global admixture fractions

Global admixture fractions can be calculated for each individual as simply the proportion of chromosomes painted with each path. Figure 4.3 shows how the fractions change in the different populations. Yamnaya samples are painted as expected with primarily EHG and CHG paths, and Neolithic farmers by WHG and Anatolian paths. Only in Bronze Age samples do we see substantial steppe ancestry (EHG and CHG path) components in Western European samples, after the migration of steppe pastoralists into Europe.

There are minor components from other paths. Many of these are likely due to noise introduced in the RELATE inference and classification, but also potentially true signals are present from gene flow between groups that fits with the geographic ranges of these groups. In particular, WHG and EHG have a greater proportion of their genomes taking path 4 and 3 respectively, likely due to gene flow east to west between each other across the continent. There is also a substantial presence of path 1 in CHG genomes, likely due to gene flow between early Anatolians and CHG across the Caucasus mountains. The path 2 component in Anatolian genomes is smaller than the path 1 component in CHG, suggesting directional gene flow from Anatolia into the CHG. EHG genomes contain a proportion taking path 2, probably due to gene flow from CHG ancestry in eastern Europe/Western Asia.

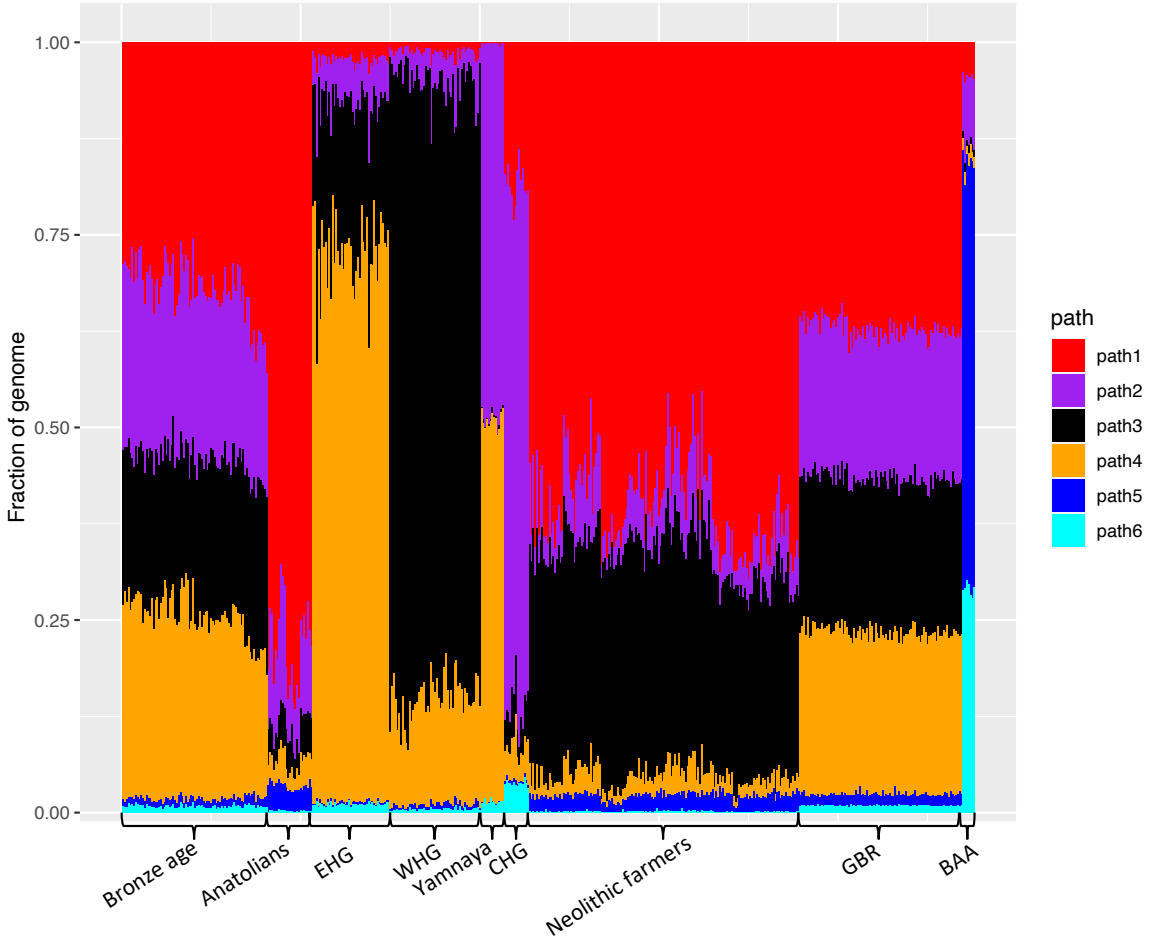


Fig. 4.3 Barplot showing for each simulated sample, the proportion of the genome taking each path. Analysis was done over all autosomes of the MesoNeo subset. Colours correspond to paths matching those used in 2.3

Figure 4.4 shows the admixture fractions of three admixed populations separately, ordered by increasing sample age of individuals. Component paths are added and coloured as one and presumed noise is not coloured. The level of WHG path ancestry in Neolithic farmers shows a slight resurgence as samples become younger at a rate of 0.027% increase per generation (Pearson r = 0.22, p-value= 0.00582). There appears to be no corresponding decrease in Anatolian path ancestry with time. This result is more consistent with a slow migration of Anatolians into Europe where the amount of admixture and fractions involve varies across the continent and at different times in different geographies. The resurgence in WHG ancestry in

the middle to late Neolithic has previously been shown [66, 15], a signal I have recovered. Perhaps the decrease in Anatolian path ancestry is masked when grouping all samples from across the continent and ordering by sample age alone.

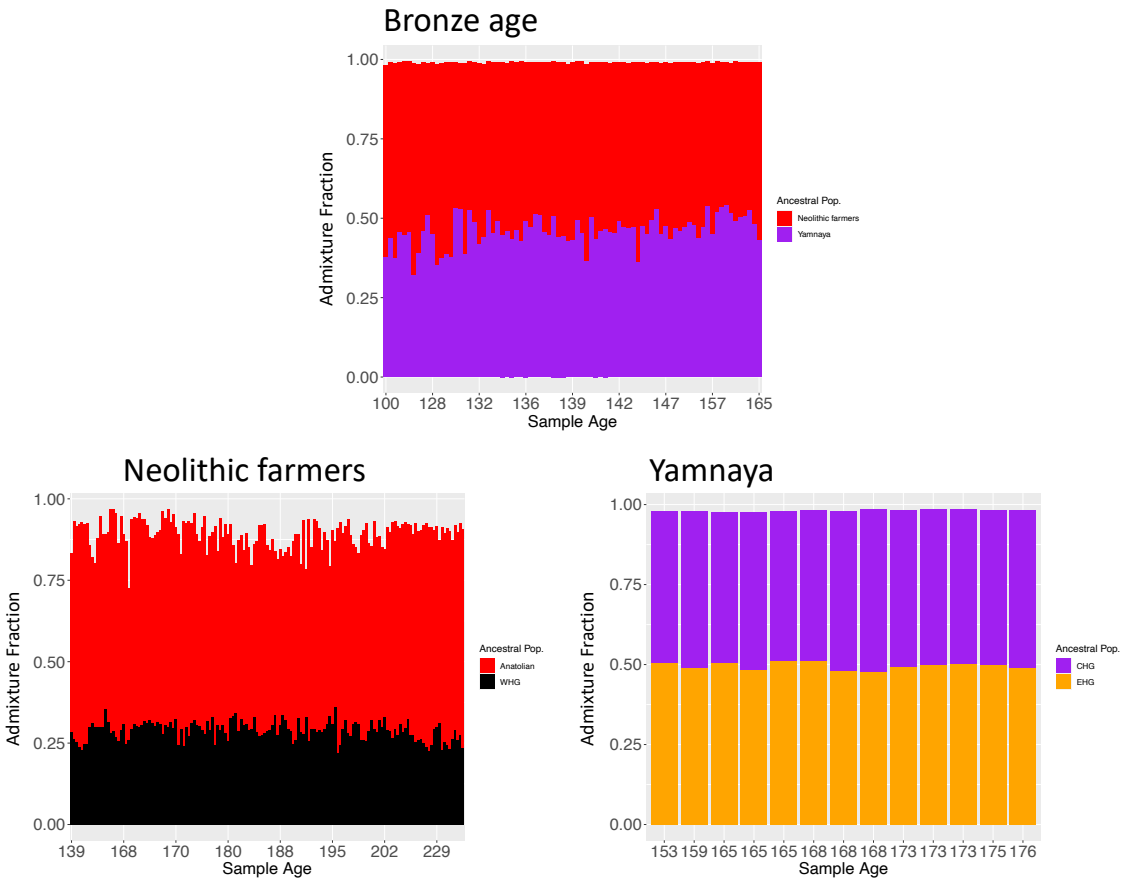


Fig. 4.4 Bar plot showing for each ancient admixed group the admixture fractions calculated by adding together the fractions of component paths if necessary from Figure 4.3. Samples are ordered along the x axis by increasing sample age.

The fraction of Yamnaya paths decreases significantly as Bronze Age samples become younger at a rate of 0.23% decrease per generation. (Pearson r = 0.45, p-value=2.8e-5). Correspondingly, the fraction of Neolithic farmer paths increases at an equal rate of 0.23% per generation (Pearson r = 0.45, p-value=2.6e-5). This result is consistent with previous work [89], implying a punctuated rapid migration of steppe herders into Europe, and subsequent

admixture with the Neolithic farmers present in Europe at similar rates across the continent. The following resurgence of Neolithic farmer ancestry represents admixture with persisting Neolithic farmer populations locally or alternatively movement of people with high Neolithic farmer ancestry around the continent. Over generations, the ancestry levels of both Yamnaya and Neolithic farmers homogenise in present day individuals (Figures 4.3 and 4.5).

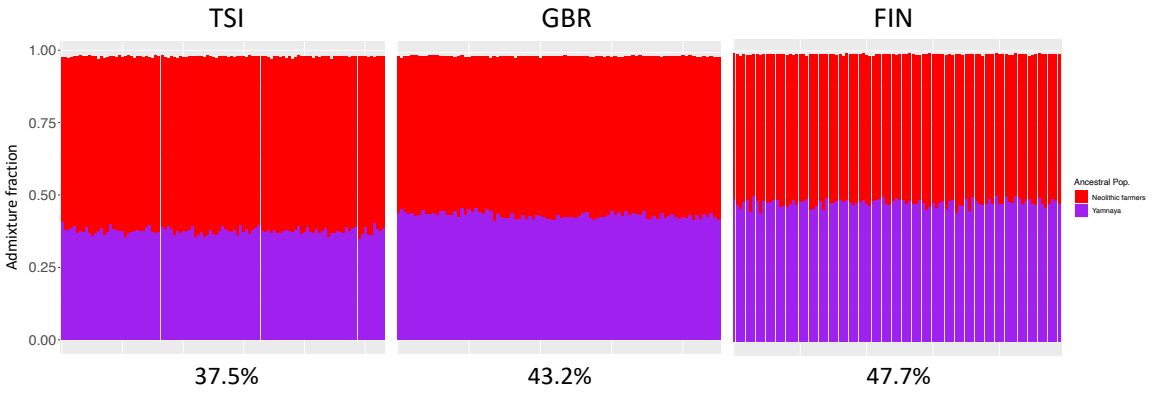


Fig. 4.5 Bar plot showing for three 1000 Genomes EUR populations the admixture fractions of Neolithic farmers and Yamnaya calculated by adding together the fractions of component paths. The mean Yamnaya path fraction is displayed under each population plot.

I calculated the Yam/Neo admixture fractions for three present day 1000 Genomes populations TSI, FIN and GBR. The results shown in Figure 4.5 show how the relative contribution of these two ancient groups varies in different modern European populations. These results fit with previously published results in that as you move further north from Italians (TSI) in southern Europe through British (GBR) in central western Europe to Finns (FIN) in northern Europe, the relative proportion of Yamnaya path ancestry increases.

4.3 Estimation of time since admixture in Europe

Next, I estimated the time of admixture for each sample in the three admixed MesoNeo populations; Neolithic farmers, Steppe Yamnaya and the Bronze Age population. The number of genomes over which this was performed was increased from my original 476 to include

samples that fall in between groups in the PCA, to explore whether these samples were the result of more recent admixture. The new total number of samples was 963, including all European 1000 Genomes samples. In Section 2.4.4 I showed that there was little decrease in accuracy for painting extra samples if I use the same set of 'reference' samples in the GNNs.

I calculated a date of admixture as the sample age plus the estimated time since admixture for each individual as described in Chapter 3, combining the standard error of the radiocarbon age estimate and the time since admixture estimate. Previously, the age of the first samples containing an ancestry are used as a proxy for the date of arrival of that ancestry to a geographical area. To test whether my inferred admixture dates are a more informative measure of the time of arrival of an ancestry, I fit spatiotemporal models of how time of admixture depends on latitude and longitude of sample archaeological sites and compared these to the same models constructed using sample age alone. I also use linear interpolation to see how both inferred admixture times and sample ages vary across the European continent.

4.3.1 Neolithic farmers

Of the 176 Neolithic farmers individuals in my subset dataset, I was able to estimate the time of admixture for 173. Figure 4.6 shows the inferred time of admixture in years against sample age. A linear regression to the points fits a model coefficient where every year younger in sample age increases the time since admixture in that sample by 0.28 years (p-value = 2.37e-06). This best fit line has a shallow gradient compared to the theoretical best fit line in which the Neolithic farmers are formed by one instantaneous admixture event happening 7,800 years ago. This suggests the migration and admixture events between Neolithic farmers and WHG was less punctuated and more of a slow process, ranging between 8,000 and 4,000 years ago.

It is important to note that admixture times are not always a proxy for movement of people. It is possible that the Anatolian Farmers moved at a faster pace into Europe, but the

subsequent integration with the local hunter gatherer populations was a slower process that continued long after migrants arrived in an area. There is evidence of persistent un-admixed hunter gatherer pockets existing long after the arrival of Anatolian migrants to the same area which is seen in parts of the World today, such as the Hadza in East Africa.

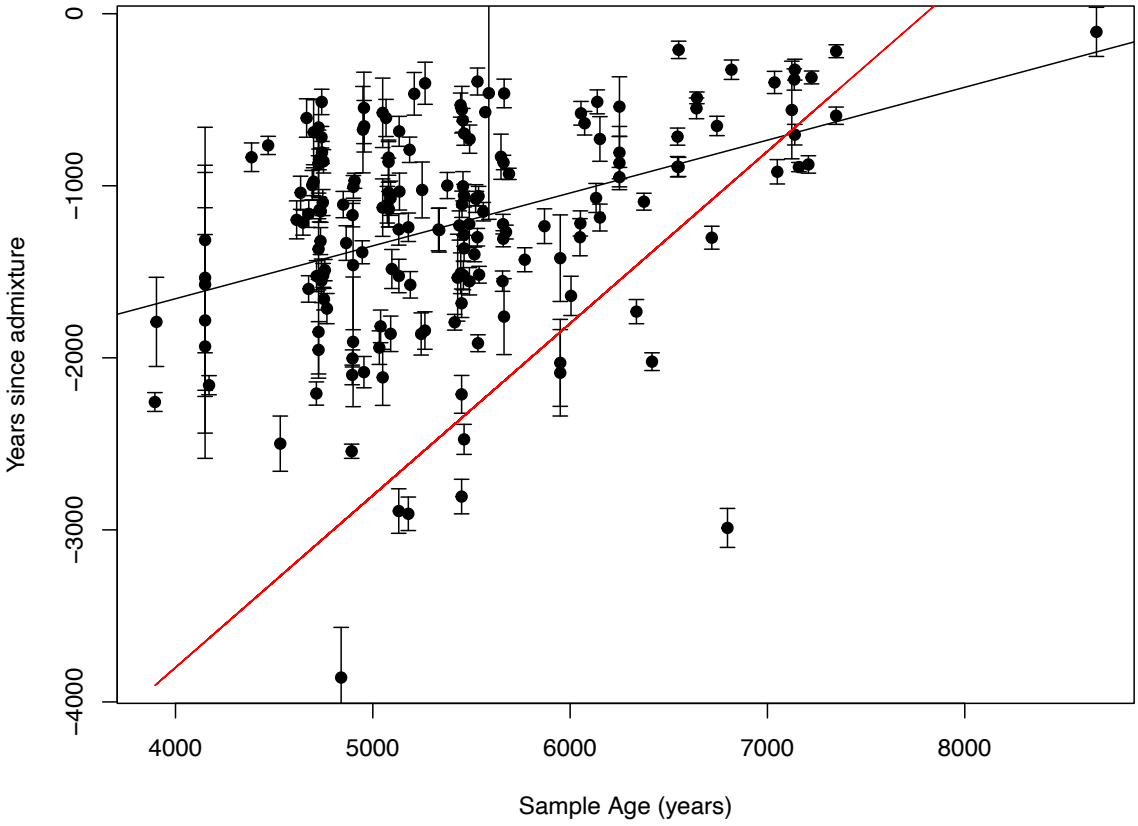


Fig. 4.6 Correlation of years since admixture and sample age over 173 Neolithic farmer samples. The best fit line is coloured black. The theoretical best fit line, in the scenario of instantaneous admixture happening at 7,800 years ago, is coloured red.

Linear models of how longitude and latitude can predict Neolithic admixture dates or sample ages alone are shown in tables 4.1. Longitude and latitude are both with highly significant predictors of admixture date. The model has a significant overall p-value of 5.947e-11 and the adjusted R-squared value (how much of the variance is explained by the model) of 23.4% is high. The predicted coefficients of longitude and latitude suggests that

Table 4.1 Summary of a linear model of Inferred Admixture time longitude + latitude for 173 Neolithic farmers samples.

Coefficients	Estimate	Std. Error	P-value
Intercept	9092.66	450.59	< 2e-16
Longitude	-25.66	4.87	4.08e-07
Latitude	-42.74	8.98	4.18e-06

Table 4.2 Summary of a linear model of Sample age longitude + latitude for 173 Neolithic farmer samples.

Coefficients	Estimate	Std. Error	P-value
Intercept	7394.92	488.34	< 2e-16
Longitude	-11.70	5.65	0.0396
latitude	-35.25	9.89	0.000471

the admixture events occurs 25 years later with every degree east and 42 years later with every degree north. In Figure 4.7 this appears as a movement north west starting in Iberia.

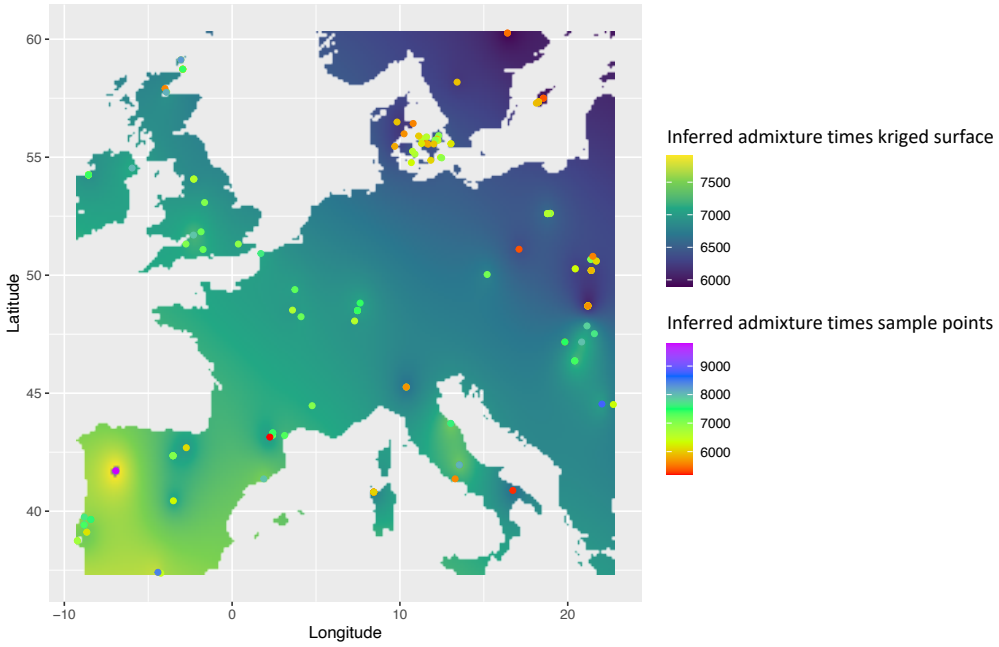


Fig. 4.7 Map of Europe with points to show the archaeological position of Neolithic farmer samples coloured by the inferred admixture time in years before present of that sample. The surface colour is of smoothed inferred admixture times in years before present created by linear interpolation between inferred sample points. This was done by kriging using the R fields package.

Two theories exist for the route that Neolithic farmers took from their origin in Anatolian into the European continent. One is a path along the northern Mediterranean coast to Iberia first before expanding north and the second is an inland route following the Danube River, north and west. My results fit with the coastal route, appearing to start in Iberia and radiating north and east from there. The MesoNeo dataset contains few samples of Neolithic farmers from along the Mediterranean coast east of Spain, due to low sample availability, or perhaps poor preservation, and so I may not be capturing the earliest admixed samples. Alternatively, the first migrants moving along the coast may not have admixed with local hunter gatherer populations until they reached Iberia. Both scenarios would be consistent with our results and in the future, more samples from critical regions could help provide clarity.

In contrast, in the model predicting sample ages alone (Table 4.2), the p-values for coefficients and the overall model, while still significant, are larger than those for inferred admixture ages and the adjusted R-square value is only 8.4%. This model suggests a more gentle south east to north west gradient of older to younger sample ages. Overall, a model involving longitude and latitude to predict inferred admixture time is more informative than one predicting sample age, in terms of understanding the impact of Anatolian farmers in Europe.

4.3.2 Bronze Age population

The migration of the Yamnaya associated ancestry into Europe is typically thought to have been a fast migration, possibly accompanied by violence [109] and substantial alteration the environment [99].

Figure 4.8 shows the inferred time since admixture in years for Bronze age individuals plotted against sample age. Constituent paths for Yamanya ancestry (paths 2 and 4) and Neolithic farmer ancestry (paths 1 and 3) were counted as one path when calculating the probability of remaining in the same path with genetic distance (Section 3.2). Of the 105

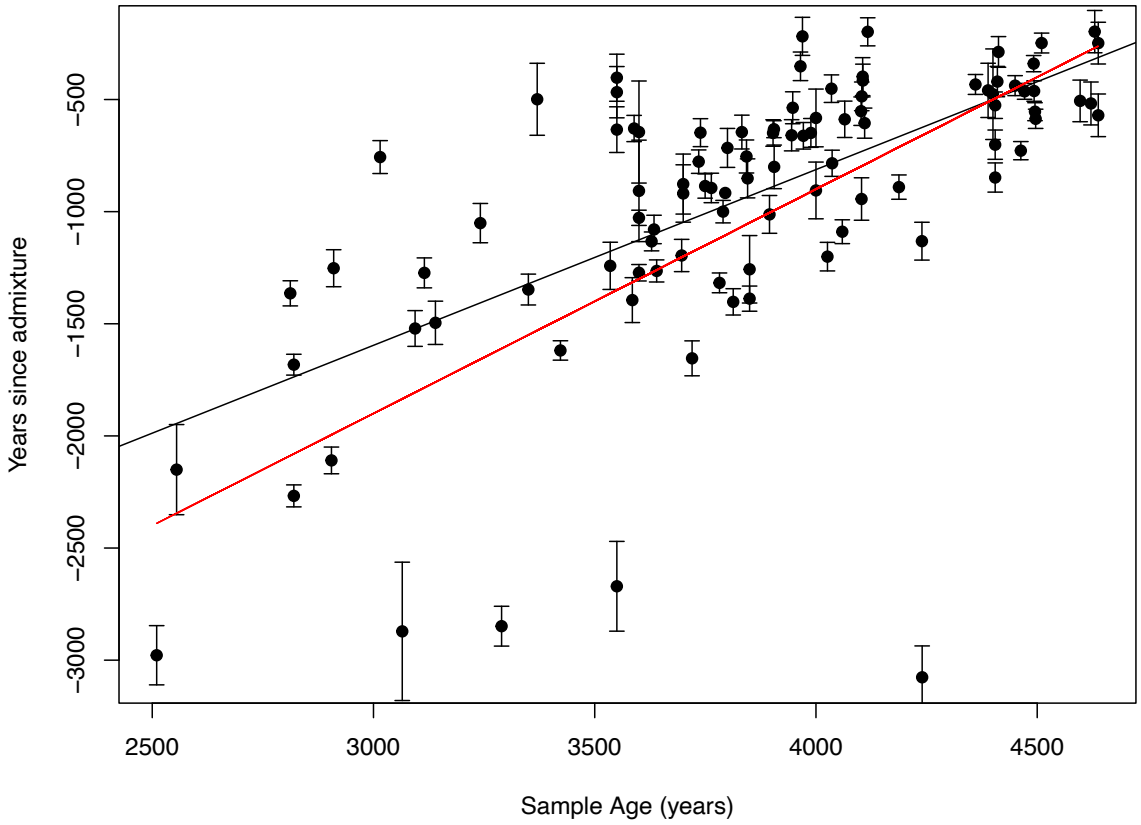


Fig. 4.8 Correlation of years since admixture and sample age over 97 Bronze Age individuals. The best fit line to the points is coloured black. The theoretical best fit line, in the scenario of instantaneous admixture happening at 4,900 years ago, is coloured red.

Bronze Age individuals in my MesoNeo subset, exponential curves could be fit to 97. A linear regression model implies that every year younger a sample's archaeological age, the time since admixture in that sample increases by 0.78 years (p-value = 1.536e-11). A coefficient value of 1 would mean near contemporaneous admixture over the whole continent, so as shown by the theoretical best fit line in red, such a high coefficient found for the Bronze Age admixture is a consistent with a rapid migration of the Yamnaya across the continent between 4000 and 5500 years ago, with admixture events quickly following.

Likewise, linear models incorporating longitude and latitude do not find longitude or latitude as a significant coefficients. This is again evidence of rapid movement of Yamnaya into Europe from the steppe with admixture following soon after the migration started, and

Table 4.3 Summary of a linear model of Inferred admixture time longitude + latitude for 97 Bronze Age samples.

Coefficients	Estimate	Std. Error	P-value
Intercept	3894.05	443.23	$7.57e-14$
Longitude	2.98	5.77	0.607
latitude	16.15	9.08	0.0785

then the admixture continuing with admixed individuals, producing no detectable variance in inferred admixture time with geography. The variance in the data explained by a model of longitude and latitude alone is small at only 4.9%. Other factors not accounted for here must explain the variance of admixture times across the continent better, unlike the Neolithic Farmers whose admixture times are substantially explained by geography (Table 4.3).

While the R-squared adjusted for the inferred admixture times is small, models using sample ages alone explain the data even less well at only 2.1%.

4.3.3 Yamnaya population

The MesoNeo dataset provides more Yamnaya related individuals than were previously available, adding more data points for dating the admixture event between EHG and CHG groups. While this is a substantially boost in sample to size, there are still relatively few samples and therefore not enough power to detect any trends in admixture time with sample age or geography. This is compounded by the samples being from disparate geographic locations in Ukraine, Poland, Kazakhstan and across Russia.

Despite this, I can date the samples I have and draw some conclusions. The genetic formation of the Yamnaya population is not well understood. Culturally they appear in the archaeological records 3300-2600 BC [85]. I dated the genetic formation of the Yamnaya from 16 individuals of the 17 present in my MesoNeo subset (Table 4.4). Most estimated admixture dates are 5000-6000 years before present, a millennium or so before their believed cultural formation. Three outlier samples have inferred admixture times of more than 7000

Table 4.4 Table of the results from inferring admixture time in 16 Yamnaya individuals.

Sample name	Sample age	Inferred Admixture Times (years BP)	Standard error (years)	Country
Latvia_LN1	4832	5651	123	Latvia
MJ-06	4630	7134	206	Ukraine
MJ-09	4286	5016	75	Ukraine
NEO175	4606	5342	93	Russia
NEO212	7390	7880	103	Russia
poz81	4705	5535	80	Poland
RISE240	4706	5554	73	Russia
RISE509	4732	6961	66	Russia
RISE511	4744	5787	73	Russia
RISE546	4850	7564	238	Russia
RISE547	4710	6645	76	Russia
RISE548	4850	6558	203	Russia
RISE550	4934	5589	181	Russia
RISE552	4446	5893	190	Russia
RISE555	4627	5370	93	Russia
Yamnaya	4902	5761	63	Kazakhstan

years ago. These all have large standard errors on the estimates and so are less reliable. However, even allowing for larger confidence intervals based on higher standard errors, their admixture times are placed well before 6,000 years before present. NEO212 is the oldest of these samples and falls between the other Yamnaya individuals and Eastern hunter gatherers in PC space which suggests it might be a relatively early hybrid.

These admixture dates are much older than the archaeological appearance of the Yamnaya and so, subject to technical artefacts they suggest extended prior genetic contact between EHG and CHG well before the emergence of the Yamnaya culture, potentially early as 7000 years ago.

4.4 Conclusion

I have shown from the Neolithic Farmers and Bronze Age populations in Europe that my estimates of admixture time are more informative than using the sample ages alone for assessing spatiotemporal dynamics of these population movements and mixing. While admixture times may not be a proxy for movements of people, it is an interesting metric in itself, and one that varied significantly between the different major admixture events involved in the formation of the modern European population.

Given that the MesoNeo genomes have been imputed, there is possible phasing errors. In the local ancestry painting phasing error will appear as switches of ancestry across homologous chromosomes. As a result, phase errors will cause ancestral tracts to appear smaller than in truth and therefore produce admixture times that are older than in truth. It is possible that my calculated admixtures times have overestimated absolute times as I have not accounted for phase error in my method. However, my conclusions drawn from the relative admixture times between samples are still valid if I assume phase errors to happen at a constant rate between samples.

A strategy to account for phase errors is to use both haplotypes from each admixed sample when measuring the ancestry correlation with genetic distance and not treat the two chromosomes from the same sample independently. Switches in ancestry across the homologous chromosomes will be accounted by measuring over both haplotypes.

Chapter 5

Using coalescence events

Contents

A coalescence is the event where two haplotypes share a common ancestor in the past, at which point two lineages coalesce to become one lineage. The rate that coalescences occur at over a time period in the past between a set of sampled lineages is related to demographic

and evolutionary processes acting within that population during that time period (Section 1.3).

Knowledge of coalescence rates through time enables inference of the parameters of these processes. Specifically, coalescence rates are inversely proportional to effective population size. When the population was small for some time period in the past, the number of coalescences occurring in that time period will be high. Similarly, alleles under selection during a past time period will have a disproportionately high coalescence rate compared to the genome-wide average, during that time period.

While both a population bottleneck and positive selection create an excess of coalescences, the difference in inference is that population size changes act on the whole genome, all trees in the sequence, but selection will only alter the trees covering the selected site and the trees surrounding, with decreasing extent.

The above concepts can be applied to tree sequences, combined with local ancestry paintings, to infer parameters of evolutionary processes conditioned on haplotype ancestral background. The advantage of the concept of local ancestry as a path is that population size changes and selection signals can be inferred along paths in various populations that, with a panmictic model, might otherwise be diluted. Therefore, with paintings for both the modern 1000 Genomes samples and ancient European samples combined with the tree sequences, I propose a genealogy-based method that leverages both ancient genomes and local ancestry estimates to infer population size changes and selection along paths of ancestry.

5.1 Inference of effective population size

5.1.1 Existing approaches

Existing methods that aim to infer the parameters of complex demographic models, including population size estimates, can be broadly divided into two approaches; Site frequency spectrum-based and haplotype-based.

The site frequency spectrum (SFS) is a histogram of the number of alleles in a sampled population that have corresponding derived frequencies. The expected SFS for a panmictic population of constant size under no natural selection can be calculated using both coalescent theory [55] and diffusion theory [54], where the expected proportion of sites with allele frequency i is proportional to $\frac{1}{i}$. If the SFS deviates from this expectation it indicates that the population history deviates from the assumed panmixia of constant size and no selection. The shape is sensitive to the population history including population size change, population structure, migration and selection. For example a bottleneck in population size will produce an excess of low frequency variants and so skew the SFS. Selection will affect the SFS around the selected site only, while population size changes will change the combined SFS across all sites.

Many methods have been developed that aim to infer the population history from the observed SFS. The observed SFS can be compared to the expected SFS calculated under a given model of population history and the goodness of fit of that model to the data evaluated. A maximum likelihood set of parameters can be found by searching through different parameters until the best fitting set is found.

Early methods used full likelihood approaches [9, 56] or approximate the posterior of the SFS under a model of population history using techniques such as MCMC [42, 41]. The problem with full likelihood approaches is that the likelihood function is very difficult to evaluate or is intractable and methods that approximate the posterior become very slow

when complexity of demographic history increases. Additionally, most are limited to non-recombining sequences.

To combat this, "likelihood-free" Approximate Bayesian Computation (ABC) methods were introduced [8]. These methods approximate the likelihood function by simulations and comparing the outcome of those simulations to the observed data. The parameter set used to simulate is discarded if the 'distance' between the SFS of the simulated data and the observed data is greater than a threshold amount. Simulating the SFS is possible using the coalescent or diffusion theory and so ABC methods have allowed inference of more complex demographic methods [83, 21, 105].

However, ABC can be very slow as computation time increases with the number of loci. Composite likelihood methods have computation times that are independent of loci number. Notably $\delta\alpha\delta i$ approximates the joint multi-population SFS with a Wright-Fisher diffusion approach and compares it to the observed SFS data with a composite likelihood function [38]. Similarly fastsimcoal2 uses coalescent simulations to estimate the SFS under demographic models and compares simulations to the data, calculating a composite likelihood of the model to optimise the parameters [23]. Unlike the $\delta\alpha\delta i$ framework which can only model 3 populations, fastsimcoal2 can model handle arbitrary numbers of populations and may be more robust than $\delta\alpha\delta i$.

While all aforementioned methods are limited to demographies that are tree-like, momi2 can model the join SFS for non tree-like demographic histories involving admixture events by defining demographies on general Directed Acyclic Graphs (DAGs). The expected joint SFS is computed using the continuous time Moran model which shrinks the state space compared to the coalescent model [51].

However, SFS-based methods are limited to independent sites and do not make use of information contains in patterns of linkage between sites [108]. Additionally, there is debate around whether demographic history is statistically identifiable from the SFS [86, 10].

The demographic history influences coalescence times of haplotypes which in turn influences the mutation patterns. Haplotype-based methods aim to infer the coalescence times from the patterns of mutations and in turn infer the demographic history and so leverage the full information from linked SNPs. Recombination alters the underlying genealogies and so it is necessary to average over the possible recombination histories that may have shaped the observed haplotypes [108].

Modelling the coalescent with recombination is difficult because long-range correlations between sites makes the state space of possible genealogies very large. Building on the work of Wiuf and Hein [125] who formulated the coalescent process as along the genome rather than backwards in time, McVean and Cardin formulated the sequentially Markovian coalescent (SMC) [80], to approximate the coalescent with recombination. SMC circumvents long-range correlations between genealogies by making the current genealogy depend only on the immediately previous genealogy. This allowed the Markovian class of algorithms to be used along chromosomes to estimate population parameters. Additionally, Marjoram and Wall introduced SMC' which allows for recombination back into the same lineage at a different coalescence time, a so called 'invisible' recombination that improves the closeness of the approximation to the full coalescent model [71].

With SMC and SMC' in hand, hidden Markov models (HMMs) can be applied across chromosome, allowing the unknown genealogies to be integrated out by modelling them as the hidden states and the genetic sequence data as the observed states. Through Baum-Welch the coalescence rates can be calculated and from those the effective population size at discrete time epochs is ascertained. Pairwise sequentially Markovian coalescent (PSMC) applies this approach to a pair of haplotypes or a single diploid genome to estimate the coalescent times between them at each site along the genome [64]. Subsequently, multiple sequentially Markovian coalescent (MSMC) was developed that estimates the time to the first coalescent between multiple phased haplotypes and which pair of haplotypes are involved [107]. This

allows for estimates of population size in more recent time. SMC++ is a more recent method in the same family as PSMC and MSMC but scales to many sample haplotypes and achieves a substantial speed-up by 'skipping over' stretches of non-segregating sites [122].

These described methods can be used to infer changes in population structure. MSMC and SMC++ are more suited to this task as they can use multiple sample sequences from separate populations. MSMC calculates cross-coalescence rates between separate populations throughout the past, indicating when divergence occurred and cessation of gene flow. However, the dual ability to also infer population structure from coalescence rates as well as population size poses a problem: How to distinguish between a structured population of constant size or a panmictic population of changing size when only representative samples of one population are available? These methods have come under criticism for this reason [78]. An extension of MSMC called MSMC-IM attempts to tackle this problem. It fits an isolation-migration model to coalescence rates estimated by MCMC to infer piece-wise migration rates between populations and piece-wise population size histories within populations [124].

Unlike the methods presented in this section that treat the underlying genealogy of samples as an unobserved variable, RELATE directly estimates the genealogies of a set of samples across the genome. The shapes of trees are altered by evolutionary processes and therefore they reflect the demographic history of a sample (Section 1.3). RELATE has the advantage over PSMC and MSMC of being scalable to thousands of sample haplotypes. In its current implementation RELATE has an MCMC method for re-estimating branch lengths under a model of variable population size but assumes panmixia (Section 1.5). However, since RELATE infers the genealogies themselves, its is possible to tease apart population structure from population size change. I aim to estimate the changing population sizes on separate paths through time by placing coalescence events along paths and from the counts of coalescences in each path, calculate a maximum likelihood population size on each path at a given time.

5.1.2 A method for inferring population size along paths

Given that for each tree I have path assignments for leaf nodes, I traverse up the tree, assigning path labels to internal nodes. All internal nodes can be reached when traversing from multiple different leaf nodes. Whenever there is a discrepancy in the path assigned for an internal nodes when traversing from different painted leaf nodes, I assign it the path of the leaf node which obtained the larger maximum softmax value, output in the neural network's final layer. With all nodes in all trees assigned to a path, the diploid population size (N) for a path can then, in principle, be estimated from the number of coalescences k occurring when n lineages are present in the path. Across the whole tree sequence, I record in a 4-dimensional matrix $C^{(p,t)}_{(n,k)}$ the count of trees where there are k coalescences between n lineages on path p at time t in units of generations.

The maximum likelihood population size estimate can be calculated along each path in each generation from these counts. The number of coalescences per generation can be modelled as a Poisson distribution with mean,

$$\lambda = \frac{n(n-1)}{4N}, \tag{5.1}$$

and therefore, the probability of seeing k coalescences in a generation is

$$P(\text{Coal = k}) = \frac{\frac{n(n-1)}{4N}^{k} e^{[-\frac{n(n-1)}{4N}]}}{k!}. \tag{5.2}$$

Taking a log of equation 5.2,

$$log\,(P(\text{Coal = k})) = k \log\left(\frac{n(n-1)}{4N}\right) - \frac{n(n-1)}{4N} - \log k!. \tag{5.3}$$

To find the maximum likelihood population size N, I need to find the value of N when the derivative with respect to N is set to 0. In each generation and on each path, I sum over all n,k pairs, multiplied by the count of each n,k pair $C_{(n,k)}$,

$$log\left(P(\text{Data} \mid \text{N})\right) = \sum_n \sum_k \left[k \log\left(\frac{n(n-1)}{4N}\right) - \frac{n(n-1)}{4N} - \log k! \right] C_{(n,k)}, \tag{5.4}$$

$$log\left(P(\text{Data} \mid \text{N})\right) = \sum_n \log\left(\frac{n(n-1)}{4N}\right) \sum_k k C_{(n,k)} - \sum_n \frac{n(n-1)}{4N} \sum_k C_{(n,k)} - \sum_k \sum_n C_{(n,k)} \log k!, \tag{5.5}$$

$$log\left(P(\text{Data} \mid \text{N})\right) = \sum_n \left(\log n(n-1) - \log 4 - \log N\right) \sum_k k C_{(n,k)} - \sum_n \frac{n(n-1)}{4N} \sum_k C_{(n,k)} - \sum_k \sum_n C_{(n,k)} \log k!. \tag{5.6}$$

Taking the derivative with respect to N,

$$\frac{\partial log(P)}{\partial N} = -\frac{1}{N} \sum_n \sum_k k C_{(n,k)} + \frac{1}{N^2} \sum_n \frac{n(n-1)}{4} \sum_k C_{(n,k)}, \tag{5.7}$$

and setting this equal to 0,

$$\frac{\partial log(P)}{\partial N} = -\frac{1}{N} \sum_n \sum_k k C_{(n,k)} + \frac{1}{N^2} \sum_n \frac{n(n-1)}{4} \sum_k C_{(n,k)} = 0, \tag{5.8}$$

$$\frac{1}{N^2} \sum_n \frac{n(n-1)}{4} \sum_k C_{(n,k)} = \frac{1}{N} \sum_n \sum_k k C_{(n,k)}, \tag{5.9}$$

$$N = \frac{\sum_n \frac{n(n-1)}{4} \sum_k C_{(n,k)}}{\sum_n \sum_k k C_{(n,k)}}, \tag{5.10}$$

$$\hat{N} = \frac{\sum_n \frac{n(n-1)}{4} \sum_k C_{(n,k)}}{\text{Total number of coalescences}}. \tag{5.11}$$

The above derivation shows that the maximum likelihood population size is a weighted average of the counts of n,k pairs. With the 4-dimensional array of counts $C_{(n,k)}^{(p,t)}$ I can calculate a maximum likelihood population size along each path, in each generation. When two paths are joined in a population history, the number of lineages is the sum of the number in each path and coalescences can occur between any of those lineages.

5.1.3 Testing on simulated data

Using nine simulated tree sequences from the model of European population history and with all leaf nodes painted with the true paths, I assign internal nodes to paths by traversal and apply the above procedure for population size estimate along paths. I recover estimates close to the correct population sizes with some noise, demonstrating that my approach is sound given perfect data (Figure 5.2).

I then applied the procedure to the RELATE tree sequences that were inferred from simulated data with all internal nodes assigned a path label by traversal. Figure 5.3 shows that the coalescence time distribution along each path is similar to that of the simulated tree sequence but not an exact match. Inferred coalescence times are slightly lagging behind the simulated coalescence times and so there is a small deficit in coalescences early in time and a small surplus later in time.

Correspondingly, Figure 5.2 shows that in some parts of the structure, the correct population size can be recovered in the RELATE trees. At the start of paths, population sizes are overestimated, meaning there are too few coalescence events for the number of lineages present. Over admixture events, changes in population size estimates in RELATE are sloped rather than stepwise changes as in the true demography and estimates from simulated tree sequences. During the periods of lowest population size in all paths, population size estimates from RELATE tree sequences are close to the truth. In the deeper time periods, older than the population splits, the population sizes are underestimated, meaning there are too many coalescence events for the number of lineages present.

From these results, it is clear that population size estimation is very sensitive to small deviations from the correct path assignment and coalescence time. While I have shown that the classifier has good accuracy as a percentage of testing GNNs and the coalescence times are overall similar to the true times, mistakes in classification are multiplied two ways. The first, a path error is effective as both wrongly placed along one path and missing from its true

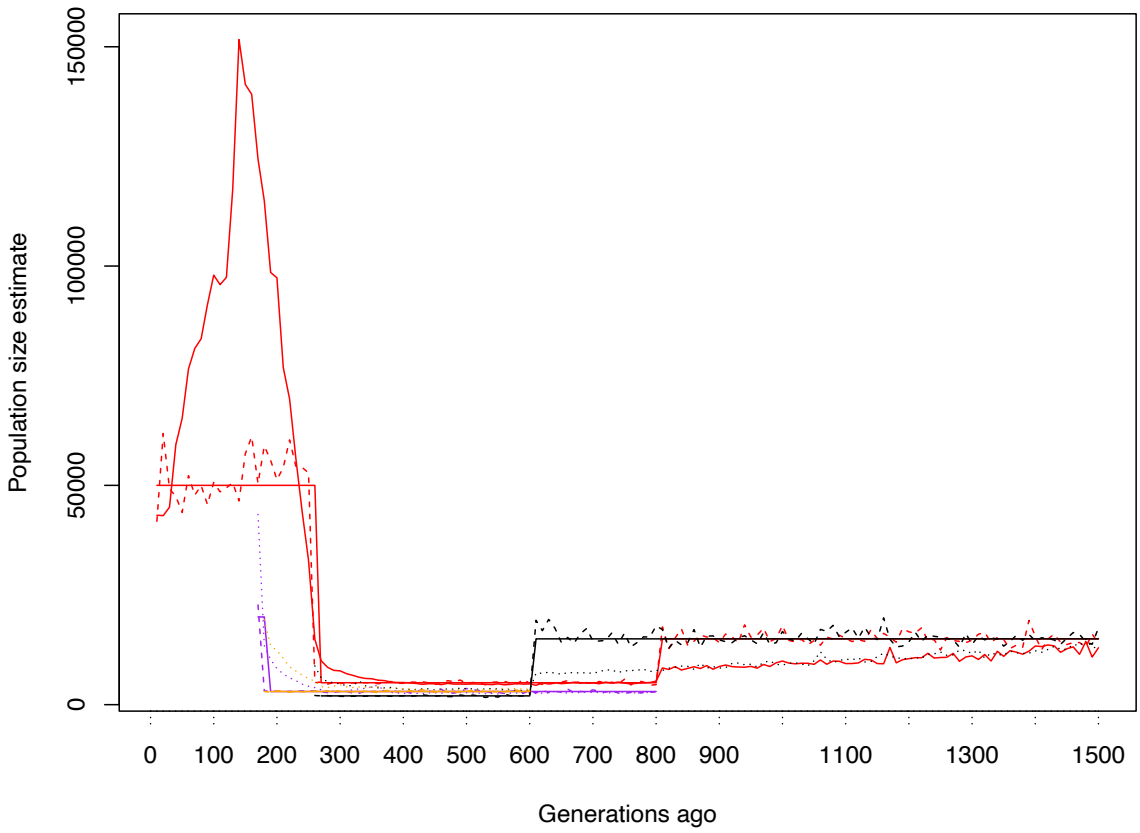


Fig. 5.1 The effective population size, calculated as in Subsection 5.1.2, over time in generations. The solid lines show the true coalescence count, dashed lines show the estimates from simulated tree sequences and the dotted lines show the estimates from RELATE inferred trees whose coalescences have been classified to paths. Lines plot an average population size estimate across 10 generations.

path and second the incorrect removal of lineages from a path is carried up the remaining part of the tree in the parameter n.

It should be noted that when the sample size is not sufficient to create any coalescence events in a given generation across all trees, the denominator is equal to zero and equation 5.8 is unbounded. This can be improved with samples that are distributed throughout time, resulting in coalescence events that are more evenly distributed through the demography and allowing an estimate of population size in more generations as well as improving power for

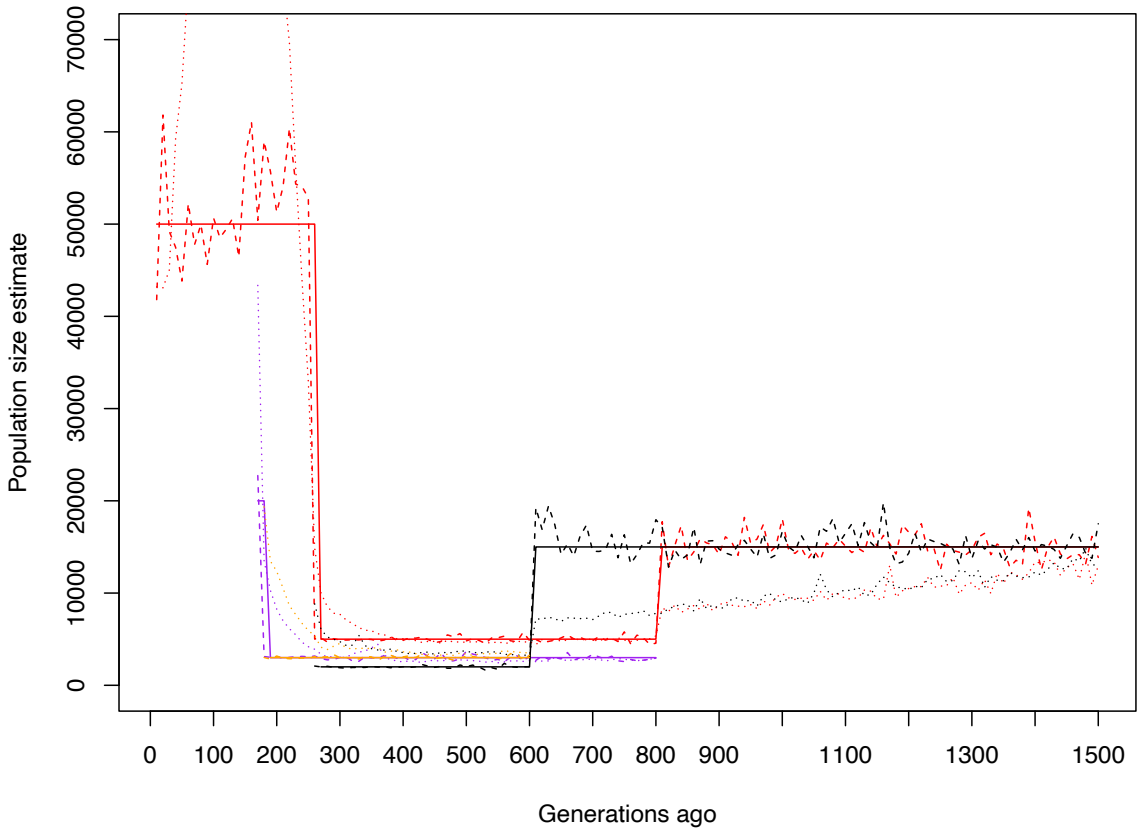


Fig. 5.2 The effective population size plotted up to 70,000, calculated as in Subsection 5.1.2, over time in generations. The solid lines show the true coalescence count, dashed lines show the estimates from simulated tree sequences and the dotted lines show the estimates from RELATE inferred trees whose coalescences have been classified to paths. Lines plot an average population size estimate across 10 generations.

inference deeper in time. This again highlights the utility of ancient samples. Likewise, more samples overall create more lineages and coalescence events.

RELATE assumes no population structure, yet it has been shown that divergence times and separation histories can be determined from coalescence patterns [117, 116]. With knowledge of the structure of a population like that of Europeans, a future endeavour is to incorporate a structure into the RELATE inference process which could be successful at improving population size estimates along paths (Section 6.4).

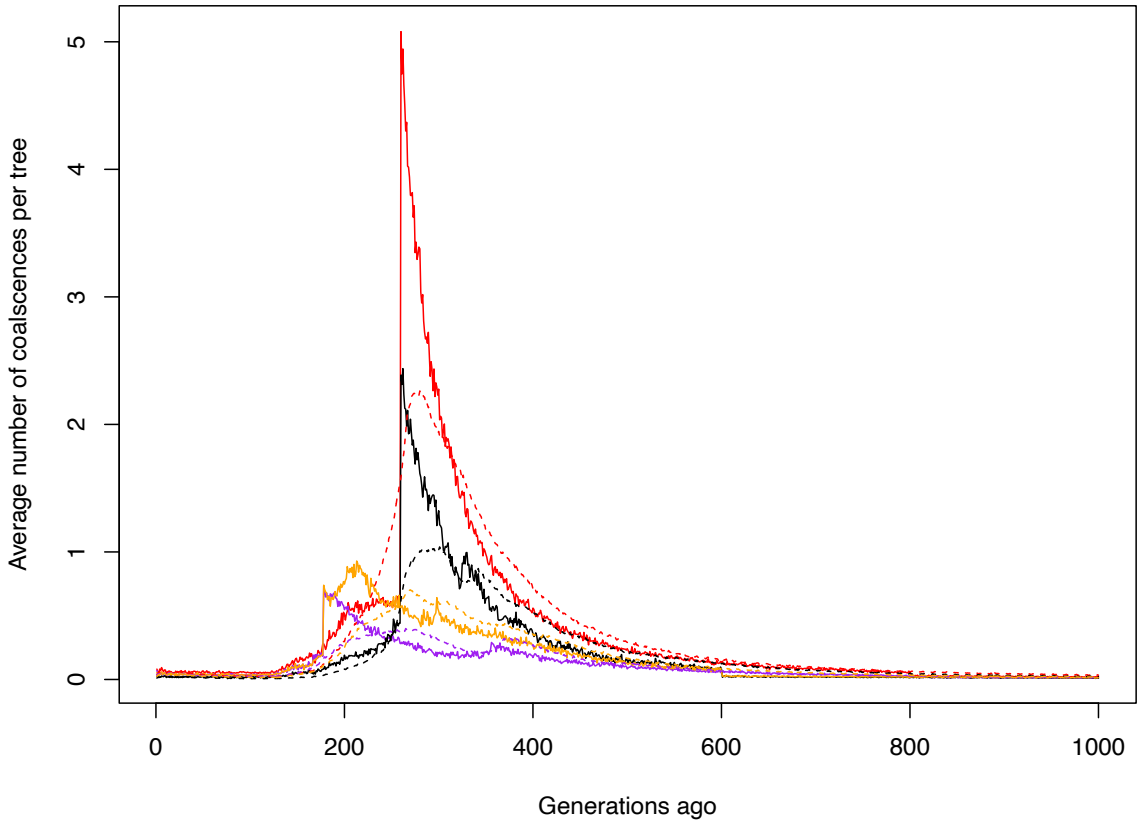


Fig. 5.3 The average number of coalescences per tree occurring along each path (coloured) over time in generations. The solid lines show the true simulated coalescence count and the dashed lines show the count from RELATE inferred trees whose coalescences have been classified to paths.

Lastly, a lucrative strategy could be to subset the trees over which inference is performed to those that are high-quality. With evenly distributed samples through time, there should be enough trees from which a subset would still have enough coalescence events to maintain power for inference.

5.2 Signals of positive selection along paths

Many examples of selection on European alleles have been demonstrated. In particular, selection signals on the lactase persistence allele have been found using multiple methods

[4, 75, 93]. Finding signals of positive selection on relevant genes in the MesoNeo dataset is of particular interest since the dataset covers a big transition in lifestyle from hunter gathering to farming. It is hypothesized that such a change would result in selective pressures on features relating to diet and metabolism and may therefore be relevant to certain chronic diseases today. In this section I will first give a review of existing methods for detecting positive selection from sequences followed by an explanation of a new method for detecting selection from painted tree sequences.

5.2.1 Existing approaches

When an allele is advantageous to individuals who carry it in a population, its frequency in the population will tend to increase in each generation due to natural selection. Eventually the allele might reach fixation in what is termed a selective sweep. Selection therefore alters the shape of the underlying genealogy at selected sites. For example, positive selection will produces a 'star-like' pattern of rapidly multiplying derived lineages. Methods to detect positive selection are therefore very similar to those presented in Section 5.1.1 for detecting population size changes in that they aim to infer selection and associated parameters from the signatures in the genomes that are caused by genealogies differing from expectation under neutrality.

One group of selection tests is based on analysing the rate of substitutions. Synonymous substitutions are mutations where one base is substituted for another but does not change the amino acid sequence. At sites with four-fold degeneracy, where every possible mutation is synonymous, it is normally assumed that synonymous mutations are neutral and therefore the substitution rate at these sites can be assumed to be the neutral substitution rate. Positive selection will increase the rate of substitution compared to the neutral rate. Therefore, comparing the ratio of the number of nonsynonymous substitutions at nonsynonymous sites to the number of synonymous substitutions at synonymous sites indicates whether a locus is

evolving neutrally or not. An increased relative nonsynonymous substitution rate indicates positive selection and a decreased relative rate indicates purifying selection [128, 30, 48].

The McDonald Kreitman [79] test also utilises synonymous vs nonsynonymous substitution rate. It compares the divergence to diversity ratio which at synonymous sites is assumed to represent the neutral ratio. Nonsynonymous sites with a higher divergence to diversity than the neutral rate indicates that a locus is evolving under positive selection as positive selection will fix alleles more rapidly than neutral alleles. However, these tests tend to only good for long term sustained selection [30] and there is now also evidence that synonymous mutations are not always neutral [17].

A second group of selection tests are based on the frequency of a selected alleles. After a selective sweep, the site frequency spectrum (SFS) is depleted of intermediate-frequency alleles, which the summary statistic Tajima's D captures:

$$D = \widehat{\theta_\pi} - \widehat{\theta_w},$$

where $\widehat{\theta_\pi}$ is the average number of pairwise differences between individuals and $\widehat{\theta_w}$ is Watterson's estimator of θ. When evolving neutrally have an expectation of $4N_e\mu$ and therefore E[D] = 0. Under positive selection E[D] < 0 [120]. Several extensions to Tajima's D have been developed such as Wu's H [25] that is more sensitive to recent selective sweeps.

A problem with allele frequency or substitution-based methods is that they rely on an expectation of neutrality which is estimated under assumptions of demography. As described in Section 4.1, population structure and changes in population size can produce the same signatures in the genome as selection, so these methods can easily be confounded by violations of demographic assumptions [88]. Moreover, in the same way as population structure estimation, these methods are limited to unlinked, independent SNPs and do not use the full information from patterns of SNP linkage in haplotypes.

During a sweep, sites that are in the region of the focal site undergoing selection will also increase in frequency as they occur on the same haplotype. Linkage decreases at sites further away from the selected site due to recombination, so the effect of nearby linked alleles increasing in frequency with the selected allele decreases with genetic distance. This process of 'genetic hitchhiking' and was first recognised by Maynard Smith and Haigh and results in a valley of genetic diversity around the selected site [114]. Very close neutral sites may also be fixed as they increase in frequency along with the selected variant. Moving either direction along the chromosome, away from the selected site there is a steady increase in diversity. Genetic hitchhiking therefore produces increase linkage disequilibrium (LD) around selected SNPs.

A more recent class of methods exploits the haplotype structure of increased LD and steadily increasing diversity around selected SNPs. The extended haplotype homozygosity (EHH) statistic [104] measures the probability that two haplotypes are identical up to a distance x from a focal SNP. This statistic converges to 0 at a sufficient distance from the focal SNP as haplotypes structure breaks down due to recombination. The rate at which the EHH decays to 0 indicates whether there has been selection. Similarly, the iHS statistic [123] is designed to detect ongoing selection and uses the integral of the EHH statistic once haplotypes are partitioned by ancestral or derived state of the focal SNP. In order to detect soft sweeps, which haplotype homozygosity has less power to do so, the H_{12} score [31] uses the frequency of haplotypes from a set of haplotypes surrounding a focal SNP to infer selection, based on the principle that under sweep conditions, the several most frequent should dominate the haplotype frequency distribution.

Lastly, coalescent based methods can be very powerful to detect selection while also allowing an estimate of the timing of selection in the past. The problem with coalescent methods to infer selection is that the space of possible genealogies is very large and so integrating over all possible genealogies is challenging. CLUES [118] is a method based on

ARG structures such as those produced by RELATE which uses an HMM and importance sampling to fully integrate out the allele frequency trajectory and in doing so estimate selection coefficient parameters.

The above methods are concerned with finding signals of past selection from modern genomes alone. With ancient genomes there is a direct sample of alleles from the past and so the allele frequency of certain SNPs can be estimated. Allele frequency trajectories can then be inferred, and selection is implied where frequencies rapidly increase [75, 93].

5.2.2 A method for identifying signals of positive selection along paths

Here I develop a method for detecting signals of positive selection in the past on paths of ancestry using my local ancestry assignments of RELATE tree sequences. Variants that increase the fitness of individuals can create a 'star-like' pattern in the genealogy of the marginal tree that covers the selected variant. Positive selection will therefore produce more coalescence events between carrier chromosome in the tree of the selected locus at the time of selection than expected for the chromosome. I use an estimate of population size calculated from across all trees in a given generation in the past to correct for changes in demography altering the genealogies.

Using the painted leaf nodes, I assign internal nodes in all trees to a paths going by traversal towards the root. Taking the marginal tree covering the potential selected site and considering only lineages that are derived, I record the number of coalescences in each generation back in time along paths. I model the number of coalescences for a neutrally evolving site, per generation per path, as a Poisson distribution with rate parameter given by,

$$\lambda = \frac{n(n-1)}{4N_e}, \tag{5.12}$$

where n is the number of derived lineages between which coalescences are occurring in that generation and path, and N_e is the effective population size for that generation and path,

calculated as above, from across all trees in the sequence. I obtain the p-value of observing a coalescence count in the selected marginal tree as the probability of sampling that number of coalescences from the Poisson PDF, given the number of lineages present and effective population size. Variants that have been selected at some time on a path will have p-values above the significance threshold, when corrected for multiple testing.

By modelling the coalescence count at each time-point in the selected tree as a sample of the concurrent chromosome-wide distribution in the same path, I correct for reductions in population size which will also increase coalescence rates but over the whole chromosome. Likewise, I also correct for the reduction in power to detect positive selection as the number of lineages decreases going back in time by sampling from a Poisson distribution with a rate parameter that is dependent on n. Even if the absolute population size calculated is not correct, the signal of a relative excess of coalescences is still valid.

5.3 Application to MesoNeo genomes

In this Chapter I have described methods for inferring both population size changes and selection along ancestry paths. Using RELATE I can directly observe plausible genealogies, which most methods for estimating both demographic history and selection must model as an unobserved variable. This means I can derive relatively simple expressions for calculating maximum-likelihood population sizes and inferring selection, without the need for highly expensive or intractable likelihood calculations that integrate over all possible genealogies or approximate the integral.

In this Section, I present results from applying the above methods for population size estimation and selection analysis to the MesoNeo genomes.

5.3.1 Population size estimates

I applied the above method for calculating population sizes to my MesoNeo subset containing 952 haplotypes (Table 2.1). Figure 5.4 shows the population size estimates for each path. There is large spike in population size at the start of the Bronze Age, once all paths have joined. This could represent an expanding population at that time followed by a decrease in the effective population size as people started to move more across the continent and admix, resulting in reduced the genetic diversity. However, a similar spike is seen when testing the method on RELATE trees inferred from simulated data with a constant population size of 50,000 at the start of the Bronze Age (Figure 5.1), so this spike could be a technical artefact.

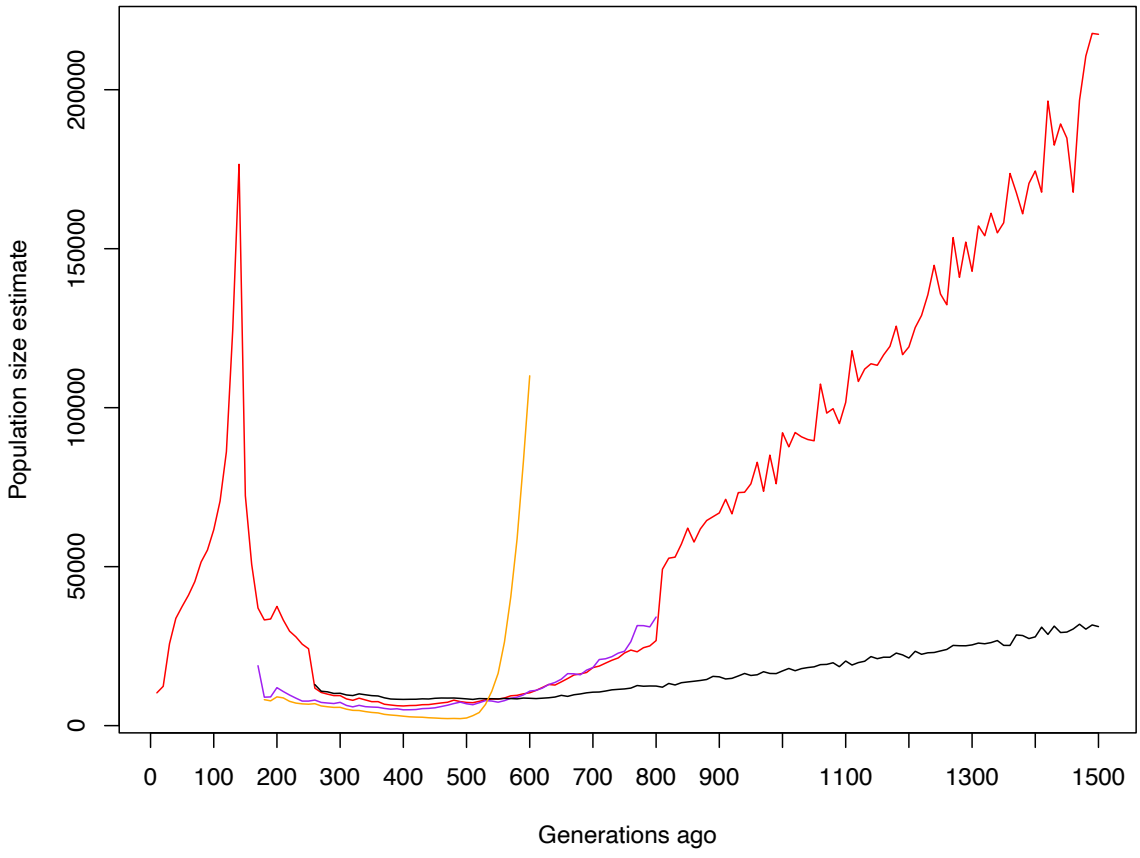


Fig. 5.4 Effective population size estimates for the MesoNeo genomes along paths 1-4. Paths are coloured as in the model of European population structure in Figure 2.3.

During the time period when all paths are separate, from 260 to 550 generations ago, all paths display low population sizes, with the EHG on path 4 reaching below 3,000 (Figure 5.5). This mirrors the low population sizes in the model of European population structure, consistent with the PCA and F_{st} results, and is likely a genuine signal of a population bottleneck in the four path populations induced by isolation in ice age refugia.

Deeper in time the population sizes increase, probably as a result of structure in these older populations that is not modeled.

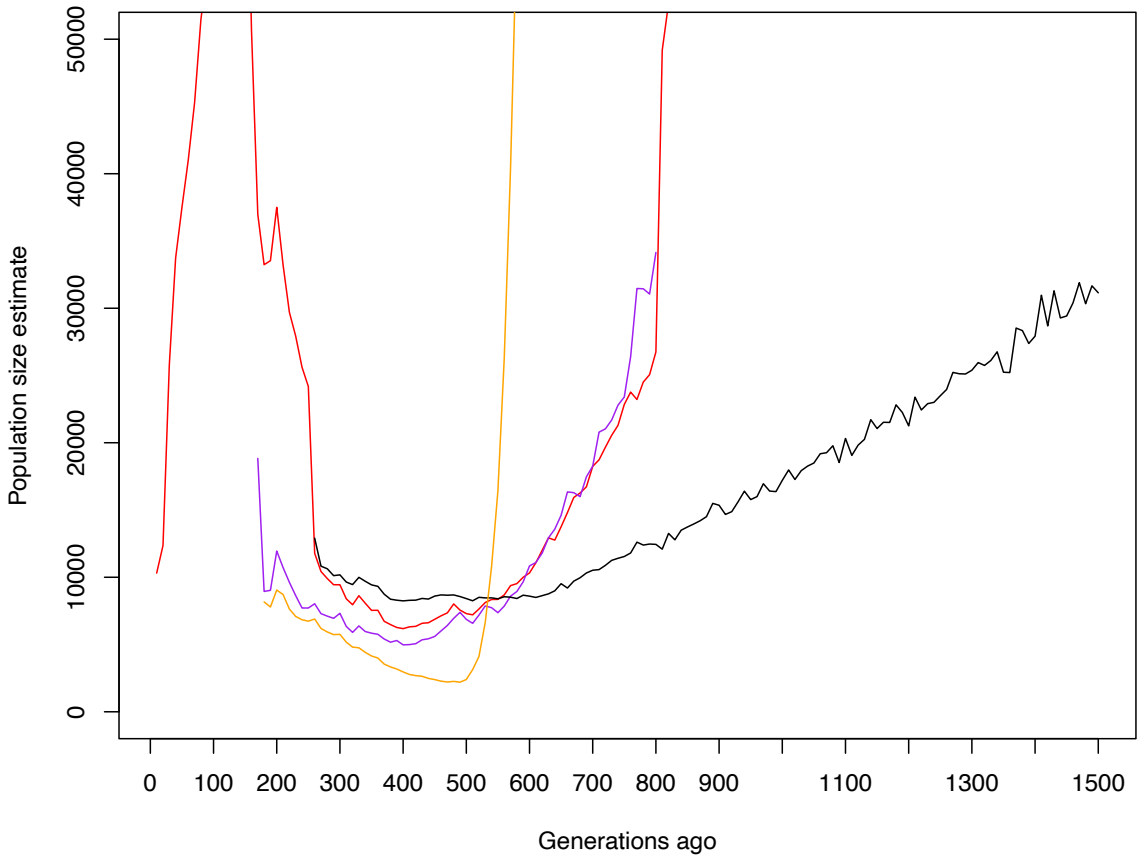


Fig. 5.5 Effective population size estimates for the MesoNeo genomes along paths 1-4, limiting the y axis at 50,000. Paths are coloured as in the model of European population structure in Figure 2.3.

5.3.2 Investigating selection of lactase persistence

Located in the MCM6 gene, rs4988235 is the marker SNP most strongly associated with lactase persistence into adulthood in Europeans. Carriers of the derived allele show expression of the LCT gene after weaning and therefore an adult with the derived allele is able to digest lactose in milk. The derived allele has been shown to have been under strong selection in Europeans in the past. However, the timing of selection and the population in which selection occurred are debated.

I constructed the RELATE tree sequence of chromosome 2 from the larger subset of samples, including all 1000 Genomes EUR samples and the additional ancient samples to explore selection at the LCT locus. The mutation lies on a branch with a parent node age of 243,208 years ago and a child node age of 228,088 years ago. The rs4988235 derived variant therefore appeared long before the beginning of my model of European population structure at around 45,000 years ago, which means any selection was on standing variation.

I applied my method of selection inference described above to the tree sequence. I calculated p-values for every generation along the four main paths up to 1500 generations before present for only derived 1000 Genomes samples. I obtained significant p-values for generations 6, 7, 9, 10, 11, 13, 21 and 159 along path 1 and generation 575 on path 4 (Figure 5.6). The results suggest selection occurred in Europeans mostly after the Bronze Age admixture, spanning 170-590 years before present with some indication of selection around 4,500 years ago near the beginning of the Bronze Age. There is also a highly significant signal around 16,100 years ago along path 4 in the EHG.

Selection on the lactase persistence allele seems to be occurring after all four paths have joined, shown in Figure 5.6 by colour red after 150 generations ago. But lineages in this time period have a path assignment and so I tested whether derived lineages are painted significantly more with one ancestry than expected. This was to help elucidate the path from which modern Europeans inherit the derived LCT allele. The number of samples painted

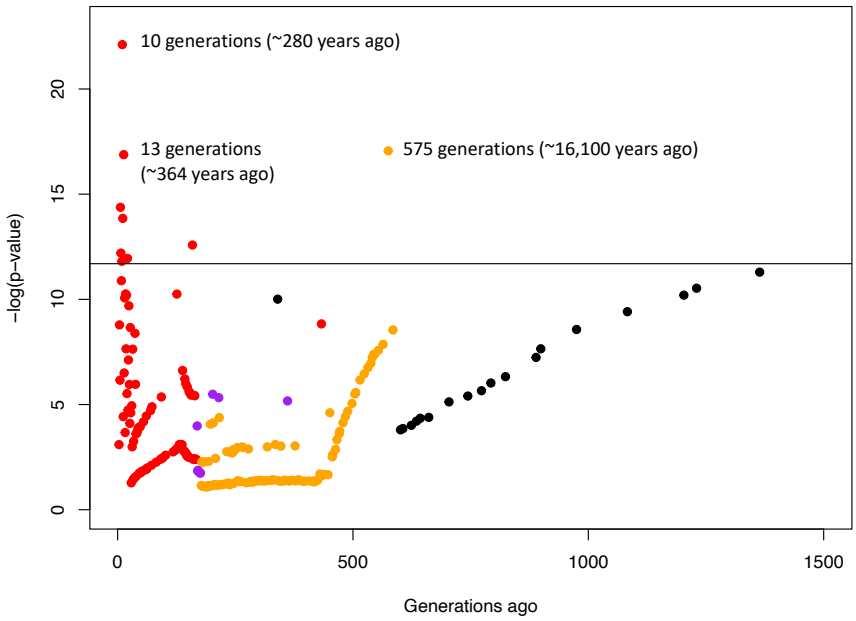


Fig. 5.6 For all 1000 Genomes EUR samples containing the derived rs4988235 variant, the -log(p-value) of the observed number of coalescences in the tree covering rs4988235 given the number of lineages and N_e estimate in each generation in each path. Points are plotted in generations where at least one coalescence event occurred. Colours correspond to the main paths 1-4, seen in Figure 2.3. The horizontal line marks the significance threshold, with a Bonferroni correction for multiple testing.

with each path is normally distributed across all trees. I counted the number of derived 1000 Genomes present day samples that are painted with each path at the rs4988235 variant. I then tested whether there is a significant enrichment and/or depletion of samples painted by each path covering the rs4988235 site given the total distribution of those same derived samples taking each path from all trees over chromosome 2.

The path labelling at the lactase persistence derived allele is highly enriched in path 4 and significantly depleted in other paths. The probability of observing 461 haploid samples painted with path 4 out of 511, given the chromosome-wide distribution of present-day samples, is 7.7e-37 (Figure 5.7A).

Moreover, I examined not just the labelling from the tree that covers the focal rs4988235 site but the surrounding +/- 25 trees which may be in linkage disequilibrium from selection.

The distribution of number of present day samples painted with path 4 in those 50 trees falls within the top tail of the total distribution across all trees (Figure 5.7B). The tree with the largest number, 499, of present day derived samples painted with path 4 is also captured in those 50 trees surrounding LCT.

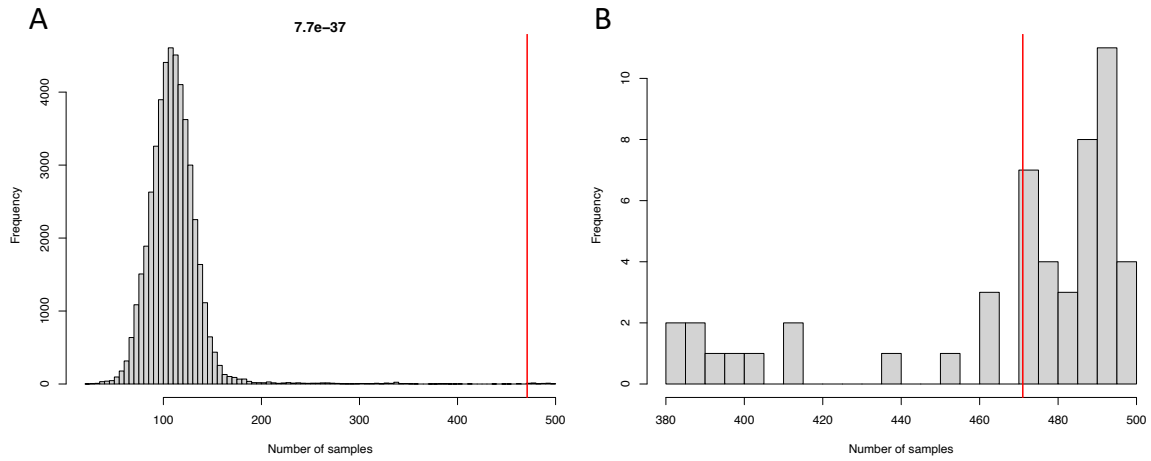


Fig. 5.7 Histograms showing the distribution of the number of present day samples painted with path 4 across all trees on chromosome 2. A: All trees across chromosome 2 counted. The red line shows where in the distribution the sample count for the tree covering rs4988235 lies. B: Counting only the 50 trees spanning across the rs4988235 site.

These results imply that the lactase persistence allele was significantly more likely to be inherited from path 4, Yamnaya/EHG lineages. There is a signal for positive selection on the derived variant in the EHG population at approximately 16,100 years ago and then again more recently in European history between 170-590 before present, after the joining of all paths in the Bronze Age.

One explanation for this is that pastoralism was initialised in the Eastern hunter gatherers with some early selection that may have helped establish the lactase persistence allele, as it gave these people the ability to extract more nutritional value from drinking milk. The

derived variant was then brought to Europe, for the most part, by the Yamnaya, after which most of the strong selection occurred in the European Bronze Age population. This suggest that the Yamnaya also brought the practice of pastoralism with them, along with the derived variant [110, 77].

I do not detect signals of positive selection in the periods in between the EHG selection and the Bronze Age selection in RELATE trees. This could be because I am unable to detect the signals due to a small selection coefficient. Alternatively, a small population size during this period might have been enough for drift to reduce the efficacy of selection and only later in the Bronze Age was selection on the allele strong enough to detect by my method.

However there are a couple of points to caveat these results. One is that I cannot rule out that the derived allele was preferentially selected for on the path 4 background in the Bronze Age population and EHG, perhaps because of some polygenic effect that is not related to pastoralism [22, 110]. Secondly, there have recently been concerns raise around the accuracy of imputation in highly selected regions. In particular, haplotypes in the LCT region were found to be nonrandom, specifically when there was an error in imputation, the common lactase persistence variant studied in this thesis was more likely to be imputed when not present in truth than the opposite [2].

Given that path 4 is strongly highlighted in my results of selection and painting of the LCT locus, a check for imputation error would be to check if the presence of the derived variant in Yamnaya and EHG samples are highly imputed and low coverage compared to other samples. Additionally, I could check if the allele frequency at the locus in Yamnaya samples is well-predicted by the EHG alleles. A disconnect between the groups would suggest that imputation error is having a substantial effect.

Chapter 6

Discussion and future work

Contents

In this thesis I have presented a method for inferring local ancestry in ancient and modern genomes given an explicit model of the structured population history for a set of samples. The redefinition of local ancestry as a path, rather than a static identity, back in time through various populations is more appropriate to histories involving multiple populations, admixture and split events and in the context of ancient samples. Almost all human populations have complex histories of this nature. I suggest that it is an important step to begin thinking of ancestry in this way rather than in terms of static identity. It is also this reframing of the meaning of ancestry which separates my method from pre-existing local ancestry inference tools.

My method performs as well as a leading local ancestry inference tool, GNOMix, for populations that are close to the time of admixture and performs better for populations further

since admixture. I have shown through extensive simulations that the method is robust to various demographic scenarios and mismatches between the training and testing parameters. I have used highly interpretable machine learning to perform a process that could be done manually. This helps to avoid the pitfalls of a 'black box' where it would be difficult to know which basic features are driving classification. When testing the classifiers under various scenarios, I can therefore understand why there are mistakes from the pattern of coalescence events.

The combination of the tree sequence structure and the local path painting opens the door to many downstream analyses, utilising the ability to place coalescence events along paths and/or edges. I have developed a method to infer the date of admixture for individual samples from their painted chromosomes. Applying this method to the MesoNeo genomes has revealed spatiotemporal patterns of admixture of different populations in Europe and I have showed how the inferred admixture date is more informative of this than using the archaeological sample age. Results from Neolithic farmer genomes suggest that the movement of Anatolian farmers into Europe and their subsequent admixture with WHG was slow and can be explained by geography to a substantial extent. Admixture appears to occur first in Iberia and moves north west over time, potentially supporting a route along the north Mediterranean coast of Anatolian farmers into Europe. I also show a small signal of the resurgence in WHG ancestry in the middle to late Neolithic.

In contrast, results from Bronze Age genomes suggest that the movement of steppe Yamnaya into Europe was fast with admixture occurring soon after the migration started followed by an increase in Neolithic farmer ancestry towards the late Bronze Age.

I found that the genetic formation of the Yamnaya is 5000-6000 years before present, a millennium or so before their inferred cultural formation determined from archaeology which is consistent with other recent results citedates. I also show evidence of potential contact between the EHG and CHG in the eighth millennium before present.

Lastly, I show evidence that the lactase persistence variant was inherited by present day Europeans from the EHG/Yamnaya path and that selection occurred on this variant, primarily recently in history during the last 600 years, but potentially also near the start of the Bronze Age, 4,500 years ago, and deeper in time in the EHG around 16,100 years ago. These results might suggest the initialisation of pastoralism in the EHG with placing the variant under some selection and that the Yamnaya brought the variant to Europe along with pastoralism.

Next, I will discuss some of the drawbacks of the method, how these might be compensated for, and then how, in the future, the method can be improved.

6.1 Caveats

The first caveat to mention is that the method has its foundations the ability of RELATE to correctly construct a tree sequence. In essence I am embedding a tree sequence within a population structure and thereby assuming that the trees fit within the model constraints i.e that illegal coalescences are minimal. There are two reasons why this may not happen: 1. The model is far from the true population structure or 2. RELATE is not inferring the tree sequences correctly, despite the model representing the true structure well. The first point I will address later. For the second point, I have tried to compensate for biases in the RELATE inference by training the classifier with GNNs extracted from RELATE inferred trees with labels from the true simulated trees. Results on simulations indicate this is successful.

A second, connected point is that imputation of ancient genomes is a fairly new approach, only possible recently due to the large reference panels of modern high-quality genomes now available. While extensive analysis was carried out to test the ability of GLIMPSE to recover down-sampled high coverage ancient genomes, the branch length inference of RELATE is vulnerable to imputation. Imputation is essentially the process of borrowing information from closely related, higher coverage individuals. This has the effect of making samples that coalescence most recently more similar to each other than their true sequences actually are.

Intuitively this effect should help the topology building process by making identification of nearest neighbours easier. However, the same effect will also bias the coalescence times younger as sequences will have fewer differences than in truth. Additionally, a shortfall in imputation is there is no way to recover variants that are private to ancient groups and so not present in the reference panel. This may both hamper neighbour identification within ancient groups and also downward bias the estimated coalescence times.

One solution to this downward bias is to use the unimputed data and only calculate the branch lengths for well genotyped, high coverage individuals. If the topology is kept the same as that inferred from the full imputed set, then the low-coverage samples could be added back to the trees constrained by the branch lengths computed on high coverage samples, perhaps using the coalescent prior. Another solution could be to use the genotype posteriors to estimate low coverage branch lengths and integrate over all possible placements of mutations on branches.

Even if we accept that inferring tree sequences with RELATE from imputed data is likely to produce coalescence times that are biased downward, there is no reason to think this should affect the accuracy of inferring tree topologies. Given that my method of classification using GNN structures relies on the topology alone, this might explain why my painting results appeared to be more accurate and stronger than my analyses based on coalescence timing.

The familiar quote ‘all models are wrong, some are useful’ is applicable to the method described in this thesis. Any model of the past demographic structure of a population will be inaccurate in many ways. In Section 2.4.3 I demonstrated how the accuracy of classification of paths drops as the number of paths increases so in most cases the structure will be simplified to represent only major population events and ignore smaller scale migrations Yet even with a simplified structure we can still obtain results that are interesting and meaningful, answering the questions posed. However, for some populations, we may have no pre-existing knowledge of the population structure and results from PCA and ADMIXTURE analysis

may be unclear. While we have shown in Section 2.4.4 that misspecifications in the model can be compensated for by generalisation of the neural network, with too little understanding of the history we may not obtain results of much confidence or meaning.

6.2 Adapting to other populations

Nevertheless, with populations where the past population structure is well understood or analyses such as PCA and ADMIXTURE produce good results, the method can be adapted to these populations. It is becoming more apparent, especially from ancient DNA analysis, that the history of human populations across the globe is characterised by multiple population split, admixture, migration and isolation events and an appropriate analogy is a braided river rather than a tree. This means that paths going back in time can pass through multiple interrelated populations. Along with the growing amount of ancient DNA data becoming available, the path concept of ancestry is applicable to many more cases other than Europeans.The model presented in this thesis of European population structure is in the stdpopsim catalogue and some other populations already have populations structures available in the stdpopsim catalogue from which to simulate data: https://popsim-consortium.github.io/stdpopsim-docs/stable/catalog.html. For those populations where demographic model is not available one must be custom built with msprime from the results of other analysis tools. The level of complexity and parameters will depend on the level of accuracy the user is willing to accept and the particular questions that are posed. The total process to apply the method to a new population is as follows:

1. If not already available in stdpopsim catalogue, construct a model of the population structure from the samples available.

2. Define the paths going backwards in time through the model.

3. Simulate chromosomes from the model.

4. Run RELATE on the simulated data.

5. Extract GNNs and train a neural network against the true path labels from the simulated data.

6. Run RELATE on the real samples.

7. Classify paths in the real data using the neural network.

6.3 Future improvements in ARG inference

Here I have used RELATE to infer tree sequences but other tools that infer tree sequences or ARGs exist, notably ARGweaver [102], tsinfer/tsdate[53] and more recently ARG needle [129]. The pros and cons of these tools have been reviewed, analysing both coalescence times and topology accuracy. Brandt et al. [127] showed that ARGweaver tends to have more accurate coalescence time inference than RELATE and tsinfer with RELATE slightly more accurate than tsinfer. Kelleher et al. [53] showed that the Kendall-Colijn tree distance, a metric of how similar inferred to simulated trees are, is similar in tsinfer and ARGweaver.

ARGweaver is substantially slower, cannot scale to sample sizes of more than around 40 and cannot incorporate ancient DNA, making it the least appropriate for use with my method, despite it producing the most accurate coalescence times. tsinfer/tsdate can scale to thousands of haplotypes and can incorporate ancient DNA but is marginally less accurate in coalescence time inference than RELATE. Additionally, it allows for pervasive multifurcating nodes which change the GNN structure. At the time of writing I would recommend using RELATE to produce the tree sequences for our method. It has strictly bifurcating nodes, can scale to thousands of haplotypes and incorporate ancient DNA with less of a trade-off in accuracy than tsinfer.

Further advances in the ability of these and future tools to infer more accurate tree sequences will likely improve our method. In particular, the application of these methods to

imputed data will allow us to exploit the great advantages of imputation to boost sample size and SNPs data. While I recommend RELATE for the reasons described, the tree sequence inference step of the method can be performed by any tool that produces a tree sequence structure from variant data. Any superior future tools that also produce tree sequences will simply be 'plug-and-play' due to the modularity of my method.

6.4 RELATE inference and population structure

In its current implementation, RELATE assumes a panmictic population where all individuals randomly mate at all times. In reality, most if not all species display population structure to some degree. Given a model of population structure such as the one I have constructed for Europeans, the branch length re-estimation steps of RELATE could be altered to incorporate this structure.

Once the topologies and initial branch length estimates are calculated, time is split into epochs and an algorithm is applied that iteratively calculates piecewise constant coalescence rates in each epoch and re-estimates branch lengths using an MCMC approach. With a model of ancestral population structure, there can also be multiple edges in each time epoch, with a separate coalescence rate (inverse effective population size) for each edge.

The maximum likelihood coalescence rates are therefore expanded to be piece-wise by time epoch and structure edge. Given that $\lambda^{e,ed}$ is the constant piecewise coalescence rate in epoch e and edge ed and t_k is the time of the coalescence event, given path labelling, reducing the number of lineages in epoch e and edge ed from k to $k-1$, then the likelihood for tree z is

$$P(z) = \prod_{e}\prod_{ed}\prod_{k}\left(\lambda^{e,ed}\exp\left(-\binom{k}{2}\lambda^{e,ed}(t_k - t_{k-1})\right)\right).$$

Taking logs on both sides,

$$log(P(z)) = \sum_{e}\sum_{ed}\left(\sum_{k} log(\lambda^{e,ed}) - \sum_{k}\binom{k}{2}\lambda^{e,ed}(t_k - t_{k-1})\right).$$

Assuming independence across trees, the log-likelihood for the whole genome is given by $\sum_z log(P(z))$. By differentiating with respect to $\lambda^{e,ed}$, setting equal to 0 and rearranging, the maximum likelihood coalescence rate in epoch e and edge ed given by,

$$\hat{\lambda}^{e,ed} = \frac{n_{e,ed}}{\sum_k \binom{k}{2} T^k_{e,ed}},$$

where $n_{e,ed}$ is the number of coalescences between two haplotypes in epoch e and edge ed, multiplied by the length of the tree along the genome, across all trees and $T^k_{e,ed}$ is the total time spent with k lineages in that epoch and edge [117].

Assuming that branches and coalescences that occur on one edge can be treated independently from the rest of the tree, MCMC proposals can be made to move nodes locally within epochs and between edges and the likelihood ratio of these proposals calculated locally. When moving nodes between epoch and edges in the MCMC branch length re-estimation step, both the coalescence rates per epoch and edge must be used and also the correct number of lineages used to calculate the ratio of coalescent priors (equation 1.3). For example, if a proposed change moves a node to before a population split and into a new time epoch, both the coalescence rate it experiences will change and the number of lineages will increase from the time of the split. Penalisation of proposed moves that would change a 'legal' coalescence event into an 'illegal' events could also be implemented.

The idea is to define a reversible Markov chain with the stationary distribution $P(\mathbf{t}|\mathbf{m})$, where $\mathbf{t}$ is the collection of all branch lengths and $\mathbf{m}$ all mutations, given a population structure of paths and a variable population size. While this approach is parametric in that the population structure, topology and split/admixture times are supplied, it could help to

refine population size estimates and selection analysis on edges. I started to implement this model but ran into technical difficulties.

References

[1] Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655–1664.

[2] Ali, A. T., Liebert, A., Lau, W., Maniatis, N., and Swallow, D. M. (2022). The hazards of genotype imputation in chromosomal regions under selection: A case study using the lactase gene region. *Ann Hum Genet*, 86(1):24–33.

[3] Allentoft, M. E., Sikora, M., Refoyo-Martínez, A., Irving-Pease, E. K., Fischer, A., Barrie, W., Ingason, A., Stenderup, J., Sjögren, K.-G., Pearson, A., Mota, B., Paulsson, B. S., Halgren, A., Macleod, R., Schjellerup Jørkov, M. L., Demeter, F., Novosolov, M., Sørensen, L., Nielsen, P.-O., Henriksen, R. H. A., Vimala, T., McColl, H., Margaryan, A., Ilardo, M., Vaughn, A., Mortensen, M. F., Nielsen, A. B., Hede, M. U., Rasmussen, P., Vinner, L., Renaud, G., Stern, A., Trolle Jensen, T. Z., Johannsen, N. N., Scorrano, G., Schroeder, H., Lysdahl, P., Ramsøe, A. D., Skorobogatov, A., Schork, A. J., Rosengren, A., Ruter, A., Outram, A., Timoshenko, A. A., Buzhilova, A., Coppa, A., Zubova, A., Silva, A. M., Hansen, A. J., Gromov, A., Logvin, A., Gotfredsen, A. B., Nielsen, B. H., González-Rabanal, B., Lalueza-Fox, C., McKenzie, C. J., Gaunitz, C., Blasco, C., Liesau, C., Martinez-Labarga, C., Pozdnyakov, D. V., Cuenca-Solana, D., Lordkipanidze, D. O., En'shin, D., Salazar-García, D. C., Price, T. D., Borić, D., Kostyleva, E., Veselovskaya, E. V., Usmanova, E. R., Cappellini, E., Petersen, E. B., Kannegaard, E., Radina, F., Yediay, F. E., Duday, H., Gutiérrez-Zugasti, I., Potekhina, I., Shevnina, I., Altinkaya, I., Guilaine, J., Hansen, J., Tortosa, J. E. A., Zilhão, J., Vega, J., Pedersen, K. B., Tunia, K., Zhao, L., Mylnikova, L. N., Larsson, L., Metz, L., Yeppiskoposyan, L., Pedersen, L., Sarti, L., Orlando, L., Slimak, L., Klassen, L., Blank, M., González-Morales, M., Silvestrini, M., Vretemark, M., Nesterova, M. S., Rykun, M., Rolfo, M. F., Szmyt, M., Przybyła, M., Calattini, M., Sablin, M., Dobisíková, M., Meldgaard, M., Johansen, M., Berezina, N., Card, N., Saveliev, N. A., Poshekhonova, O., Rickards, O., Lozovskaya, O. V., Uldum, O. C., Aurino, P., Kosintsev, P., Courtaud, P., Ríos, P., Mortensen, P., Lotz, P., Persson, P. Å., Bangsgaard, P., Damgaard, P. d. B., Petersen, P. V., Martinez, P. P., Włodarczak, P., Smolyaninov, R. V., Maring, R., Menduiña, R., Badalyan, R., Iversen, R., Turin, R., Vasilyiev, S., Wåhlin, S., Borutskaya, S., Skochina, S., Sørensen, S. A., Andersen, S. H., Jørgensen, T., Serikov, Y. B., Molodin, V. I., Smrcka, V., Merz, V., Appadurai, V., Moiseyev, V., Magnusson, Y., Kjær, K. H., Lynnerup, N., Lawson, D. J., Sudmant, P. H., Rasmussen, S., Korneliussen, T., Durbin, R., Nielsen, R., Delaneau, O., Werge, T., Racimo, F., Kristiansen, K., and Willerslev, E. (2022). Population Genomics of Stone Age Eurasia. *bioRxiv*.

[4] Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan,

A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Casa, P. D., Dąbrowski, P., Duffy, P. R., Ebel, A. V., Epimakhov, A., Frei, K., Furmanek, M., Gralak, T., Gromov, A., Gronkiewicz, S., Grupe, G., Hajdu, T., Jarysz, R., Khartanovich, V., Khokhlov, A., Kiss, V., Kolář, J., Kriiska, A., Lasak, I., Longhi, C., McGlynn, G., Merkevicius, A., Merkyte, I., Metspalu, M., Mkrtchyan, R., Moiseyev, V., Paja, L., Pálfi, G., Pokutta, D., Pospieszny, Ł., Price, T. D., Saag, L., Sablin, M., Shishlina, N., Smrčka, V., Soenov, V. I., Szeverényi, V., Tóth, G., Trifanova, S. V., Varul, L., Vicze, M., Yepiskoposyan, L., Zhitenev, V., Orlando, L., Sicheritz-Pontén, T., Brunak, S., Nielsen, R., Kristiansen, K., and Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, 522(7555):167–172.

[5] Atkinson, E. G., Maihofer, A. X., Kanai, M., Martin, A. R., Karczewski, K. J., Santoro, M. L., Ulirsch, J. C., Kamatani, Y., Okada, Y., Finucane, H. K., Koenen, K. C., Nievergelt, C. M., Daly, M. J., and Neale, B. M. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nature Genetics*, 53(2):195–204.

[6] Ausmees, K., Sanchez-Quinto, F., Jakobsson, M., and Nettelblad, C. (2022). An empirical evaluation of genotype imputation of ancient DNA. *G3 (Bethesda)*, 12(6).

[7] Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., and Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10):1359–1367.

[8] Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

[9] Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98(8):4563–4568.

[10] Bhaskar, A. and Song, Y. S. (2014). Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann Stat*, 42(6):2469–2493.

[11] Bloss, C. S., Darst, B. F., Topol, E. J., and Schork, N. J. (2011). Direct-to-consumer personalized genomic testing. *Human Molecular Genetics*, 20(R2):R132–R141.

[12] Brace, S., Diekmann, Y., Booth, T. J., van Dorp, L., Faltyskova, Z., Rohland, N., Mallick, S., Olalde, I., Ferry, M., Michel, M., Oppenheimer, J., Broomandkhoshbacht, N., Stewardson, K., Martiniano, R., Walsh, S., Kayser, M., Charlton, S., Hellenthal, G., Armit, I., Schulting, R., Craig, O. E., Sheridan, A., Parker Pearson, M., Stringer, C., Reich, D., Thomas, M. G., and Barnes, I. (2019). Ancient genomes indicate population replacement in Early Neolithic Britain. *Nature Ecology & Evolution*, 3(5):765–771.

[13] Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., and Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621.

[14] Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet*, 103(3):338–348.

[15] Brunel, S., Bennett, E. A., Cardin, L., Garraud, D., Barrand Emam, H., Beylier, A., Boulestin, B., Chenal, F., Ciesielski, E., Convertini, F., Dedet, B., Desbrosse-Degobertiere, S., Desenne, S., Dubouloz, J., Duday, H., Escalon, G., Fabre, V., Gailledrat, E., Gandelin, M., Gleize, Y., Goepfert, S., Guilaine, J., Hachem, L., Ilett, M., Lambach, F., Maziere, F., Perrin, B., Plouin, S., Pinard, E., Praud, I., Richard, I., Riquier, V., Roure, R., Sendra, B., Thevenet, C., Thiol, S., Vauquelin, E., Vergnaud, L., Grange, T., Geigl, E.-M., and Pruvost, M. (2020). Ancient genomes from present-day France unveil 7,000 years of its demographic history. *Proc Natl Acad Sci U S A*, 117(23):12791–12798.

[16] Cassidy, L. M., Martiniano, R., Murphy, E. M., Teasdale, M. D., Mallory, J., Hartwell, B., and Bradley, D. G. (2016). Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proceedings of the National Academy of Sciences*, 113(2):368–373.

[17] Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, 7(2):98–108.

[18] Chintalapati, M., Patterson, N., and Moorjani, P. (2022). The spatiotemporal patterns of major human admixture events during the european holocene. *Elife*, 11.

[19] Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., Mitrovica, J. X., Hostetler, S. W., and McCabe, A. M. (2009). The Last Glacial Maximum. *Science*, 325(5941):710–714.

[20] Cunha, E. and Ubelaker, D. H. (2020). Evaluation of ancestry from human skeletal remains: a concise review. *Forensic Sci Res*, 5(2):89–97.

[21] Currat, M., Arenas, M., Quilodràn, C. S., Excoffier, L., and Ray, N. (2019). SPLATCHE3: simulation of serial genetic data under spatially explicit evolutionary scenarios including long-distance dispersal. *Bioinformatics*, 35(21):4480–4483.

[22] Evershed, R. P., Davey Smith, G., Roffet-Salque, M., Timpson, A., Diekmann, Y., Lyon, M. S., Cramp, L. J. E., Casanova, E., Smyth, J., Whelton, H. L., Dunne, J., Brychova, V., Šoberl, L., Gerbault, P., Gillis, R. E., Heyd, V., Johnson, E., Kendall, I., Manning, K., Marciniak, A., Outram, A. K., Vigne, J.-D., Shennan, S., Bevan, A., Colledge, S., Allason-Jones, L., Amkreutz, L., Anders, A., Arbogast, R.-M., Bălăşescu, A., Bánffy, E., Barclay, A., Behrens, A., Bogucki, P., Carrancho Alonso, Á., Carretero, J., Cavanagh, N., Claßen, E., Collado Giraldo, H., Conrad, M., Csengeri, P., Czerniak, L., Dębiec, M., Denaire, A., Domboróczki, L., Donald, C., Ebert, J., Evans, C., Francés-Negro, M., Gronenborn, D., Haack, F., Halle, M., Hamon, C., Hülshoff, R., Ilett, M., Iriarte, E., Jakucs, J., Jeunesse, C., Johnson, M., Jones, A. M., Karul, N., Kiosak, D., Kotova, N., Krause, R., Kretschmer, S., Krüger, M., Lefranc, P., Lelong, O., Lenneis, E., Logvin, A., Lüth, F., Marton, T., Marley, J., Mortimer, R., Oosterbeek, L., Oross, K., Pavúk, J., Pechtl, J., Pétrequin, P., Pollard, J., Pollard, R., Powlesland, D., Pyzel, J., Raczky, P., Richardson, A., Rowe, P., Rowland, S., Rowlandson, I., Saile, T., Sebők, K., Schier, W., Schmalfuß, G., Sharapova, S., Sharp, H., Sheridan, A., Shevnina, I., Sobkowiak-Tabaka, I., Stadler, P., Stäuble, H., Stobbe, A., Stojanovski, D., Tasić, N., van Wijk, I., Vostrovská, I., Vuković, J., Wolfram, S., Zeeb-Lanz, A., and Thomas, M. G. (2022). Dairying, diseases and the evolution of lactase persistence in Europe. *Nature*, 608(7922):336–345.

[23] Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., and Sousa, V. C. (2021). fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*, 37(24):4882–4885.

[24] Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

[25] Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413.

[26] Fleming, J. O. (2011). Helminths and multiple sclerosis: will old friends give us new treatments for MS? *J Neuroimmunol*, 233(1-2):3–5.

[27] Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., Nickel, B., Peltzer, A., Rohland, N., Slon, V., Talamo, S., Lazaridis, I., Lipson, M., Mathieson, I., Schiffels, S., Skoglund, P., Derevianko, A. P., Drozdov, N., Slavinsky, V., Tsybankov, A., Cremonesi, R. G., Mallegni, F., Gély, B., Vacca, E., Morales, M. R. G., Straus, L. G., Neugebauer-Maresch, C., Teschler-Nicola, M., Constantin, S., Moldovan, O. T., Benazzi, S., Peresani, M., Coppola, D., Lari, M., Ricci, S., Ronchitelli, A., Valentin, F., Thevenet, C., Wehrberger, K., Grigorescu, D., Rougier, H., Crevecoeur, I., Flas, D., Semal, P., Mannino, M. A., Cupillard, C., Bocherens, H., Conard, N. J., Harvati, K., Moiseyev, V., Drucker, D. G., Svoboda, J., Richards, M. P., Caramelli, D., Pinhasi, R., Kelso, J., Patterson, N., Krause, J., Pääbo, S., and Reich, D. (2016). The genetic history of Ice Age Europe. *Nature*, 534(7606):200–205.

[28] Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229.

[29] Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., Whittle, A., Dani, J., Raczky, P., Higham, T. F. G., Hofreiter, M., Bradley, D. G., and Pinhasi, R. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5(1):5257.

[30] Garrigan, D. and Hedrick, P. W. (2003). Perspective: Detecting adaptive molecular polymorphism: Lessons from the MHC. *Evolution*, 57(8):1707–1722.

[31] Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLOS Genetics*, 11(2).

[32] Geibel, J., Reimer, C., Pook, T., Weigend, S., Weigend, A., and Simianer, H. (2021). How imputation can mitigate SNP ascertainment Bias. *BMC Genomics*, 22(1):340.

[33] Geza, E., Mugo, J., Mulder, N. J., Wonkam, A., Chimusa, E. R., and Mazandu, G. K. (2019). A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. *Brief Bioinform*, 20(5):1709–1724.

[34] Goldberg, A., Günther, T., Rosenberg, N. A., and Jakobsson, M. (2017). Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations. *Proceedings of the National Academy of Sciences*, 114(10):2657–2662.

[35] González-Fortes, G., Jones, E. R., Lightfoot, E., Bonsall, C., Lazar, C., Grandal-d'Anglade, A., Garralda, M. D., Drak, L., Siska, V., Simalcsik, A., Boroneanţ, A., Vidal Romaní, J. R., Vaqueiro Rodríguez, M., Arias, P., Pinhasi, R., Manica, A., and Hofreiter, M. (2017). Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. *Curr Biol*, 27(12):1801–1810.

[36] Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, 191(2):607–619.

[37] Griffiths, R. C. and Tavare, S. (1994). Simulating Probability Distributions in the Coalescent. *Theoretical Population Biology*, 46(2):131–159.

[38] Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, 5(10).

[39] Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S. L., Risch, R., Rojo Guerra, M. A., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K. W., and Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211.

[40] Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. *Science*, 343(6172):747–751.

[41] Hey, J. (2009). Isolation with Migration Models for More Than Two Populations. *Molecular Biology and Evolution*, 27(4):905–920.

[42] Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A*, 104(8):2785–2790.

[43] Hilmarsson, H., Kumar, A. S., Rastogi, R., Bustamante, C. D., Montserrat, D. M., and Ioannidis, A. G. (2021). High Resolution Ancestry Deconvolution for Next Generation Genomic Data. *bioRxiv*.

[44] Hobolth, A., Uyenoyama, M. K., and Wiuf, C. (2008). Importance sampling for the infinite sites model. *Stat Appl Genet Mol Biol*, 7(1).

[45] Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-Del-Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V., Kirsanow, K., Cassidy, L. M., Martiniano, R., Strobel, M., Scheu, A., Kotsakis, K., Halstead, P., Triantaphyllou, S.,

Kyparissi-Apostolika, N., Urem-Kotsou, D., Ziota, C., Adaktylou, F., Gopalan, S., Bobo, D. M., Winkelbach, L., Blöcher, J., Unterländer, M., Leuenberger, C., Çilingiroğlu, Ç., Horejs, B., Gerritsen, F., Shennan, S. J., Bradley, D. G., Currat, M., Veeramah, K. R., Wegmann, D., Thomas, M. G., Papageorgopoulou, C., and Burger, J. (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc Natl Acad Sci U S A*, 113(25):6886–6891.

[46] Howie, B. N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*, 5(6):e1000529.

[47] Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23(2):183–201.

[48] Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in genetics: TIG*, 18(9):486–486.

[49] Jones, E. R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R. L., Gallego Llorente, M., Cassidy, L. M., Gamba, C., Meshveliani, T., Bar-Yosef, O., Müller, W., Belfer-Cohen, A., Matskevich, Z., Jakeli, N., Higham, T. F. G., Currat, M., Lordkipanidze, D., Hofreiter, M., Manica, A., Pinhasi, R., and Bradley, D. G. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, 6(1):8912.

[50] Jones, E. R., Zarina, G., Moiseyev, V., Lightfoot, E., Nigst, P. R., Manica, A., Pinhasi, R., and Bradley, D. G. (2017). The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. *Curr Biol*, 27(4):576–582.

[51] Kamm, J., Terhorst, J., Durbin, R., and Song, Y. S. (2020). Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *Journal of the American Statistical Association*, 115(531):1472–1487.

[52] Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, 12(5):e1004842–.

[53] Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nat Genet*, 51(9):1330–1338.

[54] Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232.

[55] Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248.

[56] Kuhner, M. K., Beerli, P., Yamato, J., and Felsenstein, J. (2000). Usefulness of Single Nucleotide Polymorphism Data for Estimating Population Parameters. *Genetics*, 156(1):439–447.

[57] Kuhner, M. K., Yamato, J., and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140(4):1421–1430.

[58] Lachance, J. and Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays*, 35(9):780–786.

[59] Lazaridis, I., Alpaslan-Roodenberg, S., Acar, A., Açıkkol, A., Agelarakis, A., Aghikyan, L., Akyüz, U., Andreeva, D., Andrijašević, G., Antonović, D., Armit, I., Atmaca, A., Avetisyan, P., Aytek, A. İ., Bacvarov, K., Badalyan, R., Bakardzhiev, S., Balen, J., Bejko, L., Bernardos, R., Bertsatos, A., Biber, H., Bilir, A., Bodružić, M., Bonogofsky, M., Bonsall, C., Borić, D., Borovinić, N., Morante, G. B., Buttinger, K., Callan, K., Candilio, F., Carić, M., Cheronet, O., Chohadzhiev, S., Chovalopoulou, M.-E., Chryssoulaki, S., Ciobanu, I., Čondić, N., Constantinescu, M., Cristiani, E., Culleton, B. J., Curtis, E., Davis, J., Davtyan, R., Demcenco, T. I., Dergachev, V., Derin, Z., Deskaj, S., Devejyan, S., Djordjević, V., Carlson, K. S. D., Eccles, L. R., Elenski, N., Engin, A., Erdoğan, N., Erir-Pazarcı, S., Fernandes, D. M., Ferry, M., Freilich, S., Frînculeasa, A., Galaty, M. L., Gamarra, B., Gasparyan, B., Gaydarska, B., Genç, E., Gültekin, T., Gündüz, S., Hajdu, T., Heyd, V., Hobosyan, S., Hovhannisyan, N., Iliev, I., Iliev, L., Iliev, S., İvgin, İ., Janković, I., Jovanova, L., Karkanas, P., Kavaz-Kındığılı, B., Kaya, E. H., Keating, D., Kennett, D. J., Kesici, S. D., Khudaverdyan, A., Kiss, K., Kılıç, S., Klostermann, P., Valdes, S. K. B. N., Kovačević, S., Krenz-Niedbała, M., Škrivanko, M. K., Kurti, R., Kuzman, P., Lawson, A. M., Lazar, C., Leshtakov, K., Levy, T. E., Liritzis, I., Lorentz, K. O., Łukasik, S., Mah, M., Mallick, S., Mandl, K., Martirosyan-Olshansky, K., Matthews, R., Matthews, W., McSweeney, K., Melikyan, V., Micco, A., Michel, M., Milašinović, L., Mittnik, A., Monge, J. M., Nekhrizov, G., Nicholls, R., Nikitin, A. G., Nikolov, V., Novak, M., Olalde, I., Oppenheimer, J., Osterholtz, A., Özdemir, C., Özdoğan, K. T., Öztürk, N., Papadimitriou, N., Papakonstantinou, N., Papathanasiou, A., Paraman, L., Paskary, E. G., Patterson, N., Petrakiev, I., Petrosyan, L., Petrova, V., Philippa-Touchais, A., Piliposyan, A., Kuzman, N. P., Potrebica, H., Preda-Bălănică, B., Premužić, Z., Price, T. D., Qiu, L., Radović, S., Aziz, K. R., Šikanjić, P. R., Raheem, K. R., Razumov, S., Richardson, A., Roodenberg, J., Ruka, R., Russeva, V., Şahin, M., Şarbak, A., Savaş, E., Schattke, C., Schepartz, L., Selçuk, T., Sevim-Erol, A., Shamoon-Pour, M., Shephard, H. M., Sideris, A., Simalcsik, A., Simonyan, H., Sinika, V., Sirak, K., Sirbu, G., Šlaus, M., Soficaru, A., Söğüt, B., Sołtysiak, A., Sönmez-Sözer, Ç., Stathi, M., Steskal, M., Stewardson, K., Stocker, S., Suata-Alpaslan, F., Suvorov, A., Szécsényi-Nagy, A., Szeniczey, T., Telnov, N., Temov, S., Todorova, N., Tota, U., Touchais, G., Triantaphyllou, S., Türker, A., Ugarković, M., Valchev, T., Veljanovska, F., Videvski, Z., Virag, C., Wagner, A., Walsh, S., Włodarczak, P., Workman, J. N., Yardumian, A., Yarovoy, E., Yavuz, A. Y., Yılmaz, H., Zalzala, F., Zettl, A., Zhang, Z., Çavuşoğlu, R., Rohland, N., Pinhasi, R., and Reich, D. (2022). The genetic history of the Southern Arc: A bridge between West Asia and Europe. *Science*, 377(6609):eabm4247.

[60] Lazaridis, I., Mittnik, A., Patterson, N., Mallick, S., Rohland, N., Pfrengle, S., Furtwängler, A., Peltzer, A., Posth, C., Vasilakis, A., McGeorge, P. J. P., Konsolaki-Yannopoulou, E., Korres, G., Martlew, H., Michalodimitrakis, M., Özsait, M., Özsait, N., Papathanasiou, A., Richards, M., Roodenberg, S. A., Tzedakis, Y., Arnott, R., Fernandes, D. M., Hughey, J. R., Lotakis, D. M., Navas, P. A., Maniatis, Y., Stamatoyannopoulos, J. A., Stewardson,

K., Stockhammer, P., Pinhasi, R., Reich, D., Krause, J., and Stamatoyannopoulos, G. (2017). Genetic origins of the Minoans and Mycenaeans. *Nature*, 548(7666):214–218.

[61] Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E. R., Roodenberg, S. A., Lengyel, G., Bocquentin, F., Gasparian, B., Monge, J. M., Gregg, M., Eshed, V., Mizrahi, A.-S., Meiklejohn, C., Gerritsen, F., Bejenaru, L., Blüher, M., Campbell, A., Cavalleri, G., Comas, D., Froguel, P., Gilbert, E., Kerr, S. M., Kovacs, P., Krause, J., McGettigan, D., Merrigan, M., Merriwether, D. A., O'Reilly, S., Richards, M. B., Semino, O., Shamoon-Pour, M., Stefanescu, G., Stumvoll, M., Tönjes, A., Torroni, A., Wilson, J. F., Yengo, L., Hovhannisyan, N. A., Patterson, N., Pinhasi, R., and Reich, D. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424.

[62] Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., Sawyer, S., Posth, C., Haak, W., Hallgren, F., Fornander, E., Rohland, N., Delsate, D., Francken, M., Guinet, J.-M., Wahl, J., Ayodo, G., Babiker, H. A., Bailliet, G., Balanovska, E., Balanovsky, O., Barrantes, R., Bedoya, G., Ben-Ami, H., Bene, J., Berrada, F., Bravi, C. M., Brisighelli, F., Busby, G. B. J., Cali, F., Churnosov, M., Cole, D. E. C., Corach, D., Damba, L., van Driem, G., Dryomov, S., Dugoujon, J.-M., Fedorova, S. A., Gallego Romero, I., Gubina, M., Hammer, M., Henn, B. M., Hervig, T., Hodoglugil, U., Jha, A. R., Karachanak-Yankova, S., Khusainova, R., Khusnutdinova, E., Kittles, R., Kivisild, T., Klitz, W., Kučinskas, V., Kushniarevich, A., Laredj, L., Litvinov, S., Loukidis, T., Mahley, R. W., Melegh, B., Metspalu, E., Molina, J., Mountain, J., Näkkäläjärvi, K., Nesheva, D., Nyambo, T., Osipova, L., Parik, J., Platonov, F., Posukh, O., Romano, V., Rothhammer, F., Rudan, I., Ruizbakiev, R., Sahakyan, H., Sajantila, A., Salas, A., Starikovskaya, E. B., Tarekegn, A., Toncheva, D., Turdikulova, S., Uktveryte, I., Utevska, O., Vasquez, R., Villena, M., Voevoda, M., Winkler, C. A., Yepiskoposyan, L., Zalloua, P., Zemunik, T., Cooper, A., Capelli, C., Thomas, M. G., Ruiz-Linares, A., Tishkoff, S. A., Singh, L., Thangaraj, K., Villems, R., Comas, D., Sukernik, R., Metspalu, M., Meyer, M., Eichler, E. E., Burger, J., Slatkin, M., Pääbo, S., Kelso, J., Reich, D., and Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413.

[63] Lazaridis, I. and Reich, D. (2017). Failure to replicate a genetic signal for sex bias in the steppe migration into central Europe. *Proceedings of the National Academy of Sciences*, 114(20):E3873–E3874.

[64] Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

[65] Liang, M. and Nielsen, R. (2014). The lengths of admixture tracts. *Genetics*, 197(3):953–967.

[66] Lipson, M., Szécsényi-Nagy, A., Mallick, S., Pósa, A., Stégmár, B., Keerl, V., Rohland, N., Stewardson, K., Ferry, M., Michel, M., Oppenheimer, J., Broomandkhoshbacht, N., Harney, E., Nordenfelt, S., Llamas, B., Gusztáv Mende, B., Köhler, K., Oross, K., Bondár, M., Marton, T., Osztás, A., Jakucs, J., Paluch, T., Horváth, F., Csengeri, P., Koós, J.,

Sebők, K., Anders, A., Raczky, P., Regenye, J., Barna, J. P., Fábián, S., Serlegi, G., Toldi, Z., Gyöngyvér Nagy, E., Dani, J., Molnár, E., Pálfi, G., Márk, L., Melegh, B., Bánfai, Z., Domboróczki, L., Fernández-Eraso, J., Antonio Mujika-Alustiza, J., Alonso Fernández, C., Jiménez Echevarría, J., Bollongino, R., Orschiedt, J., Schierhold, K., Meller, H., Cooper, A., Burger, J., Bánffy, E., Alt, K. W., Lalueza-Fox, C., Haak, W., and Reich, D. (2017). Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*, 551(7680):368–372.

[67] Logan, A. C. and Jacka, F. N. (2014). Nutritional psychiatry research: an emerging discipline and its intersection with global urbanization, environmental challenges and the evolutionary mismatch. *Journal of Physiological Anthropology*, 33(1):22.

[68] Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., and Berger, B. (2013). Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, 193(4):1233–1254.

[69] Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet*, 93(2):278–288.

[70] Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511.

[71] Marjoram, P. and Wall, J. D. (2006). Fast" coalescent" simulation. *BMC genetics*, 7(1):1–9.

[72] Marnetto, D., Pärna, K., Läll, K., Molinaro, L., Montinaro, F., Haller, T., Metspalu, M., Mägi, R., Fischer, K., and Pagani, L. (2020). Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nature Communications*, 11(1):1628.

[73] Martiniano, R., Cassidy, L. M., Ó'Maoldúin, R., McLaughlin, R., Silva, N. M., Manco, L., Fidalgo, D., Pereira, T., Coelho, M. J., Serra, M., Burger, J., Parreira, R., Moran, E., Valera, A. C., Porfirio, E., Boaventura, R., Silva, A. M., and Bradley, D. G. (2017). The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLOS Genetics*, 13(7).

[74] Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olalde, I., Broomandkhoshbacht, N., Candilio, F., Cheronet, O., Fernandes, D., Ferry, M., Gamarra, B., Fortes, G. G., Haak, W., Harney, E., Jones, E., Keating, D., Krause-Kyora, B., Kucukkalipci, I., Michel, M., Mittnik, A., Nägele, K., Novak, M., Oppenheimer, J., Patterson, N., Pfrengle, S., Sirak, K., Stewardson, K., Vai, S., Alexandrov, S., Alt, K. W., Andreescu, R., Antonović, D., Ash, A., Atanassova, N., Bacvarov, K., Gusztáv, M. B., Bocherens, H., Bolus, M., Boroneanţ, A., Boyadzhiev, Y., Budnik, A., Burmaz, J., Chohadzhiev, S., Conard, N. J., Cottiaux, R., Čuka, M., Cupillard, C., Drucker, D. G., Elenski, N., Francken, M., Galabova, B., Ganetsovski, G., Gély, B., Hajdu, T., Handzhyiska, V., Harvati, K., Higham, T., Iliev, S., Janković, I., Karavanić, I., Kennett, D. J., Komšo, D., Kozak, A., Labuda, D., Lari, M., Lazar, C., Leppek, M., Leshtakov, K., Vetro, D. L., Los, D., Lozanov, I., Malina, M., Martini, F., McSweeney, K., Meller, H., Menđušić, M., Mirea, P., Moiseyev, V., Petrova, V., Price, T. D., Simalcsik,

A., Sineo, L., Šlaus, M., Slavchev, V., Stanev, P., Starović, A., Szeniczey, T., Talamo, S., Teschler-Nicola, M., Thevenet, C., Valchev, I., Valentin, F., Vasilyev, S., Veljanovska, F., Venelinova, S., Veselovskaya, E., Viola, B., Virag, C., Zaninović, J., Zäuner, S., Stockhammer, P. W., Catalano, G., Krauß, R., Caramelli, D., Zariņa, G., Gaydarska, B., Lillie, M., Nikitin, A. G., Potekhina, I., Papathanasiou, A., Borić, D., Bonsall, C., Krause, J., Pinhasi, R., and Reich, D. (2018). The genomic history of southeastern Europe. *Nature*, 555(7695):197–203.

[75] Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J. L., de Castro, J. B., Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M. A. R., Roodenberg, J., Vergès, J. M., Krause, J., Cooper, A., Alt, K. W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R., and Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503.

[76] Mathieson, I. and Scally, A. (2020). What is ancestry? *PLOS Genetics*, 16(3):e1008624–.

[77] Mathieson, S. and Mathieson, I. (2018). FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution*, 35(12):2957–2970.

[78] Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, 116(4):362–371.

[79] McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654.

[80] McVean, G. A. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393.

[81] Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic Maps of Human Gene Frequencies in Europeans. *Science*, 201(4358):786–792.

[82] Mittnik, A., Wang, C.-C., Pfrengle, S., Daubaras, M., Zariņa, G., Hallgren, F., Allmäe, R., Khartanovich, V., Moiseyev, V., Tõrv, M., Furtwängler, A., Andrades Valtueña, A., Feldman, M., Economou, C., Oinonen, M., Vasks, A., Balanovska, E., Reich, D., Jankauskas, R., Haak, W., Schiffels, S., and Krause, J. (2018). The genetic prehistory of the Baltic Sea region. *Nat Commun*, 9(1):442.

[83] Mondal, M., Bertranpetit, J., and Lao, O. (2019). Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, 10(1):246.

[84] Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A. L., and Reich, D. (2011). The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLOS Genetics*, 7(4).

[85] Morgunova, N. L. and Khokhlova, O. S. (2013). Chronology and Periodization of the Pit-Grave Culture in the Region Between the Volga and Ural Rivers Based on Radiocarbon Dating and Paleopedological Research. *Radiocarbon*, 55(3):1286–1296.

[86] Myers, S., Fefferman, C., and Patterson, N. (2008). Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73(3):342–348.

[87] NEEL, J. V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet*, 14(4):353–362.

[88] Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res*, 15(11):1566–1575.

[89] Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., Altena, E., Lipson, M., Lazaridis, I., Harper, T. K., Patterson, N., Broomandkhoshbacht, N., Diekmann, Y., Faltyskova, Z., Fernandes, D., Ferry, M., Harney, E., de Knijff, P., Michel, M., Oppenheimer, J., Stewardson, K., Barclay, A., Alt, K. W., Liesau, C., Ríos, P., Blasco, C., Miguel, J. V., García, R. M., Fernández, A. A., Bánffy, E., Bernabò-Brea, M., Billoin, D., Bonsall, C., Bonsall, L., Allen, T., Büster, L., Carver, S., Navarro, L. C., Craig, O. E., Cook, G. T., Cunliffe, B., Denaire, A., Dinwiddy, K. E., Dodwell, N., Ernée, M., Evans, C., Kuchařík, M., Farré, J. F., Fowler, C., Gazenbeek, M., Pena, R. G., Haber-Uriarte, M., Haduch, E., Hey, G., Jowett, N., Knowles, T., Massy, K., Pfrengle, S., Lefranc, P., Lemercier, O., Lefebvre, A., Martínez, C. H., Olmo, V. G., Ramírez, A. B., Maurandi, J. L., Majó, T., McKinley, J. I., McSweeney, K., Mende, B. G., Modi, A., Kulcsár, G., Kiss, V., Czene, A., Patay, R., Endrődi, A., Köhler, K., Hajdu, T., Szeniczey, T., Dani, J., Bernert, Z., Hoole, M., Cheronet, O., Keating, D., Velemínský, P., Dobeš, M., Candilio, F., Brown, F., Fernández, R. F., Herrero-Corral, A.-M., Tusa, S., Carnieri, E., Lentini, L., Valenti, A., Zanini, A., Waddington, C., Delibes, G., Guerra-Doce, E., Neil, B., Brittain, M., Luke, M., Mortimer, R., Desideri, J., Besse, M., Brücken, G., Furmanek, M., Hałuszko, A., Mackiewicz, M., Rapiński, A., Leach, S., Soriano, I., Lillios, K. T., Cardoso, J. L., Pearson, M. P., Włodarczak, P., Price, T. D., Prieto, P., Rey, P.-J., Risch, R., Rojo Guerra, M. A., Schmitt, A., Serralongue, J., Silva, A. M., Smrčka, V., Vergnaud, L., Zilhão, J., Caramelli, D., Higham, T., Thomas, M. G., Kennett, D. J., Fokkens, H., Heyd, V., Sheridan, A., Sjögren, K.-G., Stockhammer, P. W., Krause, J., Pinhasi, R., Haak, W., Barnes, I., Lalueza-Fox, C., and Reich, D. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*, 555(7695):190–196.

[90] Padhukasahasram, B. (2014). Inferring ancestry from population genomic data and its applications. *Front Genet*, 5:204.

[91] Pasaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–21.

[92] Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J., and Reich, D. (2004). Methods for high-density admixture mapping of disease genes. *Am J Hum Genet*, 74(5):979–1000.

[93] Patterson, N., Isakov, M., Booth, T., Büster, L., Fischer, C.-E., Olalde, I., Ringbauer, H., Akbari, A., Cheronet, O., Bleasdale, M., Adamski, N., Altena, E., Bernardos, R., Brace, S., Broomandkhoshbacht, N., Callan, K., Candilio, F., Culleton, B., Curtis, E., Demetz, L., Carlson, K. S. D., Edwards, C. J., Fernandes, D. M., Foody, M. G. B., Freilich, S., Goodchild, H., Kearns, A., Lawson, A. M., Lazaridis, I., Mah, M., Mallick, S., Mandl, K., Micco, A., Michel, M., Morante, G. B., Oppenheimer, J., Özdoğan, K. T., Qiu, L., Schattke, C., Stewardson, K., Workman, J. N., Zalzala, F., Zhang, Z., Agustí, B., Allen, T., Almássy, K., Amkreutz, L., Ash, A., Baillif-Ducros, C., Barclay, A., Bartosiewicz, L., Baxter, K., Bernert, Z., Blažek, J., Bodružić, M., Boissinot, P., Bonsall, C., Bradley, P., Brittain, M., Brookes, A., Brown, F., Brown, L., Brunning, R., Budd, C., Burmaz, J., Canet, S., Carnicero-Cáceres, S., Čaušević-Bully, M., Chamberlain, A., Chauvin, S., Clough, S., Čondić, N., Coppa, A., Craig, O., Črešnar, M., Cummings, V., Czifra, S., Danielisová, A., Daniels, R., Davies, A., de Jersey, P., Deacon, J., Deminger, C., Ditchfield, P. W., Dizdar, M., Dobeš, M., Dobisíková, M., Domboróczki, L., Drinkall, G., Đukić, A., Ernée, M., Evans, C., Evans, J., Fernández-Götz, M., Filipović, S., Fitzpatrick, A., Fokkens, H., Fowler, C., Fox, A., Gallina, Z., Gamble, M., González Morales, M. R., González-Rabanal, B., Green, A., Gyenesei, K., Habermehl, D., Hajdu, T., Hamilton, D., Harris, J., Hayden, C., Hendriks, J., Hernu, B., Hey, G., Horňák, M., Ilon, G., Istvánovits, E., Jones, A. M., Kavur, M. B., Kazek, K., Kenyon, R. A., Khreisheh, A., Kiss, V., Kleijne, J., Knight, M., Kootker, L. M., Kovács, P. F., Kozubová, A., Kulcsár, G., Kulcsár, V., Le Pennec, C., Legge, M., Leivers, M., Loe, L., López-Costas, O., Lord, T., Los, D., Lyall, J., Marín-Arroyo, A. B., Mason, P., Matošević, D., Maxted, A., McIntyre, L., McKinley, J., McSweeney, K., Meijlink, B., Mende, B. G., Menđušić, M., Metlička, M., Meyer, S., Mihovilić, K., Milasinovic, L., Minnitt, S., Moore, J., Morley, G., Mullan, G., Musilová, M., Neil, B., Nicholls, R., Novak, M., Pala, M., Papworth, M., Paresys, C., Patten, R., Perkić, D., Pesti, K., Petit, A., Petriščáková, K., Pichon, C., Pickard, C., Pilling, Z., Price, T. D., Radović, S., Redfern, R., Resutík, B., Rhodes, D. T., Richards, M. B., Roberts, A., Roefstra, J., Sankot, P., Šefčáková, A., Sheridan, A., Skae, S., Šmolíková, M., Somogyi, K., Somogyvári, Á., Stephens, M., Szabó, G., Szécsényi-Nagy, A., Szeniczey, T., Tabor, J., Tankó, K., Maria, C. T., Terry, R., Teržan, B., Teschler-Nicola, M., Torres-Martínez, J. F., Trapp, J., Turle, R., Ujvári, F., van der Heiden, M., Veleminsky, P., Veselka, B., Vytlačil, Z., Waddington, C., Ware, P., Wilkinson, P., Wilson, L., Wiseman, R., Young, E., Zaninović, J., Žitňan, A., Lalueza-Fox, C., de Knijff, P., Barnes, I., Halkon, P., Thomas, M. G., Kennett, D. J., Cunliffe, B., Lillie, M., Rohland, N., Pinhasi, R., Armit, I., and Reich, D. (2022). Large-scale migration into Britain during the Middle to Late Bronze Age. *Nature*, 601(7894):588–594.

[94] Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics*, 192(3):1065–1093.

[95] Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences*, 111(7):2632–2637.

[96] Pool, J. E. and Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719.

[97] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

[98] Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLOS Genetics*, 5(6).

[99] Racimo, F., Woodbridge, J., Fyfe, R. M., Sikora, M., Sjögren, K.-G., Kristiansen, K., and Vander Linden, M. (2020). The spatiotemporal spread of human migrations during the European Holocene. *Proceedings of the National Academy of Sciences*, 117(16):8989–9000.

[100] Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T. W. J., Orlando, L., Metspalu, E., Karmin, M., Tambets, K., Rootsi, S., Mägi, R., Campos, P. F., Balanovska, E., Balanovsky, O., Khusnutdinova, E., Litvinov, S., Osipova, L. P., Fedorova, S. A., Voevoda, M. I., DeGiorgio, M., Sicheritz-Ponten, T., Brunak, S., Demeshchenko, S., Kivisild, T., Villems, R., Nielsen, R., Jakobsson, M., and Willerslev, E. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 505(7481):87–91.

[101] Ralph, P., Thornton, K., and Kelleher, J. (2020). Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics*, 215(3):779–797.

[102] Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genetics*, 10(5).

[103] Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J., and Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, 53(1):120–126.

[104] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837.

[105] Sanchez, T., Cury, J., Charpiat, G., and Jay, F. (2021). Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660.

[106] Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008). Estimating local ancestry in admixed populations. *Am J Hum Genet*, 82(2):290–303.

[107] Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925.

[108] Schraiber, J. G. and Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740.

[109] Schroeder, H., Margaryan, A., Szmyt, M., Theulot, B., Włodarczak, P., Rasmussen, S., Gopalakrishnan, S., Szczepanek, A., Konopka, T., Jensen, T. Z. T., Witkowska, B., Wilk, S., Przybyła, M. M., Pospieszny, Ł., Sjögren, K.-G., Belka, Z., Olsen, J., Kristiansen, K., Willerslev, E., Frei, K. M., Sikora, M., Johannsen, N. N., and Allentoft, M. E. (2019). Unraveling ancestry, kinship, and violence in a Late Neolithic mass grave. *Proceedings of the National Academy of Sciences*, 116(22):10705–10710.

[110] Segurel, L., Guarino-Vignon, P., Marchi, N., Lafosse, S., Laurent, R., Bon, C., Fabre, A., Hegay, T., and Heyer, E. (2020). Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. *PLoS Biol*, 18(6):e3000742.

[111] Seldin, M. F., Pasaniuc, B., and Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nat Rev Genet*, 12(8):523–528.

[112] Slatkin, M. (2008). Linkage disequilibrium —understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485.

[113] Slatkin, M. and Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, 113(23):6380–6387.

[114] Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1):23–35.

[115] Souilmi, Y., Tobler, R., Johar, A., Williams, M., Grey, S. T., Schmidt, J., Teixeira, J. C., Rohrlach, A., Tuke, J., Johnson, O., Gower, G., Turney, C., Cox, M., Huber, C. D., and Cooper, A. (2020). Ancient human genomes reveal a hidden history of strong selection in Eurasia. *bioRxiv*.

[116] Speidel, L., Cassidy, L., Davies, R. W., Hellenthal, G., Skoglund, P., and Myers, S. R. (2021). Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. *Molecular Biology and Evolution*, 38(9):3497–3511.

[117] Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329.

[118] Stern, A. J., Wilton, P. R., and Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLOS Genetics*, 15(9).

[119] Suarez-Pajes, E., Díaz-de Usera, A., Marcelino-Rodríguez, I., Guillen-Guio, B., and Flores, C. (2021). Genetic ancestry inference and its application for the genetic mapping of human diseases. *International Journal of Molecular Sciences*, 22(13):6962.

[120] Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

[121] Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet*, 79(1):1–12.

[122] Terhorst, J., Kamm, J. A., and Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet*, 49(2):303–309.

[123] Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLOS Biology*, 4(3).

[124] Wang, K., Mathieson, I., O'Connell, J., and Schiffels, S. (2020). Tracking human population structure through time from whole genome sequences. *PLOS Genetics*, 16(3).

[125] Wiuf, C. and Hein, J. (1999). Recombination as a Point Process along Sequences. *Theoretical Population Biology*, 55(3):248–259.

[126] Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354.

[127] Y. C. Brandt, D., Wei, X., Deng, Y., Vaughn, A. H., and Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, 221(1). iyac044.

[128] Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*, 15(5):568–573.

[129] Zhang, B. C., Biddanda, A., and Palamara, P. F. (2021). Biobank-scale inference of ancestral recombination graphs enables genealogy-based mixed model association of complex traits. *bioRxiv*, page 2021.11.03.466843.

[130] Zhou, Y., Yuan, K., Yu, Y., Ni, X., Xie, P., Xing, E. P., and Xu, S. (2017). Inference of multiple-wave population admixture by modeling decay of linkage disequilibrium with polynomial functions. *Heredity*, 118(5):503–510.