Genealogy estimation for thousands of samples

Leo Speidel
Mansfield College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2019

Acknowledgements

I would like to express my greatest gratitude towards Simon, whose natural, unrivalled enthusiasm from day one convinced me to do my DPhil with him. I have been extremely fortunate to have had such an exciting research topic, accompanied by a continuous stream of his brilliant ideas, as well as his patience while I tried to decipher their true meaning. Our many meetings were extremely fruitful, including our regular off-topic chats on Brexit (naturally less enjoyable but still important), his many attempts at explaining the rules of Cricket to me (which I still don't understand), and our future plans of reviving the Oxford Dodo. I am also very grateful to all members of the Myers group and my office mates; in particular to Daniel Wells who has had to put up with my endless questions on biology. Our excursion to New York City following Probgen 2018 is a great memory from this DPhil.

I would also like to thank friends and family who have supported me for so many years. In particular, I am grateful to my sister Saya, who is soon to become a real doctor and to my grandparents for their support. While home has been far for the last four years, whenever I did manage to come home, my parents always made sure to recharge my batteries with exceptional food and hospitality. This has admittedly always triggered a small amount of jealousy about their lives in Tokyo, but it should be said that food in Oxford has been good as well. Here, I need to especially acknowledge the never ending stream of generous parcels from my grandparents filled with Japanese delicatessen, and also Seoul Plaza for their constant supply of essential Japanese food ingredients which has made life in Oxford a small Japanese Oasis.

Last, but by no means least, I am extremely grateful to Clare, who has been there for me throughout this DPhil and has supported me especially in some of the more difficult periods. Our joint DPhil journey in Oxford has been the best time of my life so far. She had to proof-read almost every word I ever wrote in my DPhil and I hope that she won't skip these.

Abstract

Abstract

A key and fundamental concept that captures our shared genetic history is the genealogy, which traces the genetic relationships of present-day individuals to their most-recent common ancestors. Knowledge of the genealogy would, in principle, capture all evolutionary forces that modified the genetic material ancestral to our DNA, and would hence simplify - and enhance - many inference problems about past demography and evolution. Despite their importance, estimation of genealogies has remained unsolved even for moderately sized data sets, with existing methods unable to handle sample sizes beyond a few hundred samples, yet modern data sets often exceed tens of thousands of samples.

In this thesis, I present a method, Relate, that estimates such genealogies for thousands of samples. I demonstrate on a variety of population genetic applications that Relate-based inferences improve in accuracy, resolution, or statistical power on state-of-the-art alternatives. I then reconstruct the genealogy of 2478 humans from 26 populations. I infer historical population sizes and population split times with higher resolution than previously possible and identify highly diverged lineages, reflecting Neanderthal and Denisovan introgression in non-Africans, and unknown events in Africans. I report regions that show evidence of being under strong positive selection that were previously unreported and identify multi-allelic traits likely to be under selection. I additionally apply Relate to 50 wild mice sampled in France, India, and Taiwan and demonstrate that the estimated genealogies contain rich information about their demographic history, mutation rate trends consistent with GC biased gene conversion, as well as strong indications of selective sweeps in each population.

Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in the text.

Leo Speidel

27th September 2019

Contents

List of Figures

List of Tables

1 Introduction

Contents

The development of next-generation DNA sequencing technologies has led to the sequencing of thousands of genomes for many species. The largest data sets have now sequenced >100,000 human samples, capturing the genetic diversity of modern-day humans on an unprecedented scale. Combined with extensive genomic annotations, such as associations of mutations or genes with phenotypes, they provide a powerful resource for decoding the molecular blueprints of living organisms.

The DNA of modern people not only inform us about present-day diversity; they also contain rich information about our genetic past. For instance, a variety of tools have detected genetic structure on very fine scales (e.g., within the British Isles) that is in good agreement with independent evidence, such as geopolitical separation over generations [87, 100, 123]. Admixture events of ancestral populations have been identified and dated [3, 37, 68]. Genetic and anthropological evidence clearly trace the origin of *homo sapiens* to the African continent [20, 46] and

deep population size bottlenecks have been found in populations that subsequently migrated out of Africa [88, 128, 136, 155].

The complete evolutionary record of genetic material ancestral to our DNA is captured by a genealogy. Looking backwards in time, we can relate present-day samples through their most-recent common ancestors (MRCAs), some of which may be millions of years old. Genealogical trees organise these MRCA relationships using historical coalescence, recombination, and mutation events. They inform us about how our DNA evolved through time, capturing ideally all information contained in observed DNA about the sample's genetic past. If known, genealogy-based inferences have the potential to be maximally powerful.

The reconstruction of genealogies from present-day samples is a fundamental problem in population genetics. The numerous existing methods have been constrained by the enormous number of possible genealogies and are applicable to no more than tens (or at most a few hundred) of samples. Modern data sets can comprise many thousands of samples and are rapidly increasing in size. As a result, genealogy estimation has become infeasible and is often avoided. Instead, efficient algorithms based on representations of the data in lower dimensions [100, 113] and techniques that model only some properties of the data, such as allele frequencies [123], and downsampling [69, 88], have been popular. Many of these approaches indirectly measure changes in the genealogy, by first predicting how genealogies are changed and then translating this to how we expect these changes to affect present-day variation patterns. Knowledge of the underlying genealogy would therefore remove the need of indirectly (and often imperfectly) measuring effects of the genealogy on a given summary statistic and would immediately simplify and improve most existing inference schemes.

In this thesis, I propose a scalable method, Relate, to estimate genealogies for thousands of sampled genomes. Relate combines a hidden Markov model (HMM) similar to that proposed by Li and Stephens [90], and a hierarchical clustering algorithm. The HMM models mutation and recombination events by reconstructing one sample as a mosaic of other samples in the data set. It then outputs a

distance matrix at each locus containing information about the order in which sampled haplotypes coalesced with ancestors of other sampled haplotypes. Using these distance matrices, Relate constructs rooted binary trees along the genome, where each binary tree describes the genetic history of a subsection of the genome. Because the distance matrix is position specific, these binary trees adapt to changes in local genetic ancestry that arise due to recombination. Under certain idealistic approximations of real scenarios, such as if the "infinite-sites" model is satisfied so that each site has mutated at most once in history, our approach is guaranteed to generate genealogies exactly producing the observed data, in the limiting cases where either there is no recombination, where the recombination rate is very large, or where all recombination occurs in intense widely spaced hotspots. For non-zero recombination rates, it outperforms other methods in both runtime and accuracy across a variety of simulation scenarios.

We explore different applications of Relate. We develop algorithms for estimating mutation and coalescence rates through time, as well as a tree-based statistic for detecting positive selection. For these applications, we demonstrate that a genealogy-based inference approach is both powerful and flexible and has the benefit of being a coherent, self-contained framework in which all inferences are derived from the same genealogy.

To demonstrate the utility of our methods, we apply Relate to 4956 haplotypes of the 1000 Genomes Project data set [1, 150]. We obtain age estimates for almost all biallelic SNPs and their frequencies through time. We estimate population sizes of all 26 populations in the data set and their split times using cross-coalescence rates between populations. In agreement with a previous study, we identify a remarkable increase in the mutation rate of TCC to TTC mutations, which we date at around 5,000 to 30,000 years ago and which is strongest in Europeans [65, 66]. We then demonstrate that the estimated genealogy contains a strong signal of introgression between Neanderthals and modern humans in Eurasia. We also detect a weaker introgression signal between modern East and Southern Asians and Denisovans. Finally, we identify genomic regions that have potentially evolved under

strong positive selection which were previously unreported. We find a remarkable enrichment of SNPs identified in genome-wide association studies (GWAS) among such potential targets of selection and study directional polygenic adaptation using SNP-trait associations identified in GWAS.

Additionally, we apply Relate to 50 wild mice sampled in France, India, and Taiwan. Compared to humans, these mice have a lower per generation mutation rate [61] and computational phasing is expected to be less accurate, due to the lack of reference panels. Both of these factors potentially complicate genealogy estimation. Nevertheless, we demonstrate that Relate-estimated genealogies are rich in information about the genetic past of these mice, including their demographic history, split times, and introgression, as well as signatures of GC biased gene conversion in mutation rate estimates and evolutionary adaptation.

1.1 Ancestral recombination graphs

A genealogy records historical events acting on genomes ancestral to our DNA, including mutations and recombinations. Mutation are the source of genetic variation and typically replace one base-pair by another, and may be induced by replication errors, environmental damage, or an interaction of the two. In humans, we typically inherit 10 to 20 novel mutations from the maternal lineage, and 30 to 60 novel mutations from the paternal lineage in every generation, with parental age being a major factor influencing the de-novo mutation count [47]. The average per base per generation mutation rate is estimated to be around $1.0 - 2.0 \times 10^{-8}$, depending on the technique used [110, 139]. Mutation rates are highly context dependent; for instance, CpG dinucleotides, which are sites at which a C nucleotide is followed by a G nucleotide, are known to have a 12 fold higher rate of CG to TG changes [153].

Meiotic recombination describes the process by which two homologous copies of the same chromosome cross over, forming new recombined chromosomes that are

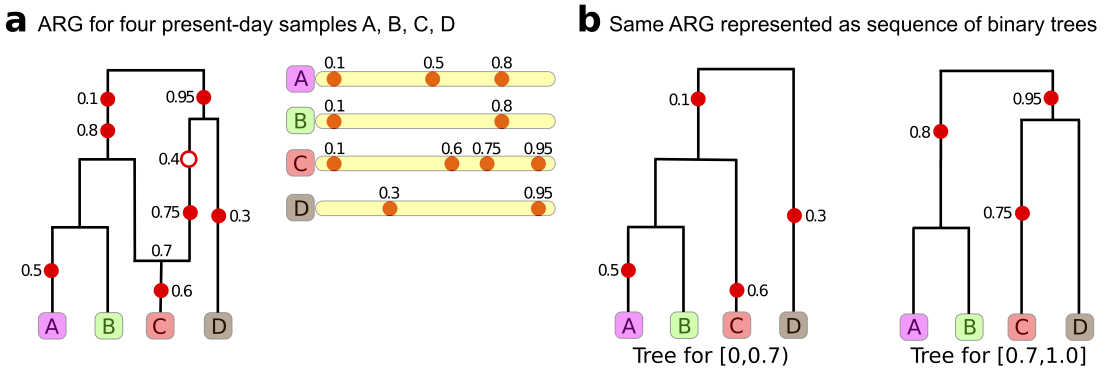


Figure 1.1: Schematic of an ancestral recombination graph (ARG). **a**, Schematic of an ARG for four present-day samples (haplotypes). For this example, we assume a continuous genome with base-pair positions given by numbers in the interval $[0, 1]$. Red circles indicate mutations, with the position of the mutation indicated alongside. The ARG contains one recombination event in the ancestor of sample C, which occurred at position 0.7. At any genomic position < 0.7, genetic material is inherited from the lineage to the left of the recombination event, whereas at any genomic position ≥ 0.7, genetic material is inherited from the lineage to the right. We notice that a mutation at position 0.4 has been lost due to this recombination event and is not observed in samples A – D. **b**, Representation of the same ARG as a sequence of binary trees.

mosaics of the parental chromosomes. A new recombined chromosome will carry a unique mixture of mutations inherited by its two parental chromosomes. A direct consequence of recombination is that any two mutations on the same chromosome may not be inherited together in future generations; more generally, genetic linkage between any two genomic regions, defined as their tendency to be inherited on the same chromosome, fluctuates over time and typically decreases. This has important biological consequences. For instance, an advantageous mutation would otherwise be unable to be separated from potential deleterious background mutations. In any one generation, the overall recombination count is on average 35 in females [26] and 25 in males [151], with typically at least one event required on each pair of sister chromatids to minimise the chance of aneuploidy [42].

In clonally reproducing organisms without recombination, the genealogy of a sample is described by a (binary) tree, with each internal node representing the most-recent common ancestor (MRCA) of samples falling underneath each daughter branch. Recombination causes different genomic regions to be inherited via different sets of ancestors, implying that the MRCA of any two chromosomes generally varies

along the genome; they can be closely related in some parts of the genome and highly diverged in others. The general genealogical history of a sample is therefore captured by an ancestral recombination graph (ARG) [53] (see Fig. 1.1 (a)). An ARG reduces to a binary tree, in the absence of recombination. With recombination, an internal node can have two ancestral nodes, indicating that the node inherited genetic material from its two ancestors that recombined.

At any position along the genome, we can trace ancestral relationships in the ARG to obtain a marginal binary tree describing the local genetic ancestry. An alternative (but not equivalent) representation of a genealogy is therefore a collection of binary trees, where each binary tree represents the genetic ancestry of a subregion without recombination (see Fig. 1.1 (b)). Typically, binary trees describing adjacent subregions are very similar and can be transformed into one another by a single recombination event. Representing ARGs using marginal trees is not lossless, because explicit information on how trees are transformed into one another by recombination, as well as recombination events that do not affect marginal trees are lost.

1.2 Genetic variation data sets contain information about the underlying ARG

In the vast majority of scenarios, we cannot directly observe ARGs. However, genetic variation data sets record polymorphisms within a sample of DNA sequences that can inform us about the true ARG underlying this data. Here, we only focus on single nucleotide polymorphisms (SNPs), which are single base-pair mutations changing an ancestral allele to a derived allele. In humans and many other species, we can generally assume that at most one mutation has occurred at any given position of the genome since the divergence from other species (e.g., chimpanzees for humans). This assumption, known as the infinite sites model [84], is justified by a small average mutation rate in these species. If this assumption is valid

and the data is free of errors, any mutation indicates the existence of a branch with exactly the descendants determined by carriers of the derived allele and so resolves a branch in the genealogy. In the absence of recombination, the genealogy is given by a binary tree and it immediately follows that all mutations can be partially ordered by inclusion [57].

With recombination, the data may not be consistent with a single binary tree. If two SNPs cannot be ordered by inclusion, an assertion also known as the four-gamete test [76], then a recombination event must have occurred between these two SNPs. Violation of the four-gamete test is therefore a sufficient condition for a recombination. It is important to note that in practice, the vast majority of recombination events are undetectable because of insufficient tagging by SNPs. For this reason, the true number of historical recombinations is unobtainable and furthermore, determining the minimum number of recombination events required to construct an ARG consistent with the data is NP-hard [15, 161]. Instead, a series of lower bounds have been developed that first generate an incompatibility matrix storing lower bounds on the number of recombinations between any pairs of SNPs (e.g., using the four-gamete test) and then combine these bounds to obtain an overall bound [76, 104]. These incompatibility matrices have revealed block-like patterns in the number of recombinations along the genome, suggesting a non-uniform distribution of recombination events.

To estimate more accurate maps of recombination rates genome-wide, methods based on full likelihood principles were developed [39, 52]. These methods use importance sampling; they evaluate the likelihood function of recombination maps by first sampling plausible genetic histories under a simpler approximate model, and subsequently reweight each history according to how likely it was in the correct model given any recombination map. Because sampling of genome-wide genealogies conditional on observed data has been computationally infeasible on all but the smallest data sets even when using approximate approaches, these full likelihood approaches have been computationally intractable and instead more efficient methods have been based on composite likelihood schemes [5, 75, 99]. These

methods reduce the complexity of the problem by calculating the likelihood function only for pairs of SNPs and then multiply these across SNP pairs to construct a pseudo-likelihood. Recombination maps estimated in this way have revealed that recombination events tend to cluster within intense hotspots, with 50% of all events occurring in less than 10% of the genome [99, 106]. These hotspots were subsequently identified to be enriched for certain motifs that are recognised by the zinc-finger protein PRDM9, a protein that binds to DNA and recruits the necessary machinery to initiate the recombination process [8, 105, 107].

1.3 Probabilistic models for the genetic history of a sample

Observed genetic variation only partially narrows down the space of possible ARGs with a large number of possible ARGs remaining that may generate the observed data (even in the absence of errors). This is partially because in humans (and in many other species), mutation and recombination rates are of similar magnitude [28], implying that the expected number of mutations on any branch is approximately one. We therefore need ways of aggregating information about the true underlying genealogy from nearby regions, leveraging the fact that a recombination event only partially breaks up haplotype patterns, such that mutations mapping to different but correlated branches of the genealogy are still informative. It is useful to approach this task by defining generative probabilistic models for a sample's genetic history; these are useful for understanding how observed variation patterns arise. These probabilistic models define probability distributions on the space of possible ARGs conditional on the data which we use for their statistical inference.

The Wright-Fisher model can be seen as the foundation of many such models. This model was developed in the 1920s and 1930s by R. A. Fisher [41] and S. Wright [165] to understand how evolutionary processes, such as mutation or

recombination, shape genetic variation. At the core of the Wright-Fisher model is the assumption that genetic variation arises by random forces, such as random mutation and recombination events, acting on the genetic material of individuals. The Wright-Fisher model assumes discrete generations, where individuals are simultaneously replaced by offspring who choose their parent uniformly at random from the previous generation. Generations are assumed to be stochastically independent.

While the Wright-Fisher model describes the evolution of a whole population forwards in time, we can use it to derive a backwards-in-time model describing the genetic history of a random sample drawn from a population. Backwards in time, we are interested in the time it takes for two chromosomes to coalesce. Assuming a constant population size of N_e, the probability for two chromosomes to choose the same parent (i.e., coalesce) in one generation is given by $1/N_\mathrm{e}$. Because generations are independent, the number of generations τ_g until two chromosomes coalesce is given by the geometric distribution

$$P(\tau_\mathrm{g} = k) = \left(1 - \frac{1}{N_\mathrm{e}}\right)^{k-1} \frac{1}{N_\mathrm{e}}. \tag{1.1}$$

By rescaling time to $t = \tau_\mathrm{g}/N_\mathrm{e}$ and considering the limit of an infinitely large population ($N_\mathrm{e} \to \infty$), the time to coalescence is given by an exponential distribution with rate 1.

We can generalise this process to more than two chromosomes. The resulting model, which describes the genetic history of N present-day chromosomes drawn at random from a larger population, is known as the standard coalescent model [85]. In this model, the time to the first coalescence is given by the sum of all pairwise coalescence rates. Therefore, if all pairs coalesce at a rate of 1, the time to the first coalescence has an exponential distribution with rate $\binom{N}{2}$. After the first coalescence, $N-1$ lineages remain, and the time to the next coalescence event has an exponential distribution with rate $\binom{N-1}{2}$. In general, the time between the kth and $k+1$st coalescence has an exponential distribution with rate $\binom{N-k}{2}$.

In the Wright-Fisher model, recombination can be incorporated by allowing offspring to choose two (instead of one) parents with some probability. In the

coalescent, a recombination is represented by a branching event of one lineage into two lineages backwards in time, and occurs at some specified rate [74].

So far, we have assumed a constant population size and uniform probability of choosing a parent in any generation. Both assumptions can be relaxed. For instance, if the size of a population expands back in time, it is less likely for two chromosomes to meet and coalesce, resulting in a lower rate of coalescence in these time periods. To model population structure, we can alter the probability of choosing a parent depending on its affiliation to a population, such that it is more likely to choose a parent in the same population than another population. As a result, the coalescence rate depends on the affiliation of both chromosomes.

1.4 Properties of the coalescent and implications for observed genetic diversity

It is useful to study properties of genealogies generated by the coalescent as these directly emulates properties of observed genetic variation. Here, we focus on a few properties of the coalescent without recombination and a constant population size, as this is the easiest possible case. We note that recombination does not affect marginal trees at single genomic locations, and so the study of the coalescent without recombination is valid for understanding properties of marginal trees.

First, let us consider the time to the most recent common ancestor (TMRCA) of a sample of size N. The TMRCA is given by the sum of times τ_k while k ancestors remain in the tree, which are exponentially distributed with rate parameter $\binom{k}{2}$. By taking the expectation, we obtain

$$E\left[\sum_{k=2}^{N} \tau_k\right] = \sum_{k=2}^{N} E[\tau_k] = \sum_{k=2}^{N} \frac{2}{k(k-1)} = 2\left(1 - \frac{1}{N}\right). \tag{1.2}$$

As we increase the sample size N, the expected TMRCA converges to 2, implying that increasing the sample size is not expected to reveal an arbitrarily deep genetic history.

We can explicitly calculate the probability that an additional sequence coalesces into an existing coalescent tree of N samples while k lineages remain in that tree ($k = 2, \dots N$). Let us define by I_k the indicator random variable that equals 1 if the additional sequence coalesces into the tree while k lineages remain and 0 otherwise. If the additional sequence has not coalesced while $>k$ lineages remained ($I_N = 0, \dots, I_{k+1} = 0$), it will coalesce at rate k while k lineages remain in the tree. By additionally conditioning on τ_k, we therefore have

$$P(I_k = 1 | \tau_k, I_N = 0, \dots, I_{k+1} = 0) = 1 - e^{-k\tau_k}. \tag{1.3}$$

Removing the condition on τ_k, we obtain

$$P(I_k = 1 | I_N = 0, \dots, I_{k+1} = 0) = \int_0^\infty (1 - e^{-k\tau_k}) \binom{k}{2} e^{-\binom{k}{2}\tau_k} d\tau_k = \frac{k}{\binom{k+1}{2}}. \tag{1.4}$$

It follows that the unconditional probability of $I_k = 1$ is given by

$$P(I_k = 1) = \frac{\binom{N}{2}}{\binom{N+1}{2}} \frac{\binom{N-1}{2}}{\binom{N}{2}} \cdots \frac{k}{\binom{k+1}{2}} = \frac{k}{\binom{N+1}{2}}. \tag{1.5}$$

Summing over $k \geq K$ gives the distribution function, representing the probability that the additional sequence coalesces while at least K lineages remain, which is given by

$$P\left(\sum_{\ell=N}^{K} I_\ell = 1\right) = 1 - \frac{\binom{K}{2}}{\binom{N+1}{2}}. \tag{1.6}$$

For large N and $a \in [0, 1]$, Eq. (1.6) can be rewritten as

$$P\left(\sum_{\ell=N}^{\lfloor \sqrt{a}N \rfloor} I_\ell = 1\right) = 1 - a + O\left(\frac{1}{N}\right). \tag{1.7}$$

Choosing, for instance, $a = 0.5$ therefore implies that there is a 50% chance that the additional sequence coalesces with the remaining tree while at least $\sqrt{0.5}N \approx 0.7N$ (or 70% of all) lineages remain. Setting $K = 2$ in Eq. (1.6) gives the probability that the additional sequence coalesces with the remaining tree before all N samples have coalesced and we obtain

$$P\left(\sum_{\ell=N}^{2} I_\ell = 1\right) = 1 - \frac{1}{\binom{N+1}{2}}. \tag{1.8}$$

Equation (1.8) shows that the probability that an additional sequence reveals a deeper genetic history decreases quadratically with N and is already less than 0.02 for $N = 10$ samples. Instead, as Eq. (1.7) demonstrates, it is far more likely that an additional sequence will coalesce further down in the tree, which is why an increased sample size is expected to primarily result in increased resolution of recent ancestry.

Next, we study the shape of coalescent trees. The coalescent defines a probability distribution on the space of rooted binary trees with labelled tips. We might therefore be interested in how a typical coalescent tree looks like, e.g., do these trees tend to look more symmetric, in the sense that all lineages have similar numbers of descendants, or do they look more asymmetric, in the sense that descendant distributions are unequal across different lineages.

While k lineages remain, let us define by $Z_1, Z_2, \ldots, Z_k$ the number of samples subtending each of the k lineages. Then, it is well known that the joint distribution of $Z_1, Z_2, \ldots, Z_k$ is given by the uniform distribution over possible partitions of N samples to k lineages [54], i.e., it is given by

$$P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k) = \frac{1}{\binom{N-1}{k-1}}. \tag{1.9}$$

Equation (1.9) implies that for each k, it is equally likely to observe any number of descendants subtending each lineage, implying that coalescent trees are typically not very balanced. The proof of Eq. (1.9) is done by induction over k. For the base case, we set $k = N$, in which case there is only one possible division of N samples to $k = N$ lineages and $P(Z_1 = 1, Z_2 = 1, \ldots, Z_k = 1) = 1/\binom{N-1}{N-1} = 1$. For the induction step, we assume that Eq. (1.9) is true for $k+1$ and prove that it must then also hold for k. We use that all k lineages are equally likely to be the lineage that increases the number of lineages from k to $k+1$ forwards in time, and therefore

$$\begin{aligned} &P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k) \\ &\quad = \sum_{i=1}^{k} \sum_{a=1}^{z_i - 1} P(Z_1 = z_1, \ldots, \tilde{Z}_{i_1} = a, \tilde{Z}_{i_2} = z_i - a, \ldots, Z_k = z_k) \frac{1}{k}, \end{aligned} \tag{1.10}$$

where $\tilde{Z}_{i_1}$ and $\tilde{Z}_{i_2}$ denote the two daughter lineages of lineage i. By using the induction hypothesis for the case $k+1$, we obtain

$$\begin{aligned} P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k) &= \sum_{i=1}^{k} \sum_{a=1}^{z_i - 1} \frac{1}{\binom{N-1}{k}} \frac{1}{k} \\ &= \frac{N-k}{k} \frac{1}{\binom{N-1}{k}} = \frac{1}{\binom{N-1}{k-1}}. \end{aligned} \tag{1.11}$$

We note that the proof does not require the constant population size assumption, such that Eq. (1.9) is valid regardless of the particular demographic history of the sample.

This has interesting implications on the patterns of genetic diversity we observe. Let us define by $f(b)$ the fraction of SNPs with a derived allele frequency of b. By using Eq. (1.9), the probability that a mutation attains a DAF of b given it arose while k lineages remained is the marginal distribution of Z_1 in Eq. (1.9) and is given by

$$P(\text{DAF } = b | \text{Mutation arises while } k \text{ lineages remain}) = \frac{\binom{N-b-1}{k-2}}{\binom{N-1}{b-1}}. \tag{1.12}$$

The probability that a mutation arose while k lineages remained can be shown to equal

$$P(\text{Mutation arises while } k \text{ lineages remain}) \propto kE[\tau_k] = \frac{2}{k-1}. \tag{1.13}$$

and therefore by multiplying Eqs. (1.12) and (1.13) and summing over k, we obtain

$$f(b) = \frac{1}{H_{N-1}} \sum_{k=2}^{N-b+1} \frac{\binom{N-b-1}{k-2}}{\binom{N-1}{b-1}} \frac{1}{k-1}, \tag{1.14}$$

where $H_{N-1} = \sum_{\ell=1}^{N-1} 1/\ell$ denotes the Harmonic series. We can simplify Eq. (1.14) and obtain

$$f(b) = \frac{1}{H_{N-1}} \left[\frac{1}{b} \frac{1}{\binom{N-1}{b}} \sum_{\ell=b-1}^{N-1} \binom{\ell}{b-1} \right] = \frac{1/b}{H_{N-1}}. \tag{1.15}$$

In other words, the fraction of SNPs at frequency b is proportional to $1/b$, implying for instance that the majority of SNPs we observe are likely to be rare. We note that Eq. (1.15) is only valid if the population size is constant because Eq. (1.13)

assumes a constant size population. If for instance the population size was rapidly expanding forwards in time, younger branches would be elongated, implying that mutations are relatively more likely to occur while many lineages remain resulting in an excess of rare mutations compared to the constant size scenario.

1.5 Existing approaches for estimating genealogies

Numerous methods have been proposed for estimating ARGs, many of which are based on the coalescent. While each method has its own advantages and disadvantages, a common challenge with all existing methods has been their lack of scalability to large modern data sets comprising many thousands of samples. As a result, genealogy estimation using current methods has been infeasible for many applications, and their estimation is usually avoided.

Coalescent-based methods typically implement statistical techniques that sample from the posterior probability of an ARG given the observed data ($P(\text{ARG}|\text{Data})$). Assuming that mutations arise randomly at some rate μ, the likelihood of the data given any particular ARG ($P(\text{Data}|\text{ARG})$) is Poisson distributed and its evaluation is therefore straight forward. Using the coalescent as a prior on ARGs ($P(\text{ARG})$), we can therefore evaluate $P(\text{ARG}|\text{Data})$ up to a normalising constant using Bayes' theorem

$$P(\text{ARG}|\text{Data}) \propto P(\text{Data}|\text{ARG})P(\text{ARG}). \tag{1.16}$$

Equation 1.16 is sufficient to construct Markov-chain Monte Carlo (MCMC) algorithms that sample from $P(\text{ARG}|\text{Data})$. The effectiveness of an MCMC approach depends on how quickly the Markov chain converges to its stationary distribution and how efficiently it subsequently explores the space of likely ARGs. Designing such an algorithm has proved to be very challenging [86, 93].

To construct efficient coalescent-based MCMC samplers, one approach has focused on the representation of ARGs as a sequence of marginal binary trees, where each marginal tree describes a subregion of the genome without recombination (see Section 1.1). This approach is attractive because $P(\text{Data}|\text{ARG})$ in Eq. (1.16) only depends on marginal trees. However, when viewing the coalescent with recombination as a stochastic process on marginal trees that moves along the genome, such that we start with a binary coalescent tree at one end and modify this tree whenever a recombination occurs, it is known that this process is not Markovian [164]. In other words, the probability of a marginal tree depends on all other marginal trees and not just its adjacent marginal trees. This non-Markovian property complicates the sampling of marginal trees. In addition, there are infinitely many ARGs that can give rise to the same set of marginal trees and integrating over all of these ARGs is usually impossible [98]. Therefore, using the coalescent with recombination as an indirect model on marginal trees is impractical.

To overcome these constraints, an alternative approach has focussed on directly defining a model on marginal trees. This model, known as the Sequentially Markovian Coalescent (SMC), can be seen as an approximation of the coalescent with recombination that turns the non-Markovian model into a Markovian one [98]. The SMC constrains the space of possible ARGs by imposing the following rule: if a new recombination event is introduced to a marginal tree, which results in the removal of the branch on which the recombination occurred, this branch has to coalesce back to a branch in the remaining tree. In contrast, the coalescent with recombination allows the recombined branch to potentially coalesce with a branch not present in the marginal tree, i.e., not ancestral to the sample at this position. The benefit of this approximation is that long-range dependencies of marginal trees along the genome are removed. This results in a reduced state space and a Markovian process along the genome, leading to more efficient algorithms. Despite these advances, the state-of-the-art method sampling from the SMC model, ARGweaver, scales to at most a few hundred samples [126].

Alternatives to coalescent based statistical inference of genealogies commonly employ parsimony principles, with constructing an ARG carrying the minimum number of required recombination events being the most obvious. This line of research was pioneered by Hein and colleagues, who proposed a dynamic programming approach to find a parsimonious genealogical history [67, 144]. In their approach, a distance measure corresponding to the minimum number of recombinations (or subtree-prune-and-regraft opertations) needed to transform one tree into another tree is defined [145]. They then consider sets W_i of rooted binary trees with fixed order of coalescence events that are consistent with site i and select a sequence of trees contained in $(W_1, \ldots, W_S)$ (S being the number of segregating sites) that minimises the overall distance. The exact version of this algorithm finds the minimum number of recombination events, however any such algorithm is known to be NP hard [15, 161]. Other more heuristic methods (not all based on parsimony principles) work in polynomial time but these too are often slow and relatively poor in accuracy when compared to methods sampling from coalescent models [79, 103, 166].

In summary, inference under the coalescent is complicated by the structure of the model, uncertainty over the correct genealogy conditional on observed data, and the large resulting space of possible sample histories. Inference using more heuristic approximations often reduce in accuracy while scalability issues prevail. Striking the right balance between principled modelling of genetic history and computational short-cuts appears difficult. In this thesis we present one solution to this problem. Our method is based on principled ideas with theoretical guarantees on accuracy in idealistic scenarios, combined with heuristics for computational efficiency and robustness to errors in the data.

1.6 Citation and implementation of Relate

All analyses presented in this thesis have been conducted by myself under the supervision of Prof. Simon Myers, unless explicitly stated otherwise. Large parts of this thesis, including the method Relate and it's applications to the 1000 Genomes Project data set, are presented in Ref. [147]. An implementation of Relate along with a documentation and toy data set are available at `https://myersgroup.github.io/relate/`. This implementation is written in C++ and is freely available for academic use. An R package for parsing output files generated by Relate and an implementation of the polygenic selection test discussed in Chapter 6 is available at `https://github.com/leospeidel/relater`. Relate-estimated coalescence rates, allele ages, and selection p-values for the 1000 Genomes Project are available from `https://doi.org/10.5281/zenodo.3234689`.

1.7 Thesis overview

The remainder of this thesis is structured as follows. In Chapter 2, I describe our method, Relate, for estimating genealogies and in Chapter 3 I describe a method for jointly fitting historical population sizes and branch lengths. I then evaluate the performance of Relate on simulated data in Chapter 4. Following this, in Chapter 5, I apply Relate to the 1000 Genomes Project data set to obtain a genealogy for 2478 samples from 26 populations. In Chapter 6, I develop a statistic for detecting evidence of positive natural selection using estimated genealogies and investigate selection on single loci, as well as polygenic traits. In Chapter 7, I apply Relate to 50 wild mice from France, India, and Taiwan and demonstrate that the methods developed in this thesis are not specific to humans. Finally, in Chapter 8, I conclude the thesis with a discussion and list possible extensions of the work presented.

2

A method for estimating genealogies for thousands of samples

Contents

In this chapter, I introduce Relate, our method for estimating genealogies for thousands of samples. Relate is a highly modular method combining multiple steps; each step is introduced in a separate section of this chapter.

Relate estimates the genealogy as a sequence of rooted binary trees, where each tree is a point estimate of the genealogy in a subregion of the genome. This representation is an approximation of an Ancestral Recombination Graph (ARG), represented as a sequence of binary trees [74] (see Section 1.1). In particular, we do not enforce constraints on tree topology and branch lengths induced by adjacent

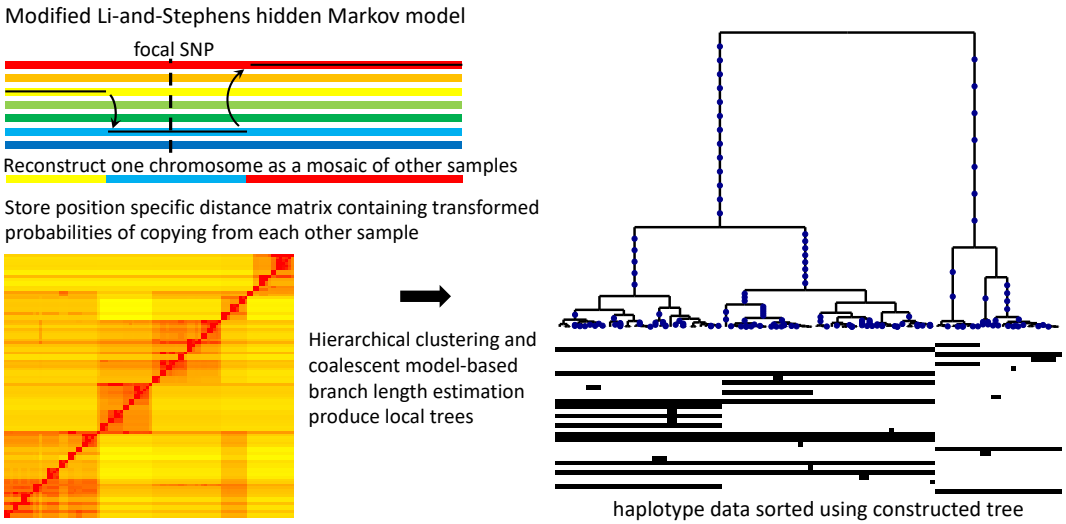


Figure 2.1: Overview of Relate.
Our method applies a version of the Li-and-Stephens hidden Markov model [90], modified to take ancestral and derived states into account, to calculate at a focal SNP (dotted vertical line) a position-specific distance matrix d (bottom left). Each entry d_{ij} of this matrix stores the rescaled log-likelihood of generating haplotype i by copying from haplotype j, which can be interpreted as the number of mutations carried by i, but not by j, locally around the focal SNP. Our tree builder uses the resulting inferred distance matrix to coalesce haplotypes (right-hand side). After mapping mutations to their corresponding branches, we estimate branch lengths using an MCMC algorithm that employs a coalescent prior model.

trees, such as enforcing identical subtrees to have identical branch lengths. By not enforcing these global constraints, Relate can be parallelised over hundreds of parallel threads, with each thread reconstructing the genealogy of a different subregion of the genome. Fig. 2.1 shows a schematic overview of Relate.

2.1 Overview

To illustrate how Relate estimates a genealogy from genetic variation data, it is useful to start with the no-recombination scenario. In this case, a single tree describes the genealogy of the whole genome. We define the number of *derived* mutations $d(i, j)$ as the number of mutations carried by haplotype i and not by haplotype j. Notice that $d(i, j) \neq d(j, i)$. Assuming that every mutation happened

exactly once, we can determine the order in which haplotype i coalesced with other haplotypes by ordering them in ascending order of derived mutations $d(i,j)$ $(j = 1, \ldots, i-1, i+1, \ldots, N)$. Once we know the relative order of coalescences, we reconstruct the tree topology using the hierarchical clustering algorithm described in Section 2.2. The resulting tree topology is guaranteed to be consistent with the truth in the sense that the constructed tree is a subtree of the gene tree describing the data (see Section 2.2.3). An example is shown in Fig. 2.2 **a**.

In the presence of recombination, the relative order of coalescences changes along the genome. We apply a modified version of a Li-and-Stephens type hidden Markov model (HMM) [90] to calculate a local number of derived mutations $d(i,j;\ell)$ at every SNP ℓ (see Section 2.3). We modified the Li-and-Stephens HMM to take ancestral and derived states into account, which is necessary for $d(i,j;\ell)$ to converge to $d(i,j)$ in the limit of no recombination. We use $d(i,j;\ell)$ to reorder haplotypes at every SNP and apply the tree building algorithm to reestimate tree topology. Our method builds trees that are consistent with the truth if $d(i,j;\ell)$ orders haplotypes correctly. This is guaranteed for a recombination map consisting of zero and infinite recombination rates, which can be seen as a limit case of a hotspot recombination map (Section 2.2.3).

It is computationally inefficient, but possible, to reestimate tree topology at every SNP, because trees are unchanged if no recombination event occurred between SNPs. Instead, Relate initially estimates the tree topology at the first SNP of the 5' end of a chromosome. It then only reestimates the tree topology if a mutation cannot be uniquely *mapped* to a branch of the tree describing the previous SNP or is *flipped*. A mutation is mapped to the branch for which the descendants coincide with the carriers of the alternative allele. Such a branch exists as long as the mutation occurred exactly once in human history and the estimated tree topology is correct. To be robust to errors in the data and the inferred tree, we relax this requirement such that the descendants of the branch only have to approximately coincide with the carriers of the mutation (see Section 2.2.2 for details). A mutation is potentially flipped, if it maps to a branch only after reinterpreting non-carriers as carriers

and vice versa. While this introduces a small (rightward) bias in the placement of tree changes relative to true recombination points, we note that this bias does not propagate along the genome because marginal trees are constructed using the distance matrix for the genomic position at which the tree topology is reestimated.

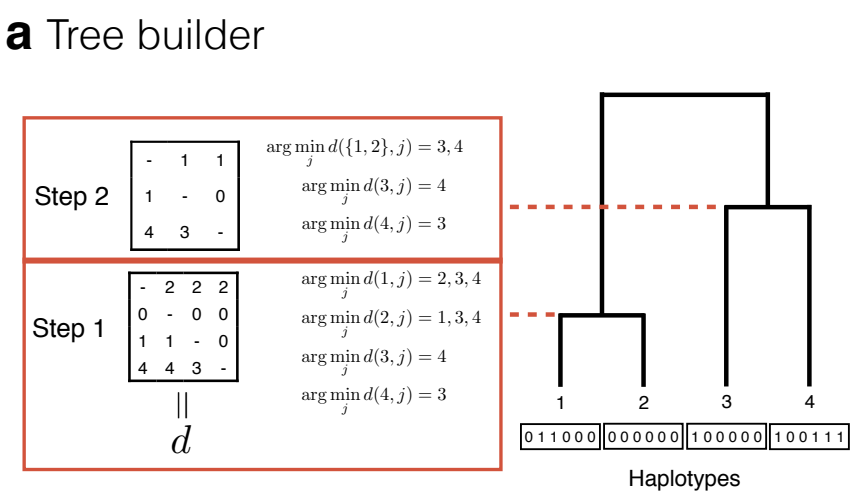
Finally, once tree topologies are estimated, we estimate the branch lengths of each tree using an MCMC approach with a coalescent prior (see Section 2.4.2, Figure 2.2 **b**). We developed an algorithm that jointly estimates coalescence rates and branch lengths if historical coalescence rates are unknown (see Chapter 3).

The backbone of the Relate approach is based on principled and well-tested concepts, such as the Li-and-Stephens HMM and coalescent theory. These guarantee that in idealistic error-free scenarios, we obtain desirable limiting behaviours. However, to maintain robustness to various types of errors in the data, we included a few heuristic steps, as shown in Figure 2.3**a**. These ensure, empirically, that genealogies remain accurate in real scenarios, which we will explore using simulations in Chapter 4.

2.2 Tree builder

In this section, we describe how we construct a tree using a distance matrix $d = (d(i,j))_{1 \leq i,j \leq N}$ as input. We note that in general d does not provide a distance metric, and $d(i,j) \neq d(j,i)$. A schematic is depicted in Fig. 2.2**a**. The choice of d is crucial, and we briefly illustrate how d is estimated (see Section 2.3 for details).

In the absence of recombination, a single rooted binary tree describes the genealogy of the whole genome. We define the number of *derived* mutations $d(i,j)$ as the number of mutations carried by haplotype i and not by haplotype j. Notice that $d(i,j) \neq d(j,i)$. We use d as the distance matrix for the hierarchical clustering algorithm described in Section 2.2. This choice of distance matrix is motivated by the fact that if we assume that every mutation happened exactly once, we



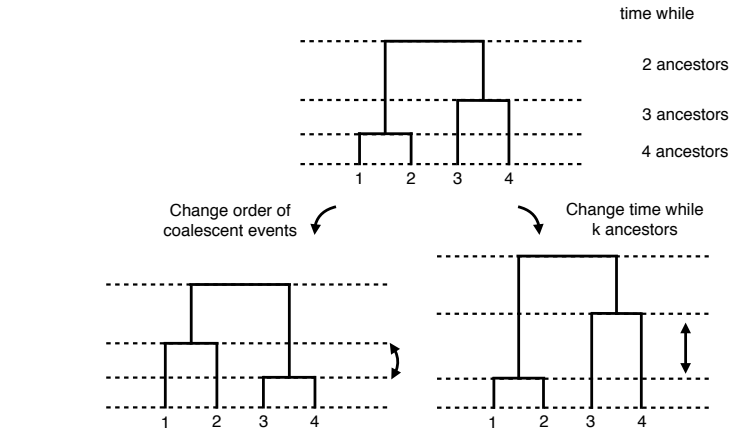


Figure 2.2: Schematics of the tree building and branch length estimation algorithms.
a, Schematic of the hierarchical clustering algorithm for estimating tree topology. In the case of no recombination, the algorithm obtains matrix d containing the number of derived mutations as input. Row i of this matrix determines the order in which haplotype i coalesced with other haplotypes. Using Eq. (2.1), the algorithm finds the pair that coalesces with each other before coalescing with any other sequence. In the example shown here, we can coalesce haplotypes 1 and 2 or haplotypes 3 and 4. We choose to coalesce haplotypes 1 and 2 first because the symmetrised distance is smaller for this pair. The resulting tree topology is consistent with the gene tree describing the data (We note that coalescing 3 and 4 first, then coalescing the resulting lineage with 2 and finally with 1 is also consistent with the gene tree and may be constructed using this algorithm if one chooses to first coalesce 3 and 4.). In contrast, when the hierarchical clustering algorithm is applied to the symmetrised matrix $(d(i,j)+d(j,i))_{i,j=1,\ldots,N}$, haplotypes 2 and 3 are coalesced first and the constructed tree topology is wrong. This is equivalent to applying the UPGMA algorithm to the symmetrised matrix of derived mutations [143].
b, Schematic of possible proposal moves in the MCMC algorithm for estimating branch lengths. We propose either a change in the order of coalescence events or a change in the time while k ancestors remain.

can determine the order in which haplotype i coalesced with other haplotypes by ordering them in ascending order of derived mutations $d(i,j)$ ($j = 1,\ldots,i-1,i+1,\ldots,N$). The resulting tree topology is guaranteed to be consistent with the truth in the sense that the constructed tree is a subtree of the gene tree describing the data (see Section 2.2.3).

In the presence of recombination, the relative order of coalescences changes along the genome. We apply a modified version of a Li and Stephens type HMM to calculate a local number of derived mutations $d(i,j;\ell)$ at every SNP ℓ [90] (see Section 2.3 for details of the HMM). We can show that $d(i,j;\ell)$ converges to $d(i,j)$ in the limit of no recombination (see Section 2.3.2). We use $d(i,j;\ell)$ as the distance matrix for the hierarchical clustering algorithm. Our method builds trees that are consistent with the truth if $d(i,j;\ell)$ orders haplotypes correctly.

2.2.1 Hierarchical clustering

The tree builder is initialised by placing each haplotype in a separate cluster. The algorithm proceeds by finding pairs of clusters that coalesce with each other before coalescing with any other haplotype. Such a pair satisfies

$$\begin{aligned}\mathcal{A} &\in \arg\min_{\mathcal{A}'} d(\mathcal{B},\mathcal{A}')\\ \mathcal{B} &\in \arg\min_{\mathcal{B}'} d(\mathcal{A},\mathcal{B}'),\end{aligned} \tag{2.1}$$

where the distance $d(\mathcal{A}',\mathcal{B}')$ between two clusters $\mathcal{A}'$ and $\mathcal{B}'$ of cardinality $|\mathcal{A}'|$ and $|\mathcal{B}'|$, respectively, is given by

$$d(\mathcal{A}',\mathcal{B}') = \frac{1}{|\mathcal{A}'||\mathcal{B}'|}\sum_{x\in\mathcal{A}'}\sum_{y\in\mathcal{B}'} d(x,y). \tag{2.2}$$

There might be more than one pair satisfying Eq. (2.1), in which case we choose the pair with the smallest symmetrised distance $d(\mathcal{A},\mathcal{B}) + d(\mathcal{B},\mathcal{A})$. The chosen pair is then combined to a new cluster comprising all haplotypes of both clusters. The algorithm is terminated when all haplotypes are in one cluster.

A pair satisfying Eq. (2.1) is guaranteed to exist as long as there exists a tree consistent with the order of coalescence events implied by matrix d. Sometimes

such a tree cannot be constructed. To make our algorithm robust to such situations, we replace Eq. (2.1) by

$$\begin{aligned} \mathcal{A} &\in \{\mathcal{A}' : |d(\mathcal{B}, \mathcal{A}') - \min_{\mathcal{C}} d(\mathcal{B}, \mathcal{C})| < \varepsilon\} \\ \mathcal{B} &\in \{\mathcal{B}' : |d(\mathcal{B}', \mathcal{A}) - \min_{\mathcal{C}} d(\mathcal{C}, \mathcal{A})| < \varepsilon\}, \end{aligned} \tag{2.3}$$

which allows for a tolerance $\varepsilon > 0$ in finding feasible pairs. In our implementation, we set $\varepsilon = 0.2$. In addition, in case pairs satisfying Eq. (2.3) cannot be found, we choose the pair with the smallest symmetrised distance.

2.2.2 Deciding when to build a new tree

Prior knowledge of recombination points would allow us to only reestimate tree topology at SNPs following a recombination event because neighbouring trees are identical otherwise. Unfortunately, this knowledge is not available to us; nevertheless we would like to avoid reestimating tree topologies at every SNP as this is computationally inefficient. We therefore employ the following strategy. We initially estimate the tree topology at the first SNP of the 5' end of a chromosome. We then reestimate the tree topology whenever we think a recombination has occurred, which we ascertain by mapping SNPs onto the tree describing the previous SNP. If the SNP maps onto the tree without having to *flip* it (i.e., reassign ancestral/derived alleles), we infer that the branch onto which the SNP maps has not recombined and we accept the tree topology of the previous SNP. Otherwise, if the SNP does not map onto the tree of the previous SNP or if it needs to be flipped, we conclude that a recombination is likely to have occurred and we therefore reestimate tree topology. This strategy will shift the placement of tree changes to the right of true recombination points.

We use present-day genome data of outgroups, such as chimpanzees and other primates for humans, to determine the ancestral and derived alleles. Occasionally, the ancestral allele can be confused with the alternative allele due to repeat mutations between species or sequencing errors. We can infer such cases if a SNP maps onto the tree only after non-carriers of the mutation are reinterpreted as carriers and

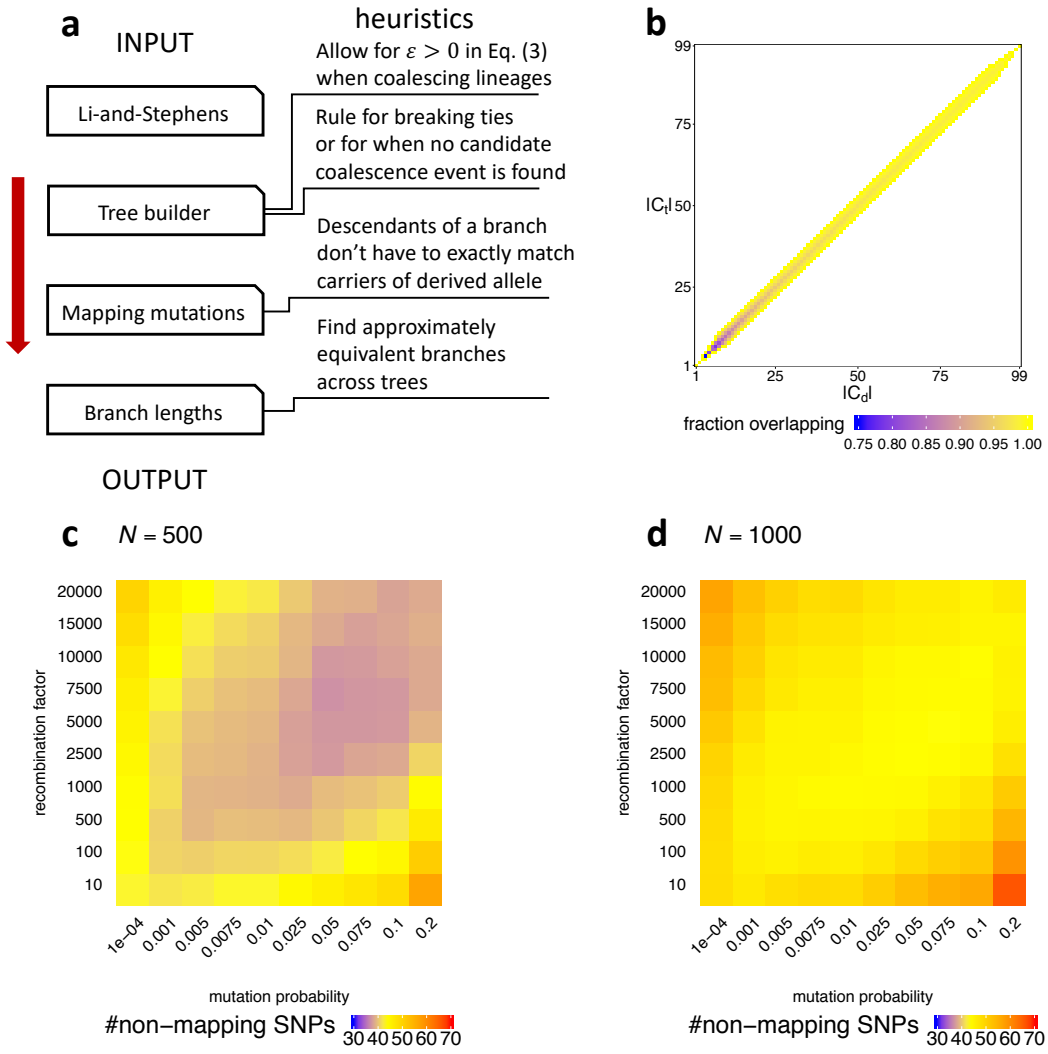


Figure 2.3: Mapping rule for mutations and sensitivity of the modified Li-and-Stephens HMM to parameter choice.
a, Schematic illustrating which parts of Relate use heuristic approaches. **b**, Heatmap showing the necessary and sufficient overlap between the set of descendants of a branch (C_t) and the set of carriers of the derived allele (C_d), given $|C_\mathrm{t}|$ and $|C_\mathrm{d}|$, with $N = 100$, as determined by Eqs. (2.4) and (2.5). Colours show $|C_\mathrm{t} \cap C_\mathrm{d}| / \min\{|C_\mathrm{t}|, |C_\mathrm{d}|\}$, where white indicates that a mutation can never be mapped for the corresponding combination of $|C_\mathrm{t}|$ and $|C_\mathrm{d}|$. **c**, **d**, Number of non-mapping SNPs for different values of p (horizontal axis) and R (vertical axis) for $N = 500$ (**c**) and $N = 1000$ (**d**). The subsets of haplotypes are chosen uniformly at random from all haplotypes. We calculated the mean over 50 randomly chosen subregions of length 1200 SNPs on chromosome 20. In our implementation, we fixed $p = 0.025$ and $R = 2500$.

vice versa. We refer to these SNPs as *flipped* SNPs and reinfer tree topology whenever we detect a potentially flipped SNP.

A mutation is mapped to the branch for which the descendants coincide with the carriers of the derived allele. Such a branch exists as long as the mutation occurred exactly once in human history and the estimated tree topology is correct. To be robust to errors in the data and the inferred tree, we relax this requirement as follows.

By placing a mutation on a branch, we indicate that all descendants below that branch carry the mutation. Let us denote the set of samples that carry the mutation by C_t and the set of samples that do not carry the mutation by N_t. Similarly, let us denote the set of samples that (do not) carry the mutation in the data set by C_d and N_d. For a mutation to map to a branch, it needs to satisfy

$$\begin{aligned} \frac{|C_\mathrm{t} \cap C_\mathrm{d}|}{\max\{|C_\mathrm{t}|, |C_\mathrm{d}|\}} &> 0.7 \\ \frac{|N_\mathrm{t} \cap N_\mathrm{d}|}{\max\{|N_\mathrm{t}|, |N_\mathrm{d}|\}} &> 0.7. \end{aligned} \tag{2.4}$$

These conditions should identify suitable candidate branches and should prevent mapping infrequent mutations, such as doubletons, to a unique branch, when in fact they cannot have arisen by a single mutation. Out of all remaining candidate branches, we calculate the fraction of missclassified haplotypes given by

$$\frac{|N_\mathrm{t} \cap C_\mathrm{d}| + |C_\mathrm{t} \cap N_\mathrm{d}|}{N}. \tag{2.5}$$

We also calculate the same quantity for branches that satisfy Eq. (2.4) after reinterpreting carriers as non-carriers and vice-versa. We then accept the branch with the minimum score given by Eq. (2.5) if it is also less than 0.03. We only flip a SNP if this leads to a smaller score (i.e., in case of a tie, we do not flip the SNP). These rules, though heuristic, allow approximate mapping for mutations in 4 or more copies in the data set (Fig. 2.3b).

If such unique branch cannot be found, we map the mutation to more than one branch. In this case, we find the smallest set of branches, such that all carriers of the mutation C_d are below one chosen branch and such that the summed score given by Eq. (2.5) equals zero. We do the same after flipping the SNP. We choose to flip the

SNP only if this leads to a smaller set of branches. For all chosen branches, we then add one over the number of branches chosen to the number of mutations on these branches. For many analyses, we only consider mutations mapping to a unique branch, however we note that e.g., CpG mutations often occur on multiple branches.

2.2.3 Consistency of estimated tree topology in the absence of recombination

We prove that the use of the number of derived mutations $d(i,j)$ as a distance matrix always guarantees estimation of a tree topology consistent with the truth, assuming that every mutation is unique in history, and no recombination. We recall that we defined the number of derived mutations $d(i,j)$ as the number of mutations carried by haplotype i and not by haplotype j. We assume that there is a gene tree that is consistent with SNPs in the data in the sense that every SNP can be mapped to a unique branch of the gene tree [7]. Such a gene tree contains polytomies reflecting branches unresolved by observed mutations, and always exists and is unique assuming the infinite-sites model. We say, that a binary coalescent (sub-)tree, is consistent with a gene tree if for all mutations, all carriers coalesce before any of them coalesces with a non-carrier of the mutation.

We prove that our tree builder described in Section 2.2 constructs a tree that is consistent with the gene tree. The input matrix is the matrix of derived mutations $(d(i,j))_{i,j=1,\ldots,N}$ and we set $\varepsilon = 0$ in the tree builder. The tree builder therefore coalesces haplotypes according to Eq. (2.1).

Proposition 2.2.1. *The tree builder cannot coalesce carriers and non-carriers before it has coalesced all carriers of any SNP.*

Proof. Assume without loss of generality (w.l.o.g.) that $1,\ldots,k$ are carriers of a SNP and $k+1,\ldots,N$ are non-carriers of the same SNP. We first observe that if a branch in the gene tree has a mutation, all descendants of that branch carry at least one more derived mutation to any non-carrier of the mutation than to a carrier of the mutation. Using the this observation, we obtain

$$d(i,j) < d(i,h) \text{ for } i,j \in \{1,\ldots,k\} \text{ and } h \in \{k+1,\ldots,N\}. \tag{2.6}$$

This property is why it is important to distinguish derived mutations in determining distances; the equivalent to Eq. (2.6) does not hold if $d(i, j)$ is determined as the number of differences between sequences i and j.

It follows that coalescing a carrier and a non-carrier in the first step of the algorithm is not feasible in Eq. (2.1). Let us assume that we have not coalesced carriers and non-carriers until the x'th step of the algorithm and that there still exist more than one cluster of carriers. We prove that it is still not feasible to coalesce a cluster of carriers and non-carriers in the $x + 1$st step. Let us denote by $\mathcal{A}$ and $\mathcal{B}$ clusters containing only carriers and by $\mathcal{C}$ a cluster containing only non-carriers. We obtain from Eq. (2.6),

$$d(\mathcal{A}, \mathcal{B}) < d(\mathcal{A}, \mathcal{C}), \tag{2.7}$$

because an average of $d(i, j)$ with $i, j \in \{1, \ldots, k\}$ is always smaller than an average of $d(i, h)$ with $i \in \{1, \ldots, k\}$ and $h \in \{k + 1, \ldots, N\}$. Therefore coalescing $\mathcal{A}$ and $\mathcal{C}$ does not satisfy Eq. (2.1). □

With Proposition 2.2.1, we know that the tree builder never violates the gene tree and therefore, if the algorithm successfully coalesces all haplotypes, we obtain a tree that is consistent with the gene tree. It remains to prove that the tree builder can always find a next coalescence event and therefore, by induction, terminates with all haplotypes in one cluster.

Proposition 2.2.2. *Assuming there is a gene tree consistent with the data, the tree builder terminates with all haplotypes in one cluster.*

Proof. Assume that we are in the xth step of the tree builder. We prove that we can coalesce a pair of clusters to proceed to the $x + 1$st step of the tree builder. We have already proved that the tree constructed until the xth step is consistent with the gene tree. We can therefore find two clusters $\mathcal{A}$ and $\mathcal{B}$, such that coalescing these clusters is consistent with the gene tree.

Let $\mathcal{C}$ be a third cluster distinct from $\mathcal{A}$ and $\mathcal{B}$. For $\mathcal{A}$ and $\mathcal{B}$ to satisfy Eq. (2.1), we require

$$\begin{aligned} d(\mathcal{A},\mathcal{B}) &\leq d(\mathcal{A},\mathcal{C}) \text{ and} \\ d(\mathcal{B},\mathcal{A}) &\leq d(\mathcal{B},\mathcal{C}). \end{aligned} \tag{2.8}$$

We show $d(\mathcal{A},\mathcal{B}) \leq d(\mathcal{A},\mathcal{C})$ and note that $d(\mathcal{B},\mathcal{A}) \leq d(\mathcal{B},\mathcal{C})$ can be shown analogously.

Let us define by $d_\ell(\mathcal{A},\mathcal{B})$ the distance considering only SNP ℓ, such that $d(\mathcal{A},\mathcal{B}) = \sum_{\ell=1}^{L} d_\ell(\mathcal{A},\mathcal{B})$ with L denoting the number of SNPs. We note that any SNP with no derived sequences in $\mathcal{A}$ yields $d_\ell(\mathcal{A},\mathcal{B}) = d_\ell(\mathcal{A},\mathcal{C}) = 0$. Let us therefore consider the possible types of SNPs with at least one derived mutation in $\mathcal{A}$. We note that we cannot have SNPs with both carriers and non-carriers in more than one cluster, because we would have violated Proposition 2.2.1. Therefore, the possible SNPs are as follows.

- If a SNP is only derived in sequences in $\mathcal{A}$, then $d_\ell(\mathcal{A},\mathcal{B}) = d_\ell(\mathcal{A},\mathcal{C})$.
- If a SNP is derived in all sequences of $\mathcal{A}$ and $\mathcal{B}$, but no sequences in $\mathcal{C}$, then

$$d_\ell(\mathcal{A},\mathcal{B}) = 0 < 1 = d_\ell(\mathcal{A},\mathcal{C}). \tag{2.9}$$

- If a SNP is derived in all sequences of $\mathcal{A}$ and $\mathcal{C}$, but no sequences of $\mathcal{B}$, then this SNP violates the assumption that coalescing $\mathcal{A}$ and $\mathcal{B}$ is consistent with the gene tree.
- If a SNP is derived in all sequences of $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, then $d_\ell(\mathcal{A},\mathcal{B}) = d_\ell(\mathcal{A},\mathcal{C}) = 0$.

By summing over all SNPs, we obtain

$$d(\mathcal{A},\mathcal{B}) = \sum_{\ell=1}^{L} d_\ell(\mathcal{A},\mathcal{B}) \leq \sum_{\ell=1}^{L} d_\ell(\mathcal{A},\mathcal{C}) = \sum_{\ell=1}^{L} d(\mathcal{A},\mathcal{C}), \tag{2.10}$$

as required. □

We conclude with a few observations. First, we notice that all branches that have a mutation in the gene tree are guaranteed to exist in the tree built by our tree builder. In particular, if all branches in the gene tree have a mutation, the tree builder is guaranteed to build the correct tree. The tree builder therefore constructs the correct tree in the limit of an infinite mutation rate (and under the infinite-sites assumption). Second, we notice that if recombination rates are only either infinite or zero, we can divide the genome into regions of zero recombination rates. In this case, the tree builder constructs trees consistent with the gene trees of these regions. Finally, we notice that our proof did not depend on the absolute values of the input distance matrix, and that the only requirement was that rows of the distance matrix are perfectly correlated to the order of coalescences of a haplotype with other haplotypes. For example, any linear transformation of each row does not impact the algorithm or the proof. Therefore, we expect our tree builder to construct an accurate tree as long as rows of the distance matrix obtained from the modified Li-and-Stephens HMM (Section 2.3) is well correlated with the order in which a haplotype coalescences with other haplotypes.

2.3 Calculating distance matrices

2.3.1 Assumptions about the input data

We apply a version of the Li-and-Stephens HMM to calculate distance matrices. To apply this algorithm, we assume haplotype SNP data as input, which can be inferred by phasing genotype data [16, 148]. We assume a high coverage of SNPs along the genome and no bias with respect to the frequency of a mutation in the population. We also assume that at most one mutation has occurred at any given position along the genome since the divergence of humans from chimpanzees and other primates. This assumption, known as the infinite sites model, is justified by a small average mutation rate in humans [84].

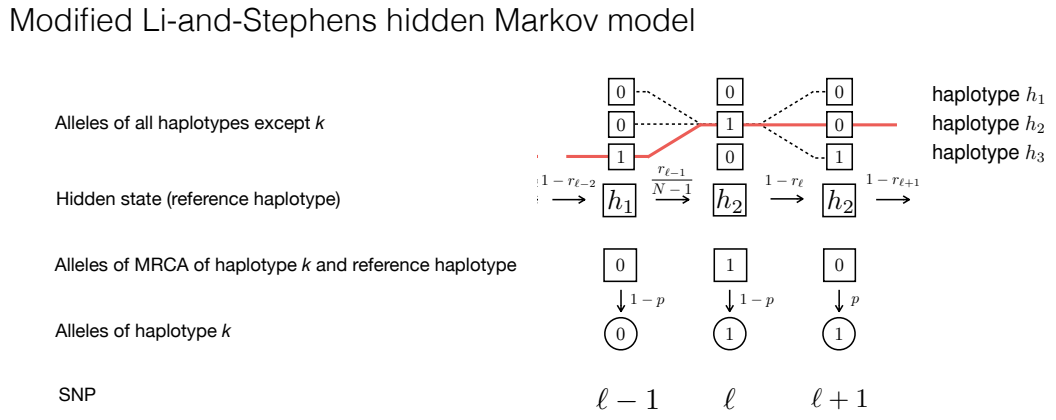


Figure 2.4: Schematic of the modified Li and Stephens HMM.
Schematic of the modified Li-and-Stephens hidden Markov model (HMM) applied to haplotype k, which has alleles 0, 1, 1 at loci $\ell-1$, ℓ, $\ell+1$. The emission and transition probabilities shown correspond to the path indicated by the red solid line. At SNP $\ell-1$, the reference haplotype is h_1 which has allele 1. Because the allele of haplotype k is 0, the allele of the MRCA with h_1 is also 0 assuming that every mutation is unique in history. Therefore, the emission probability equals $1-p$, where p is the probability of a mutation. At SNP ℓ, the reference haplotype has changed to h_2. The alleles of haplotype k and h_2 are 1. Therefore, the MRCA has allele 1 and the emission probability is given by $1-p$. At SNP $\ell+1$, haplotype k has allele 1. The allele of the reference haplotype h_2 is 0 and so is that of the MRCA, such that the emission probability equals p. Using this HMM, we calculate the probability $P_{\mathrm{m}}(H_\ell = j|D^{(k)})$. This is the probability of copying from reference haplotype j at SNP ℓ, conditional on observing $D^{(k)}$. We notice that $P_{\mathrm{m}}(H_\ell = j|D_\ell^{(k)})$ is obtained as the sum of all possible paths when $H_\ell = j$ is fixed (indicated by the dashed lines).

Additionally, we require knowledge of the ancestral allele for every recorded mutation. The ancestral allele can be determined by aligning the human genome to present day genomes of other primates [60]. Assuming that it is unlikely to observe a mutation at the same genomic position in humans and other primates, the allele carried by other primates is declared to be the ancestral allele. For humans, we use the ancestral genome that was inferred as part of the 1000 Genomes Project (see URLs in main text).

For the robustness of Relate to genotype errors and occasional confusion of ancestral and alternative alleles, please refer to Chapter 4.

Under these assumptions, we can represent SNP data as a binary matrix D with each row corresponding to one haplotype. The matrix D therefore has dimensions

$N \times L$, where N is the number of haplotypes and L is the number of SNPs. In this representation, a haplotype is a binary vector, where the ancestral allele is denoted by 0 and the alternative allele is denoted by 1. We denote row i of D by $D^{(i)}$.

2.3.2 Modified Li-and-Stephens HMM

A recombination event may change the order of coalescence events between haplotypes. Therefore, mutations close to the SNP of consideration are more informative than SNPs further away. To capture this, we apply a Li-and-Stephens type HMM [90]. We modified the original HMM by changing emission probabilities such that they take ancestral and derived states into account. A schematic of the HMM is depicted in Fig. 2.2**c**.

The original Li-and-Stephens HMM can be interpreted as a generative model for haplotype $D^{(i)}$ using all other haplotypes as inputs. To generate the allele $D_\ell^{(i)}$ at SNP ℓ, we first choose one reference haplotype H_ℓ from all other haplotypes. This reference haplotype is the hidden state of the HMM. In Ref. [90], the emission probabilities are defined by

$$P_{\mathrm{c}}(D_\ell^{(i)}|H_\ell = j) = \begin{cases} p & \text{if } D_\ell^{(i)} \neq D_\ell^{(j)}, \\ 1-p & \text{if } D_\ell^{(i)} = D_\ell^{(j)}. \end{cases} \tag{2.11}$$

Here, p is the mismatch probability. In this conventional definition of the emission probabilities, a mutation may have occurred on the branch from i to the MRCA with j, or on the branch from j to the MRCA with i. Therefore, information about on which branch the mutation occurred is lost. To preserve this information, we change the emission probabilities to

$$P_{\mathrm{m}}(D_\ell^{(i)}|H_\ell = j) = \begin{cases} p & \text{if } D_\ell^{(i)} = 1, D_\ell^{(j)} = 0, \\ 1-p & \text{if } D_\ell^{(i)} = 0, D_\ell^{(j)} = 0, \\ 1-p & \text{if } D_\ell^{(i)} = 1, D_\ell^{(j)} = 1, \\ 1-p & \text{if } D_\ell^{(i)} = 0, D_\ell^{(j)} = 1. \end{cases} \tag{2.12}$$

We can interpret p as the probability of a mutation since the MRCA. The emission probability equals p, if haplotype i carries a mutation at site ℓ which is not carried by the reference haplotype j, such that the mutation must have occurred on the

branch from i to the MRCA with j. Otherwise, no mutation occurred on this branch and the emission probability equals $1-p$ (see Fig. 2.2 **c**). We note that this modified HMM is not anymore a generative model as it requires knowledge of $D_\ell^{(j)}$ ($j \neq i$), instead it is a model of changes since the MRCA of i and j.

The hidden state may change between neighbouring SNPs according to transition probabilities r_ℓ. The transition probabilities are proportional to the recombination probabilities obtained from a recombination map. We describe how to choose p and r_ℓ in Section 2.3.3.

Using the modified Li-and-Stephens HMM, we can calculate a distance matrix at every SNP which we will use as input for the tree builder described in Section 2.2. We first derive a distance matrix for the case when transition probabilities are set to zero, which corresponds to no recombination. In this case, the reference haplotype remains the same along the genome. The likelihood of observing $D^{(i)}$ given reference haplotype $H_\ell = j$ is then given by

$$P_{\mathrm{m}}(D^{(i)}|H_\ell = j) = p^{d(i,j)}(1-p)^{L-d(i,j)}, \tag{2.13}$$

where $d(i,j)$ is the number of derived mutations defined in Section 2.1. By taking logarithms on both sides of Eq. (2.13), we obtain

$$\log P_{\mathrm{m}}(D^{(i)}|H_\ell = j) = d(i,j)\log\left(\frac{p}{1-p}\right) + L\log(1-p). \tag{2.14}$$

By rearranging Eq. (2.14), we obtain

$$d(i,j) = \frac{\log P_{\mathrm{m}}(D^{(i)}|H_\ell = j) - L\log(1-p)}{\log\left(\frac{p}{1-p}\right)}. \tag{2.15}$$

We can generalise Eq. (2.15) to the case of non-zero recombination rates to define, for each SNP ℓ, the *local* number of derived mutations

$$d(i,j;\ell) = \frac{\log P_{\mathrm{m}}(D^{(i)}|H_\ell = j) - L\log(1-p)}{\log\left(\frac{p}{1-p}\right)}. \tag{2.16}$$

Equation (2.16), corresponding to a local number of derived mutations, orders coalescence events locally at every SNP. In practice, we use $d(i,j;\ell) - \min_{j\neq i} d(i,j;\ell)$ as our distance matrix, which only affects our tie-breaking heuristic in case no pair

of haplotypes (or clusters) satisfy Eq. (2.3) and we choose a pair with minimum entry in the symmetrised matrix. By subtracting the minimum entry from each row, we remove residuals interpretable as the number of derived mutations on the tip branches, which could confound the symmetrised distance if some haplotypes have many more mutations at their tips than other haplotypes. The quantity $\log P_{\mathrm{m}}(D^{(i)}|H_\ell = j)$ in Eq. (2.16) can be calculated using the forward-backward algorithm. We describe how to efficiently implement the Li-and-Stephens HMM in Section 2.3.4.

We notice that with the conventional definition of the emission probabilities (Eq. (2.11)) and no recombination, we obtain

$$d(i,j) + d(j,i) = \frac{\log P_{\mathrm{c}}(D^{(i)}|H_\ell = j) - L\log(1-p)}{\log\left(\frac{p}{1-p}\right)}. \tag{2.17}$$

In this case, we obtain a symmetric matrix that stores the number of mutations that differ between two haplotypes. Using the matrix $d(i,j) + d(j,i)$ as input to our tree builder is equivalent to applying the UPGMA algorithm [143], which is an alternative hierarchical clustering algorithm, to the matrix $d(i,j) + d(j,i)$. We show in Fig. 2.2**a**, how this can lead to estimation of a tree topology that is inconsistent with the data. Intuitively, this is because $d(i,j) + d(j,i)$ combines the number of mutations of two branches, such that information about the number of mutations on each branch is lost. Therefore, $d(i,j)$ preserves more information about the tree topology than $d(i,j) + d(j,i)$.

2.3.3 Choosing parameters for the modified Li-and-Stephens HMM

The transition probabilities in the HMM are determined by the recombination map, multiplied by a constant factor R. The mutation probability p should reflect the true biological mutation rate and errors in the data set. In practice, we find that the tree topology constructed using distance matrix $d(i,j;\ell)$ (Eq. (2.16)) is very robust with respect to choice of p and R. To illustrate this, we evaluate the performance of our algorithm for different choices of p and R. In our implementation, we have fixed $p = 0.025$ and $R = 2500$.

To evaluate how well a pair (p, R) captures ancestry information in a subregion of the genome, we sample subregions of lengths 1200 SNPs at random. We first apply the modified Li-and-Stephens HMM on the chosen subregion. For every SNP $400 < \ell < 800$, we then do the following. We first calculate the distance matrix as described in Section 2.3. We modify the distance matrix by hiding the focal SNP ℓ which can be done by subtracting 1 from any entry (i, j) where i is a carrier and j is not a carrier of the mutation. We then build the tree with this modified distance matrix using the hierarchical clustering algorithm described in Section 2.1. Finally, we attempt to place the focal SNP on branches of the tree and record whether a branch exists such that the descendants of that branch coincide with carriers of the SNP.

We therefore evaluate a pair (p, R) by how well we can build trees at focal SNPs using information from surrounding SNPs only. As R becomes larger and p becomes smaller, only SNPs close to the focal SNP will influence the distance matrix. As R becomes smaller and p becomes larger, SNPs further away will influence the distance matrix. We should find an optimal pair (p, R) such that we include enough SNPs to be able to build a tree onto which the focal SNP can be mapped and such that we prevent SNPs that map onto different trees from distorting the signal.

We applied this method to 50 subregions of chromosome 20 of the 1000 Genomes Project data set. In Fig. 2.3c and d, we show that the number of non-mapping SNPs is relatively robust to the choice of (p, R), particularly along the axis $p/R = 10^{-5}$. We therefore fix $p = 0.025$ and $R = 2500$ in our implementation.

2.3.4 Speed-up and approximation of the modified Li-and-Stephens HMM

To calculate $d(i, j; \ell)$ defined in Eq. (2.16), we need to calculate $P_{\mathrm{m}}(D^{(i)}|H_\ell = j)$. We apply Bayes' theorem and obtain

$$P_{\mathrm{m}}(D^{(i)}|H_\ell = j) = \frac{P_{\mathrm{m}}(H_\ell = j|D^{(i)})P_{\mathrm{m}}(D^{(i)})}{P_{\mathrm{m}}(H_\ell = j)}. \tag{2.18}$$

We assume the prior probability of copying from any haplotype to be identical, such that $P_{\mathrm{m}}(H_\ell = j) = 1/(N-1)$, and calculate $P_{\mathrm{m}}(H_\ell = j|D^{(i)})$ using a forward-backward algorithm. The forward algorithm calculates $\alpha_j(\ell) = P_{\mathrm{m}}(H_\ell = j, D^{(i)}_{1:\ell})$ and the backward algorithm calculates $\beta_j(\ell) = P(D^{(i)}_{(\ell+1):L}|H_\ell = j)$ such that

$$P_{\mathrm{m}}(H_\ell = j|D^{(i)}) = \frac{\alpha_j(\ell)\beta_j(\ell)}{P_{\mathrm{m}}(D^{(i)})}. \tag{2.19}$$

By substituting Eq. (2.19) in Eq. (2.18) and taking logarithms, we obtain

$$\log P_{\mathrm{m}}(D^{(i)}|H_\ell = j) = \log\left(\alpha_j(\ell)\beta_j(\ell)\right) + \log(N-1). \tag{2.20}$$

By substituting Eq. (2.20) in Eq. (2.16), we obtain

$$d(i,j;\ell) = \frac{\log\left(\alpha_j(\ell)\beta_j(\ell)\right) + \log(N-1) - L\log(1-p)}{\log\left(\frac{p}{1-p}\right)}. \tag{2.21}$$

To speed-up the calculation of $\alpha_j(\ell)$ and $\beta_j(\ell)$ in Eq. (2.21), we calculate $P_{\mathrm{m}}(H_\ell = j|D^{(i)}) \propto \alpha_j(\ell)\beta_j(\ell)$ only for SNPs ℓ at which haplotype i is a carrier of the derived allele, i.e. $D^{(i)}_\ell = 1$. This implies that the forward-backward algorithm is applied to a different set of SNPs depending on the haplotype i.

The calculated $P_{\mathrm{m}}(H_\ell = j|D^{(i)})$ at SNPs ℓ with $D^{(i)}_\ell = 1$ are exact up to a multiplicative constant $(1-p)^{L-L^{(i)}}$, where $L^{(i)} = \sum_{\ell=1}^{L} D^{(i)}_\ell$ is the number of derived mutations carried by i. We therefore calculate Eq. (2.21) as

$$d(i,j;\ell) = \frac{\log\left(\alpha^{\mathrm{d}}_j(\ell)\beta^{\mathrm{d}}_j(\ell)\right) + \log(N-1) - L^{(i)}\log(1-p)}{\log\left(\frac{p}{1-p}\right)}, \tag{2.22}$$

where $\alpha^{\mathrm{d}}_j(\ell)$ and $\beta^{\mathrm{d}}_j(\ell)$ are the forward and backward probabilities calculated for SNPs ℓ at which $D^{(i)}_\ell = 1$.

For a site ℓ at which haplotype i is not a carrier of the derived allele, we approximate $P_{\mathrm{m}}(H_\ell = j|D^{(i)}) \propto \alpha_j(\ell)\beta_j(\ell)$ using $P_{\mathrm{m}}(H_{\ell_{\mathrm{left}}} = j|D^{(i)})$ and $P_{\mathrm{m}}(H_{\ell_{\mathrm{right}}} = j|D^{(i)})$, where ℓ_{left}, ℓ_{right} are the nearest derived sites to either side of ℓ, i.e., $D^{(i)}_{\ell_{\mathrm{left}}} = D^{(i)}_{\ell_{\mathrm{right}}} = 1$ and $D^{(i)}_\ell = 0$ for $\ell_{\mathrm{left}} < \ell < \ell_{\mathrm{right}}$. For such ℓ, we approximate $P_{\mathrm{m}}(H_\ell = j|D^{(i)})$ as a weighted average of $P_{\mathrm{m}}(H_{\ell_{\mathrm{left}}} = j|D^{(i)})$ and $P_{\mathrm{m}}(H_{\ell_{\mathrm{right}}} =$

$j|D^{(i)})$. This approximation is valid if the probability for two or more recombinations between ℓ_{left} and ℓ_{right} is sufficiently small. We expand

$$\begin{aligned} P_{\text{m}}(H_\ell = j|D^{(i)}) = P_{\text{m}}(H_\ell = j,\ &0 \text{ recombinations in } [\ell_{\text{left}}, \ell_{\text{right}})|D^{(i)}) \\ &+ P_{\text{m}}(H_\ell = j,\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}})|D^{(i)}) \\ &+ P_{\text{m}}(H_\ell = j,\ \text{more than 1 rec. in } [\ell_{\text{left}}, \ell_{\text{right}})|D^{(i)}). \end{aligned} \tag{2.23}$$

For the third term on the right hand side of Eq. (2.23), we obtain

$$P_{\text{m}}(H_\ell = j,\ \text{more than 1 rec. in } [\ell_{\text{left}}, \ell_{\text{right}})|D^{(i)}) = O((r_{\text{left}} + r_{\text{right}})^2). \tag{2.24}$$

For the second term on the right hand side of Eq. (2.23), we obtain

$$\begin{aligned} &P_{\text{m}}(H_\ell = j,\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}})|D^{(i)}) \\ &= P_{\text{m}}(H_{\ell_{\text{right}}} = j,\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell),\ 0 \text{ rec. in } [\ell, \ell_{\text{right}})|D^{(i)}) + \\ &\quad P_{\text{m}}(H_{\ell_{\text{left}}} = j,\ 0 \text{ rec. in } [\ell_{\text{left}}, \ell),\ 1 \text{ rec. in } [\ell, \ell_{\text{right}})|D^{(i)}). \end{aligned} \tag{2.25}$$

We notice that the emission probability at any site $\ell_{\text{left}} < \ell < \ell_{\text{right}}$ equals $1 - p$ regardless of the hidden state H_ℓ, because $D_\ell^{(i)} = 0$. Therefore, we find that the probability of a recombination (i.e., switch of hidden state), given the data $D^{(i)}$, is the same at any position between ℓ_{left} and ℓ_{right}. We denote the recombination distance from ℓ_{left} to ℓ by r_{left} and the recombination distance from ℓ to ℓ_{right} by r_{right}. These recombination distances correspond to the probability of a switch of hidden states between the two SNPs. We therefore obtain

$$\begin{aligned} &P_{\text{m}}(H_{\ell_{\text{right}}} = j,\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell),\ 0 \text{ rec. in } [\ell, \ell_{\text{right}})|D^{(i)}) \\ &\quad = P_{\text{m}}(\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell)|\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}))P_{\text{m}}(H_{\ell_{\text{right}}} = j,\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}})|D^{(i)}) \\ &\quad = \frac{r_{\text{left}}}{r_{\text{left}} + r_{\text{right}}} P_{\text{m}}(H_{\ell_{\text{right}}} = j,\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}})|D^{(i)}). \end{aligned} \tag{2.26}$$

Analogously, we obtain

$$\begin{aligned} &P_{\text{m}}(H_{\ell_{\text{left}}} = j,\ 0 \text{ rec. in } [\ell_{\text{left}}, \ell),\ 1 \text{ rec. in } [\ell, \ell_{\text{right}})|D^{(i)}) \\ &\quad = \frac{r_{\text{right}}}{r_{\text{left}} + r_{\text{right}}} P_{\text{m}}(H_{\ell_{\text{left}}} = j,\ 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}})|D^{(i)}). \end{aligned} \tag{2.27}$$

We substitute Eqs. (2.26) and (2.27) in Eq. (2.25) and obtain

$$\begin{aligned}
&P_{\mathrm{m}}(H_\ell = j,\ 1 \text{ rec. in } [\ell_{\mathrm{left}}, \ell_{\mathrm{right}})|D^{(i)}) \\
&\quad = \frac{r_{\mathrm{left}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} P_{\mathrm{m}}(H_{\ell_{\mathrm{right}}} = j,\ 1 \text{ rec. in } [\ell_{\mathrm{left}}, \ell_{\mathrm{right}})|D^{(i)}) \\
&\qquad + \frac{r_{\mathrm{right}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} P_{\mathrm{m}}(H_{\ell_{\mathrm{left}}} = j,\ 1 \text{ rec. in } [\ell_{\mathrm{left}}, \ell_{\mathrm{right}})|D^{(i)}).
\end{aligned} \tag{2.28}$$

For the first term on the right hand side of Eq. (2.23), we use that

$$\frac{r_{\mathrm{left}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} + \frac{r_{\mathrm{right}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} = 1, \tag{2.29}$$

to obtain

$$\begin{aligned}
&P_{\mathrm{m}}(H_\ell = j,\ 0 \text{ rec. in } [\ell_{\mathrm{left}}, \ell_{\mathrm{right}})|D^{(i)}) \\
&\quad = \frac{r_{\mathrm{left}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} P_{\mathrm{m}}(H_{\ell_{\mathrm{right}}} = j,\ 0 \text{ rec. in } [\ell_{\mathrm{left}}, \ell_{\mathrm{right}})|D^{(i)}) \\
&\qquad + \frac{r_{\mathrm{right}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} P_{\mathrm{m}}(H_{\ell_{\mathrm{left}}} = j,\ 0 \text{ rec. in } [\ell_{\mathrm{left}}, \ell_{\mathrm{right}})|D^{(i)}).
\end{aligned} \tag{2.30}$$

We substitute Eqs. (2.24), (2.28), and (2.30) in Eq. (2.23) and obtain

$$\begin{aligned}
P_{\mathrm{m}}(H_\ell = j|D^{(i)}) &= \frac{r_{\mathrm{left}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} P_{\mathrm{m}}(H_{\ell_{\mathrm{right}}} = j|D^{(i)}) \\
&\quad + \frac{r_{\mathrm{right}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} P_{\mathrm{m}}(H_{\ell_{\mathrm{left}}} = j|D^{(i)}) + O((r_{\mathrm{left}} + r_{\mathrm{right}})^2).
\end{aligned} \tag{2.31}$$

By multiplying both sides of Eq. (2.31) by $P_{\mathrm{m}}(D^{(i)})$, we obtain

$$\begin{aligned}
\alpha_j^{\mathrm{d}}(\ell)\beta_j^{\mathrm{d}}(\ell) &= \frac{r_{\mathrm{left}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} \alpha_j^{\mathrm{d}}(\ell_{\mathrm{left}})\beta_j^{\mathrm{d}}(\ell_{\mathrm{left}}) \\
&\quad + \frac{r_{\mathrm{right}}}{r_{\mathrm{left}} + r_{\mathrm{right}}} \alpha_j^{\mathrm{d}}(\ell_{\mathrm{right}})\beta_j^{\mathrm{d}}(\ell_{\mathrm{right}}) + O((r_{\mathrm{left}} + r_{\mathrm{right}})^2).
\end{aligned} \tag{2.32}$$

By substituting Eq. (2.32) in Eq. (2.22) and dropping terms $O((r_{\mathrm{left}} + r_{\mathrm{right}})^2)$, we obtain an approximation for calculating $d(i, j; \ell)$.

The complexity of the modified Li-and-Stephens HMM is reduced from $N^2 L$ to $N^2 \sum_{i=1}^{N} L^{(i)}$. While L is expected to be of order $N \log N$, we expect $L^{(i)}$ to be independent of the sample size N for large N. This is because in the standard coalescent, the expected TMRCA is, in the limit $N \to \infty$, given by $4N_{\mathrm{e}}$, such that the expected number of derived mutations carried by haplotype i since the population's TMRCA is $E[L^{(i)}] = 4N_{\mathrm{e}}\mu$. In practice, this can make the application of the algorithm about 10 times faster even for modest samples of a few thousand individuals.

2.4 Estimating branch lengths

We have now estimated the genetic ancestry of each SNP in form of a rooted binary tree. To estimate the branch lengths t_b ($b = 0, \ldots, 2N-2$) of these trees, we use a Metropolis-Hastings type MCMC algorithm. We note that while MCMC sampling is a computationally expensive approach, it is a flexible approach allowing, for instance, the specification of demographic histories. By sampling posterior branch lengths, we estimate the mean age of a coalescent event and use these to estimate branch lengths. This approach will yield branch lengths that reflect the coalescent prior for trees (or subtrees) with little information about branch lengths, which is desirable for many applications.

Before we can apply the MCMC algorithm, we notice that a recombination event changes only a few branches in adjacent trees along the genome. Some branches can persist over multiple trees. We therefore identify equivalent branches in adjacent trees along the genome (see Section 2.4.1). We then count the number of mutations across equivalent branches and calculate a cumulative mutation rate for each branch. This information is fed into the MCMC algorithm (see Section 2.4.2). We assume a coalescent prior given by the standard coalescent [85]. The effective population size is predetermined and assumed to be constant. We initialise the order of coalescence events to a random order, but such that no topological constraints are violated (see Section 2.4.3). We initialise the times while k ancestors remain using an Expectation-maximization (EM) algorithm that calculates the maximum-likelihood estimates (MLEs) of the times while k ancestors remain, where the order of coalescence events is kept fixed (see Section 2.4.4). This EM algorithm assigns a branch length of zero to any branch with no mapping mutations. Such branches are inconvenient for the MCMC sampling procedure. We therefore postprocess the branch lengths output by the EM algorithm by assigning a branch length of $1/(2N_\mathrm{e})$ to any branch of length zero, pushing any older coalescence events up accordingly to accommodate this change.

Once the constant population size model has been fitted, we can jointly infer piecewise-constant historical population sizes and branch lengths under a coalescent prior with variable historical population sizes (see Chapter 3).

2.4.1 Identifying equivalent branches in neighbouring trees

Let us take a branch $b_1 = (c_1, d_1)$ from one one tree and a branch $b_2 = (c_2, d_2)$ from another tree, where c_i and d_i $(i = 1, 2)$ denote coalescence events. We then say that branches b_1 and b_2 are equivalent if the descendants of c_1 coincide with those of c_2 and the descendants of d_1 coincide with those of d_2. To be robust to errors, we slightly relax this requirement as follows.

For every coalescence event, we store a vector of length N containing its present-day descendants, where the m'th entry of the vector equals 1 if haplotype m is below the coalescence event, and 0 otherwise. Then, for every pair of branches, with branches coming from different trees, we calculate the correlation coefficient of these vectors for the coalescence events at the lower ends of the branches and the upper ends of the branches.

Two branches are exactly equivalent if the correlation coefficient on both ends of the branches equal one. We first identify exactly equivalent pairs of branches. For the remaining branches, we require that a pair of branches is equivalent if the correlation coefficients on both ends are greater than 0.9. Notice that a branch can satisfy this condition for more than one branch in the other tree. To guarantee that each branch is associated with at most one other branch, we sort pairs of candidate branches in descending order of the correlation coefficient at the lower end of the branches. We then associate branches in the order at which they appear in this list, where we delete any entry for which one of the branches has already been associated with a different branch.

Once we identified equivalent branches, we calculate the number of mutations m_b on a branch b by adding the number of mutations across all equivalent branches. Next, we calculate the cumulative mutation rate for a branch. We assume that mutations occur at a constant rate of $\theta/2$ per base in coalescence time, where

$\theta = 4\mu N_e$ and μ is the per-generation mutation rate. For each branch b, we denote the mutation rate summed over bases for which b persists by $\theta_b/2$.

2.4.2 Metropolis-Hastings type MCMC to estimate branch lengths in a constant population size

We use a Metropolis-Hastings type MCMC algorithm to estimate branch lengths. The likelihood of observing branch lengths $\boldsymbol{t} = \{t_b\}_{b=0,\ldots,2N-2}$, conditional on the number of mutations on branches $\boldsymbol{m} = \{m_b\}_{b=0,\ldots,2N-2}$, is given by

$$P(\boldsymbol{t}|\boldsymbol{m}) \propto P(\boldsymbol{t})P(\boldsymbol{m}|\boldsymbol{t}) = P(\boldsymbol{t}) \prod_{b=0}^{2N-2} P(m_b|t_b), \tag{2.33}$$

where $P(m_b|t_b)$ is Poisson distributed with mean $\theta_b t_b/2$ and the prior $P(\boldsymbol{t})$ is given by the standard coalescent.

We can now define a reversible Markov-Chain with a unique stationary distribution which is the target distribution $P(\boldsymbol{t}|\boldsymbol{m})$. For this, we assign a label $\{1,\ldots,N-1\}$ to every coalescence event. The coalescence event that decreases the number of lineages from k to $k-1$ is stored in variable n_k $(k = 2,\ldots,N)$. We denote the time while k ancestors exist by τ_k $(k = 2,\ldots,N)$. Notice that the branch lengths are uniquely determined by $\boldsymbol{n} = \{n_k\}_{k=2,\ldots,N}$ and $\boldsymbol{\tau} = \{\tau_k\}_{k=2,\ldots,N}$. In particular, the coalescent prior is given by

$$P(\boldsymbol{t}) = P(\boldsymbol{n})P(\boldsymbol{\tau}) = P(\boldsymbol{n}) \prod_{k=2}^{N} P(\tau_k), \tag{2.34}$$

where $\boldsymbol{n}$ is uniformly distributed over all possible orders of coalescence events and τ_k is exponentially distributed with rate $\binom{k}{2}$. We initialise $\boldsymbol{n}$ and $\boldsymbol{\tau}$ as described in Sections 2.4.3 and 2.4.4. In every step of the Metropolis-Hastings algorithm, we propose a change in $\boldsymbol{n}$ with probability q and a change in $\boldsymbol{\tau}$ with probability $1-q$. In our implementation, we have chosen $q = 0.8$.

For a change in $\boldsymbol{n}$, we first choose one coalescence event $e_1 = n_k$ uniformly at random. This is the event that decreases the number of lineages from k to $k-1$. We then propose to swap e_1 with another coalescence event $e_2 = n_h$ chosen uniformly at random from any event with age between e_1's parent event and e_1's

daughter events. If the proposal is accepted, event e_1 now decreases the number of lineages from h to $h-1$ and e_2 decreases the number of lineages from k to $k-1$. We discard any proposal that would violate topological constraints of the tree if all other events retain their times. In a swap of the order of coalescence events, we keep $\boldsymbol{\tau}$ fixed. Such a swap therefore changes the lengths of six branches, but keeps all other branch lengths fixed. Denote these six branches by $b_1, \ldots, b_6$. By using Eqs. (2.33) and (2.34) and noticing that the proposal distribution is symmetric, the acceptance probability is given by

$$\min\left(1, \frac{P(\tilde{\boldsymbol{t}}|\boldsymbol{m})}{P(\boldsymbol{t}|\boldsymbol{m})}\right) = \min\left(1, \prod_{\ell=1}^{6} \frac{P(m_{b_\ell}|\tilde{t}_{b_\ell})}{P(m_{b_\ell}|t_{b_\ell})}\right), \tag{2.35}$$

where $\tilde{\boldsymbol{t}} = \{\tilde{t}_b\}_{b=0,\ldots.2N-2}$ are the proposed branch lengths.

For a change in $\boldsymbol{\tau}$, we choose one τ_k uniformly at random. We propose $\tilde{\tau}_k$ from an exponential distribution with expectation τ_k. A change in τ_k changes the length of k branches that are present at the time while k ancestors remain. Denote these k branches by $b_1, \ldots, b_k$. The remaining branch lengths remain unchanged. By using Eqs. (2.33) and (2.34), the acceptance probability is given by

$$\min\left(1, \frac{\tilde{\tau}_k}{\tau_k} \exp\left[-\frac{\tilde{\tau}_k}{\tau_k} + \frac{\tau_k}{\tilde{\tau}_k} - \binom{k}{2}(\tilde{\tau}_k - \tau_k)\right] \prod_{\ell=1}^{k} \frac{P(m_{b_\ell}|\tilde{t}_{b_\ell})}{P(m_{b_\ell}|t_{b_\ell})}\right). \tag{2.36}$$

Using this MCMC algorithm, we calculate the mean age of every coalescence event. We then calculate the branch lengths as the differences in the mean ages of coalescence events. This guarantees that the time from any tip to the root is equal, a property also known as ultrametric.

Determining convergence of an MCMC sampler to the stationary distribution is not straight-forward and can require running and comparing multiple Markov chains which is computationally intensive. In our implementation, we instead fix the number of iterations to a conservative value. We initially perform $\max\{10N, 1000\}$ burn-in iterations. We then apply the algorithm until every τ_k ($k = 2, \ldots, N$) has had at least 20 proposals and then terminate the MCMC algorithm, conditional on all branch lengths being positive, and continue until the latter condition is satisfied otherwise.

2.4.3 Initialising the order of coalescence events

We initialise the order of coalescence events $\boldsymbol{n}$ by applying a simple MCMC algorithm. In the standard coalescent, any order of coalescence events is equally likely, provided that the order does not contradict the topological constraints of the tree. We therefore propose the following swap moves. We choose a coalescence event n uniformly at random and propose a swap with another coalescence event n'. To ensure that we do not contradict tree topology after swapping the two events, we choose n' from coalescence events that are between the parent and the children of n. We then assert whether n is between the parent and children of n'. If this is the case, we accept the proposal. We accept a proposal with probability 1 because the transition probabilities are symmetric and any order of coalescence events is equally likely. We initialise the order of coalescence events by the order obtained after proposing N^2 swap moves.

2.4.4 Initialising the time while k ancestors remain

After initialising $\boldsymbol{n}$,we initialise the times τ_k while k ancestors remain using the MLE of τ_k conditional on a fixed order of coalescence events. For each k, let $b_1, \ldots, b_k$ be the k branches while there are k lineages remaining in the tree and define $\widetilde{\boldsymbol{m}}_k = \{\widetilde{m}_{b_1,k}, \ldots, \widetilde{m}_{b_k,k}\}$, where $\widetilde{m}_{b,k}$ is the number of mutations that are on branch b in the interval while there are k lineages remaining. We set up an EM algorithm to estimate the MLE of the parameters $\boldsymbol{\tau}$, given the data $\boldsymbol{m}$ and the unobserved variables $\widetilde{\boldsymbol{m}}_k = \{\widetilde{m}_{b_i,k}\}_{i=1,\ldots,k}$ $(k = 2, \ldots, N)$ corresponding to the number of mutations on branches b_i while k ancestors remain. We denote by $\hat{\boldsymbol{\tau}}^{(s)} = \{\hat{\tau}_k^{(s)}\}_{k=N,\ldots,2}$ the estimate of the MLE of τ_k after s iterations. We initialise $\hat{\tau}_k^{(0)} = \binom{k}{2}^{-1}$.

We fix k. The update rule for the EM algorithm is given by the expectation of the log likelihood function $\log P\left(\tau_k|\widetilde{\boldsymbol{m}}_k\right)$, where the expectation is taken conditional on the data and parameters from the previous iteration. We obtain,

$$\hat{\tau}_k^{(s+1)} = \arg\max_{\tau_k} E\left[\log P\left(\tau_k|\widetilde{\boldsymbol{m}}_k\right)\middle|\boldsymbol{m}, \hat{\boldsymbol{\tau}}^{(s)}\right]. \tag{2.37}$$

We first calculate $\log P\left(\tau_k|\widetilde{\boldsymbol{m}}_k\right)$. By using Bayes' theorem, we obtain

$$\begin{aligned} P\left(\tau_k|\widetilde{\boldsymbol{m}}_k\right) &\propto P(\widetilde{\boldsymbol{m}}_k|\tau_k)P(\tau_k) \\ &= P(\tau_k)\prod_{\ell=1}^{k} P(\widetilde{m}_{b_\ell,k}|\tau_k), \end{aligned} \tag{2.38}$$

where $P(\tau_k)$ is an exponential distribution with rate $\binom{k}{2}$ and $P(\widetilde{m}_{b_\ell,k}|\tau_k)$ is a Poisson distribution with mean $\theta_{b_\ell}\tau_k/2$. By using this together with Eq.(2.38), we obtain

$$P(\tau_k|\widetilde{\boldsymbol{m}}_k) \propto \tau_k^{\sum_{\ell=1}^{k}\widetilde{m}_{b_\ell,k}} \exp\left[-\left(\sum_{\ell=1}^{k}\frac{\theta_{b_\ell}}{2}+\binom{k}{2}\right)\tau_k\right]. \tag{2.39}$$

By taking logarithms on both sides of Eq. (2.39), we obtain

$$\log P(\tau_k|\widetilde{\boldsymbol{m}}_k) = \left(\sum_{\ell=1}^{k}\widetilde{m}_{b_\ell,k}\right)\log(\tau_k) - \left(\sum_{\ell=1}^{k}\frac{\theta_{b_\ell}}{2}+\binom{k}{2}\right)\tau_k + \text{const.} \tag{2.40}$$

We substitute Eq. (2.40) into Eq. (2.37) and obtain

$$\hat{\tau}_k^{(s+1)} = \arg\max_{\tau_k}\sum_{\ell=1}^{k} E\left[\widetilde{m}_{b_\ell,k}\middle|\boldsymbol{m},\hat{\boldsymbol{\tau}}^{(s)}\right]\log(\tau_k) - \left(\sum_{\ell=1}^{k}\frac{\theta_{b_\ell}}{2}+\binom{k}{2}\right)\tau_k. \tag{2.41}$$

To find the τ_k maximising Eq. (2.41), we take the derivative with respect to τ_k, we obtain the condition

$$\sum_{\ell=1}^{k} E\left[\widetilde{m}_{b_\ell,k}\middle|\boldsymbol{m},\hat{\boldsymbol{\tau}}^{(s)}\right]\frac{1}{\hat{\tau}_k^{(s+1)}} - \left(\sum_{\ell=1}^{k}\frac{\theta_{b_\ell}}{2}+\binom{k}{2}\right) = 0. \tag{2.42}$$

After reorganizing the terms, we obtain

$$\hat{\tau}_k^{(s+1)} = \frac{\sum_{\ell=1}^{k} E\left[\widetilde{m}_{b_\ell,k}\middle|\boldsymbol{m},\hat{\boldsymbol{\tau}}^{(s)}\right]}{\sum_{\ell=1}^{k}\frac{\theta_{b_\ell}}{2}+\binom{k}{2}}. \tag{2.43}$$

It therefore remains to calculate $E\left[\widetilde{m}_{b_\ell,k}|\boldsymbol{m},\hat{\boldsymbol{\tau}}^{(s)}\right]$ for $\ell = 1,\ldots,k$. In the standard coalescent, mutation events are uniformly distributed on each branch. Therefore, the likelihood that an event on branch b falls within the interval while k lineages are remaining is given by $\hat{\tau}_k^{(s)}/t_b^{(s)}$. Therefore, the likelihood that $\widetilde{m}_{b,k}$ mutations fall into that interval is given by the binomial distribution

$$P(\widetilde{m}_{b,k}|m_b,\hat{\tau}_k^{(s)}) = \binom{m_b}{\widetilde{m}_{b,k}}\left(\frac{\hat{\tau}_k^{(s)}}{t_b^{(s)}}\right)^{\widetilde{m}_{b,k}}\left(1-\frac{\hat{\tau}_k^{(s)}}{t_b^{(s)}}\right)^{m_b-\widetilde{m}_{b,k}}. \tag{2.44}$$

It follows that

$$E\left[\sum_{\ell=1}^{k}\widetilde{m}_{b_\ell,k}\middle|\boldsymbol{m},\hat{\boldsymbol{\tau}}^{(s)}\right]=\sum_{\ell=1}^{k}\frac{m_{b_\ell}\hat{\tau}_k^{(s)}}{t_{b_\ell}^{(s)}}. \tag{2.45}$$

By substituting Eq. (2.45) into Eq. (2.43), we obtain

$$\hat{\tau}_k^{(s+1)}=\frac{\sum_{\ell=1}^{k}m_{b_\ell}\frac{\hat{\tau}_k^{(s)}}{t_{b_\ell}^{(s)}}}{\sum_{\ell=1}^{k}\frac{\theta_{b_\ell}}{2}+\binom{k}{2}}. \tag{2.46}$$

We iterate $\hat{\boldsymbol{\tau}}^{(s)}$ until convergence using Eq. (2.46).

3

Estimating coalescence rates through time

Contents

Assuming a randomly mating panmictic population, the rates at which coalescence events occur in a genealogy inform us about historical population sizes, which we calculate as the inverse of the coalescence rate (see Section 1.3). In the branch length estimation step of Relate (Sections 2.4.2 to 2.4.4), we assumed a known constant population size through time. Here, I present an extension of the branch length estimation algorithm of Relate that relaxes this assumption: We jointly fit branch lengths, as well as a varying population size through time.

The method iteratively applies two steps: First, it estimates coalescence rates given a genealogy, and second, it reestimates branch lengths using the updated coalescence rates. This yields a self-contained method for inferring the demographic history and branch lengths, and should, for instance, improve age estimates of mutations. It can also be used to infer separation histories between diverged populations or track fine-scale population structure through time (see Section 3.2.1).

3.1 Existing methods

Historical population size changes can leave signatures on genetic diversity. For instance, a recent population bottleneck may result in decreased genetic diversity and dearth of rare variants, because all present-day samples are derived from a small set of ancestral individuals [111]. A number of methods utilise such effects to estimate historical population sizes from present-day DNA samples.

Exploiting the observation that population size changes leave signatures on the distribution of observed allele frequencies, a popular approach has been to reverse-engineer population sizes from site-frequency spectra. The ∂a∂i method models the site-frequency spectrum of multiple populations using a diffusion approximation [58] which is used to calculate the expected site-frequency spectrum given some specified demographic history. Assuming each site is independent of any other site and using a composite-likelihood approach, ∂a∂i calculates the likelihood of the observed site-frequency spectrum assuming sampling from the expected site-frequency spectrum. It then selects the demographic history that assigns the maximum likelihood to the observed site-frequency spectrum. We note that theoretical arguments have shown that estimation of historical population sizes from site-frequency spectra is an ill-posed problem, with population sizes being not uniquely determined even in idealistic scenarios of panmictic populations with sufficiently large population sizes and exactly known site-frequency spectra [108]. It is therefore necessary to impose reasonable smoothness constrains to restrict the space of possible histories.

An alternative line of research has focussed on inference of population size histories from whole-genome sequencing data. The pairwise sequentially Markovian coalescent (PSMC) approach infers population size histories from a single diploid genome [88]. Underlying the PSMC approach is an HMM, in which observed states are given by 0-1 observations of haplotypes, and hidden states are TMRCAs of the pair of haplotypes at a genomic position. Because the SMC model is Markovian along the genome, the probability of a TMRCA change (i.e., transition in hidden state) is also Markovian and governed by the probability of a recombination and

subsequent coalescence. Given recombination and coalescence rates, the HMM can be used to numerically calculate the probability of the observed data. Using the Baum-Welch algorithm, a version of the Expectation-maximization (EM) algorithm to determine maximum likelihood estimates of unknown parameters of an HMM, the PSMC approach infers the most likely coalescence rate given the observed data.

The multiple SMC (MSMC) approach extends the PSMC framework to more than two haplotypes by constructing an HMM in which hidden states represent the time to the first coalescence of any pair of N haplotypes [136]. This is expected to improve accuracy of population sizes particularly in the recent past, because the expected time to the most-recent coalescence scales as $1/\binom{N}{2}$, i.e., inversely with the square of the number of haplotypes N, in units of $2N_\mathrm{e}$ generations with N_e being the long-term average population size. While MSMC generalises PSMC to more haplotypes, MSMC is typically run with at most 8 diploid samples. Scalability of MSMC is constrained by a large hidden state space that grows quadratically in the number of haplotypes N, requiring integration over N^2 states per time interval and genomic position.

SMC++ extends the PSMC framework in a different direction and scales to hundreds of samples [155]. The HMM used in SMC++ has identical hidden states to PSMC, i.e., TMRCAs of a pair of haplotypes, but emits not only the haplotype states of this pair, but also the derived-allele frequency of $N-2$ unlabelled haplotypes at any focal genomic position. The emission probabilities in SMC++ are therefore based on the sample's site frequency spectrum conditional on the TMRCA of two labelled haplotypes and combines the whole-genome approach of PSMC with earlier site-frequency based approaches. ASMC is a modified version of SMC++, allowing for inference of coalescence times from genotype array data by accounting for SNP frequency or ancestry-specific ascertainment biases [117].

In summary, inference of demographic history have followed mainly two strategies, namely either using site-frequency spectra (not utilising haplotype information even if available) or whole-genome haplotype/genotype information (often relying on downsampling). Some of these methods are applicable to unphased data. For

methods based on site-frequency spectra, theoretical limitations due to identifiability of true demographic histories are well understood, however we note that issues relating to identifiability, and smoothness constraints to overcome potential issues, likely apply to all methods discussed here.

The methods presented in this section focus primarily on inference of demographic history and in particular, they do not estimate full genealogies. With Relate, we aim to jointly fit demographic histories and branch lengths without downsampling, leading to demographic histories that are consistent with any downstream population genetic analyses. We highlight that this approach aims to use all available data, but relies on accurate inference of haplotype phase and tree topologies; which method yields the most accurate demographic histories in practice is an empirical question and not the primary focus of this chapter.

3.2 Method for jointly estimating coalescence rates and branch lengths

In Relate, we initially fit a constant population size model as described in Sections 2.4.2 to 2.4.4. In practice, we use $2N_e = 30,000$ for humans. Here, I will describe how we can use these branch length estimates as the initial state for an iterative algorithm in which we repeatedly estimate coalescence rates and update branch lengths. This iterative algorithm proceeds as follows.

We first estimate a population-wide coalescence rate (see Section 3.2.1). Second, we estimate a population-wide mutation rate over time, where we divide time into epochs and calculate the quotient of the number of mutations that occurred in an epoch and the total branch length in that epoch. This estimated mutation rate is used in the third step, which is a heuristic step intended to speed-up convergence. In this step, we multiply the population-wide coalescence rate by the quotient of the pre-specified constant mutation rate μ ($\mu = 1.25 \times 10^{-8}$ for humans) and the

estimated mutation rate over time. Because we assume a constant mutation rate through time in our model, any changes in the mean mutation rate should, in theory, be fully absorbed by the coalescence rate and any remaining changes in the mean mutation rate are therefore fully confounded. This heuristic step therefore reflects this assumption of a constant mean mutation rate. Finally, using this rescaled coalescence rate estimate as input, we reestimate branch lengths, now under a variable population size model (see Section 3.2.2). We then return to the first step.

We terminate this algorithm after at least two and at most five iterations, where termination with less than five iterations happens if the mean absolute error between the inferred mutation rate through time and the pre-specified constant mutation rate μ is less than 0.01μ. This convergence criterion uses the fact that our approach tries to fit a constant average mutation rate through time (absorbing any potential fluctations in average mutation rates in population size estimates). A constant estimated mutation rate therefore reflects perfect convergence of the method.

To speed up convergence and computation time, we apply this algorithm only to trees with sufficiently many mapped mutations, which in our implementation is set as N mutations on the tree, where N is the number of haplotypes. Using the output coalescence rates, we then calculate a final estimate of the branch lengths by reestimating branch lengths for all trees for a final time.

3.2.1 Estimating the coalescence rate for a pair of haplotypes

Here, we derive an MLE for the historical coalescence rates for a pair of haplotypes, given a genealogy with estimated branch lengths. This MLE is a special case of coalescence rate estimators for arbitrary sample sizes developed by Dr Marie Forest [43]. We note that other approaches for estimating coalescence rates from genealogies exist; in particular Ref. [116] proposes such an MLE but assumes a full ARG as input and is therefore not applicable to Relate-inferred genealogies.

To obtain a population-wide estimate of the coalescence rate, we take the mean over all pairs of haplotypes in a population. We note that this is not the MLE for

the population-wide coalescence rate assuming a panmictic population. In practice, this approach, though heuristic, allows us to avoid having to assume panmixia in this step of the algorithm. It also enables us to calculate coalescence rates for any subset of haplotypes, cross-coalescence rates between subpopulations, or track fine-scale population structure through time.

To estimate pairwise coalescence rates, we divide time into epochs. Within an epoch, we assume that the coalescence rate remains constant. For every pair of haplotypes, we estimate these piecewise constant coalescence rates using an MLE. Denote epochs by $e = 0, \ldots, E$, where epoch e begins at time T_e and ends at time T_{e+1}. We denote the coalescence rate in epoch e by $\gamma(e)$.

We denote the time at which the two haplotypes coalesce in tree z by t_z. Also, we denote by e_z the index of the epoch in which the haplotypes coalesce. We therefore have

$$T_{e_z} \leq t_z < T_{e_z+1}. \tag{3.1}$$

Conditioning on tree topology and branch lengths, the probability that the two haplotypes coalesce at time t_z is given by

$$P(t_z) = \gamma(e_z) \exp\left[-\gamma(e_z)(t_z - T_{e_z})\right] \left[\prod_{e=1}^{e_z} \exp\left[-\gamma(e-1)(T_e - T_{e-1})\right]\right]. \tag{3.2}$$

By taking logarithms on both sides of Eq. (3.2), we obtain

$$\log P(t_z) = \log \gamma(e_z) - \gamma(e_z)(t_z - T_{e_z}) - \sum_{e=1}^{e_z} \gamma(e-1)(T_e - T_{e-1}). \tag{3.3}$$

Assuming independence across trees, the log-likelihood for the whole genome is given by $\sum_{z=0}^{M} \log P(t_z)$, where M is the number of trees built. By differentiating with respect to $\gamma(e)$, we obtain an MLE given by

$$\hat{\gamma}(e) = \frac{n_e}{\sum_{z:e=e_z}(t_z - T_e) + \sum_{z:e<e_z}(T_e - T_{e-1})}, \tag{3.4}$$

where n_e denotes the number of trees for which the two haplotypes coalesce in epoch e.

In Fig. 3.1, we apply the MLE derived here to four simulated demographic histories, where we simulate 200 Mb for 200 haplotypes, assigning half of the

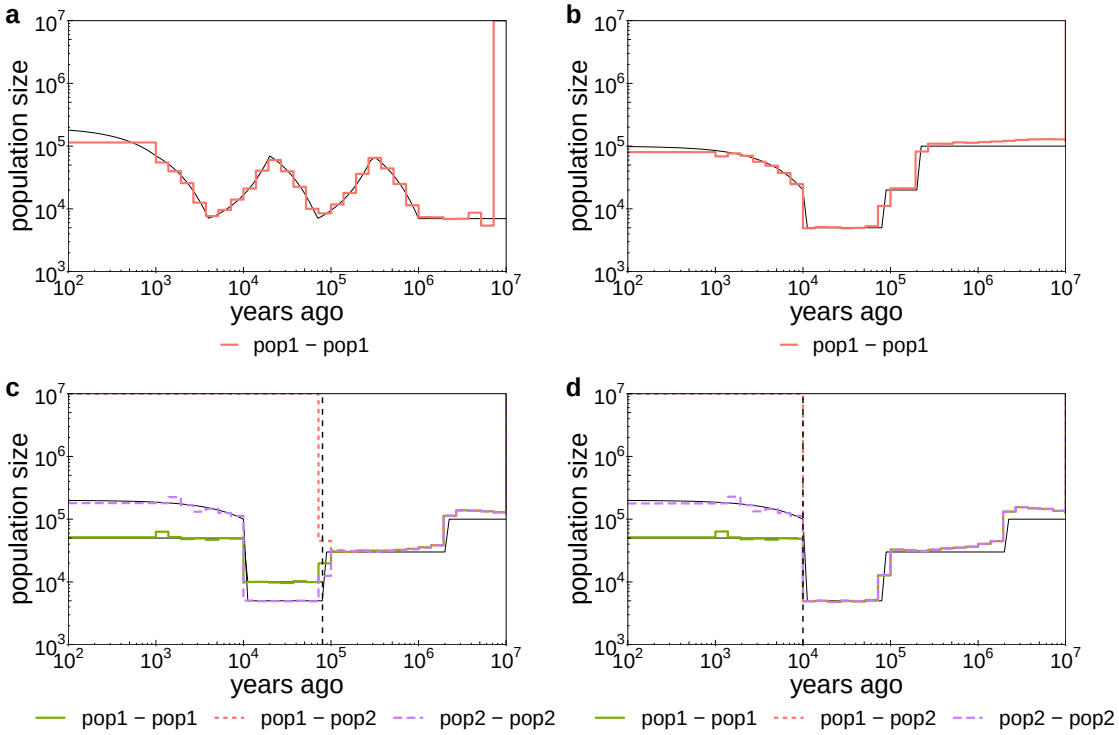


Figure 3.1: Maximum-likelihood estimator for coalescence rates.
Simulations of a single population with **a**, a sawtooth history and **b**, a discrete bottleneck. We also simulate two diverging populations that separated 80,000 years ago (in **c**) and 10,000 years ago (in **d**). In each case, we simulate one realization with $N = 200$ haplotypes and 200Mb. Recombination rates are taken from chromosome 1 and $\mu = 1.25 \times 10^{-8}$. The true ARGs are converted into Relate format and coalescence rates are calculated for every pair of haplotypes using the maximum-likelihood estimator described in Section 3.2.1 and then averaged according to affiliation of haplotypes.

samples to each population for the two examples (Fig. 3.1**c** and **d**) with two divergent populations. We find that the demographic histories are accurately recovered given the true trees, with only some minor discrepancies in the ancient past for the discrete bottleneck scenarios.

3.2.2 Reestimating branch lengths using a coalescent prior with variable population sizes

We reestimate branch lengths using an MCMC sampling identical to that described in Section 2.4.2, but with a modified Eq. (2.36) as shown below, reflecting a coalescent prior that incorporates piecewise-constant coalescence rates. The MCMC sampler is initialised using the branch lengths of the input genealogies.

Whenever τ_k is updated in an MCMC iteration by a proposed value $\tilde{\tau}_k$, all coalescence events older than this event are updated by $\Delta\tau = \tilde{\tau}_k - \tau_k$. Therefore, older events may now coalesce in a different time epoch to before, which we need to reflect in the acceptance probability of $\tilde{\tau}_k$. We modify Eq. (2.36), which states the acceptance probability of a proposed update $\tilde{\tau}_k$ of τ_k, to reflect a variable population size and obtain

$$\min\left(1, \frac{\tilde{\tau}_k}{\tau_k}\exp\left[-\frac{\tilde{\tau}_k}{\tau_k}+\frac{\tau_k}{\tilde{\tau}_k}\right]\prod_{\ell=1}^{k}\frac{P(m_{b_\ell}|\tilde{t}_{b_\ell})}{P(m_{b_\ell}|t_{b_\ell})}\prod_{m=2}^{k}c_m\right), \tag{3.5}$$

for c_m $(m = 2, \ldots, k)$ which we will derive below.

Let us define a function $\eta(t) \in \{0, \ldots, E\}$ mapping time t to its corresponding epoch. For a piecewise constant coalescence rate as defined in Section 3.2.1, the time while τ_k ancestors remain, conditional on $(\tau_\ell)_{\ell=k+1,\ldots,N}$ has density

$$f_{\tau_k|\tau_{k+1},\ldots,\tau_N} = \binom{k}{2}\gamma\left(\eta\left(\sum_{\ell=k}^{N}\tau_\ell\right)\right)\exp\left[-\binom{k}{2}\int_{\sum_{\ell=k+1}^{N}\tau_\ell}^{\sum_{\ell=k}^{N}\tau_\ell}\gamma(\eta(\tau))d\tau\right]. \tag{3.6}$$

It follows that the ratio of the prior probabilities of $\tilde{\tau}_k$ and τ_k, conditional on $\tau_{k+1}, \ldots, \tau_N$, is given by

$$c_k = \frac{\gamma\left(\eta\left(\Delta\tau + \sum_{\ell=k}^{N}\tau_\ell\right)\right)}{\gamma\left(\eta\left(\sum_{\ell=k}^{N}\tau_\ell\right)\right)}\exp\left[-\binom{k}{2}\int_{\sum_{\ell=k}^{N}\tau_\ell}^{\Delta\tau+\sum_{\ell=k}^{N}\tau_\ell}\gamma(\eta(\tau))d\tau\right]. \tag{3.7}$$

Because we also update the times of all events older than event k, we need to calculate the ratio of prior probabilities for these events, and we obtain for $m < k$,

$$c_m = \frac{\gamma\left(\eta\left(\Delta\tau + \sum_{\ell=m}^{N}\tau_\ell\right)\right)}{\gamma\left(\eta\left(\sum_{\ell=m}^{N}\tau_\ell\right)\right)}\exp\left[-\binom{m}{2}\int_{\sum_{\ell=m+1}^{N}\tau_\ell}^{\sum_{\ell=m}^{N}\tau_\ell}\left[\gamma(\eta(\tau+\Delta\tau)) - \gamma(\eta(\tau))\right]d\tau\right]. \tag{3.8}$$

Substituting Eqs. (3.7) and (3.8) in Eq. (3.6), we obtain the acceptance probability of a proposed change $\Delta\tau$ to the time while k ancestors remain. In particular, we note that if $\gamma(e) \equiv 1$ for all epochs e, we obtain $c_k = \exp\left[-\binom{k}{2}\Delta\tau\right]$ and $c_m = 1$ $(m < k)$, reducing Eq. (3.6) to Eq. (2.36).

4
Performance on simulated data

Contents

In this chapter, we test Relate on simulated data. We compare runtime and accuracy to state-of-the-art alternative approaches, in particular ARGweaver [126], Rent+ [103], as well as tsinfer [81]. ARGweaver samples from a discretised version of Sequentially Markovian Coalescent (see Section 1.5), which is an approximation of the standard coalescent with recombination. Under idealistic simulation scenarios, ARGweaver samples from approximately the correct model and we therefore expect it to perform well on simulated datasets. Rent+ is a non-parametric heuristic method that estimates genealogies more efficiently than ARGweaver. Tsinfer is a recent method scaling to potentially millions of samples, although it currently does not infer branch lengths (see Section 1.1 for a discussion on existing methods).

We then compare Relate's estimates of historical effective population sizes to those of alternative specialist methods MSMC [136] and SMC++ [155] (see Sec-

tion 3.1 for a discussion of these methods). These methods do not infer genealogies, but have been very successful in recovering the demographic history of a sample.

Finally, we evaluate Relate on a diverse set of simulation scenarios including perturbations from the infinite-sites assumption, constant mutation rates, or perfect phase. We demonstrate that Relate is robust to violation of these assumptions and can be applied to detect repeat mutations or changes in mutation rates through time.

All simulations in this chapter use msprime [80], which simulates the standard coalescent with recombination [74] assuming the infinite sites model for mutation and recombination [84].

4.1 Runtime

We evaluate runtime as functions of the number of haplotypes and number of simulated bases. For all simulations, we use $\mu = 1.25 \times 10^{-8}$, $2N_e = 30,000$, and recombination rates taken from the 1000 Genomes Project map for chromosome 1 [1]. For comparisons with ARGweaver and Rent+, we simulate $N = 200$ haplotypes, while varying the number of simulated bases L. Relate scales linearly in L (Fig. 4.3 **a**) and appears to be more than $14,000$ times faster than ARGweaver, making ARGweaver computationally infeasible for data sets of more than 200 haplotypes and 2.5Mb. For instance, with $N = 200$ haplotypes and $L = 2.5$Mb bases, Relate estimates a genealogy in approximately 30 seconds, whereas ARGweaver requires more than 85 hours. Rent+ is faster than ARGweaver but about 30 times slower than Relate on small data sets. We tested Rent+ on a data set of $N = 1000$ and $L = 2.5$Mb, where it ran out of memory after approximately 400 CPU hours on a machine with 100GB of RAM. When varying the number of simulated haplotypes N, we observe that Relate scales quadratically with N (Figure 4.3**b**) and can be applied to data sets comprising thousands of samples.

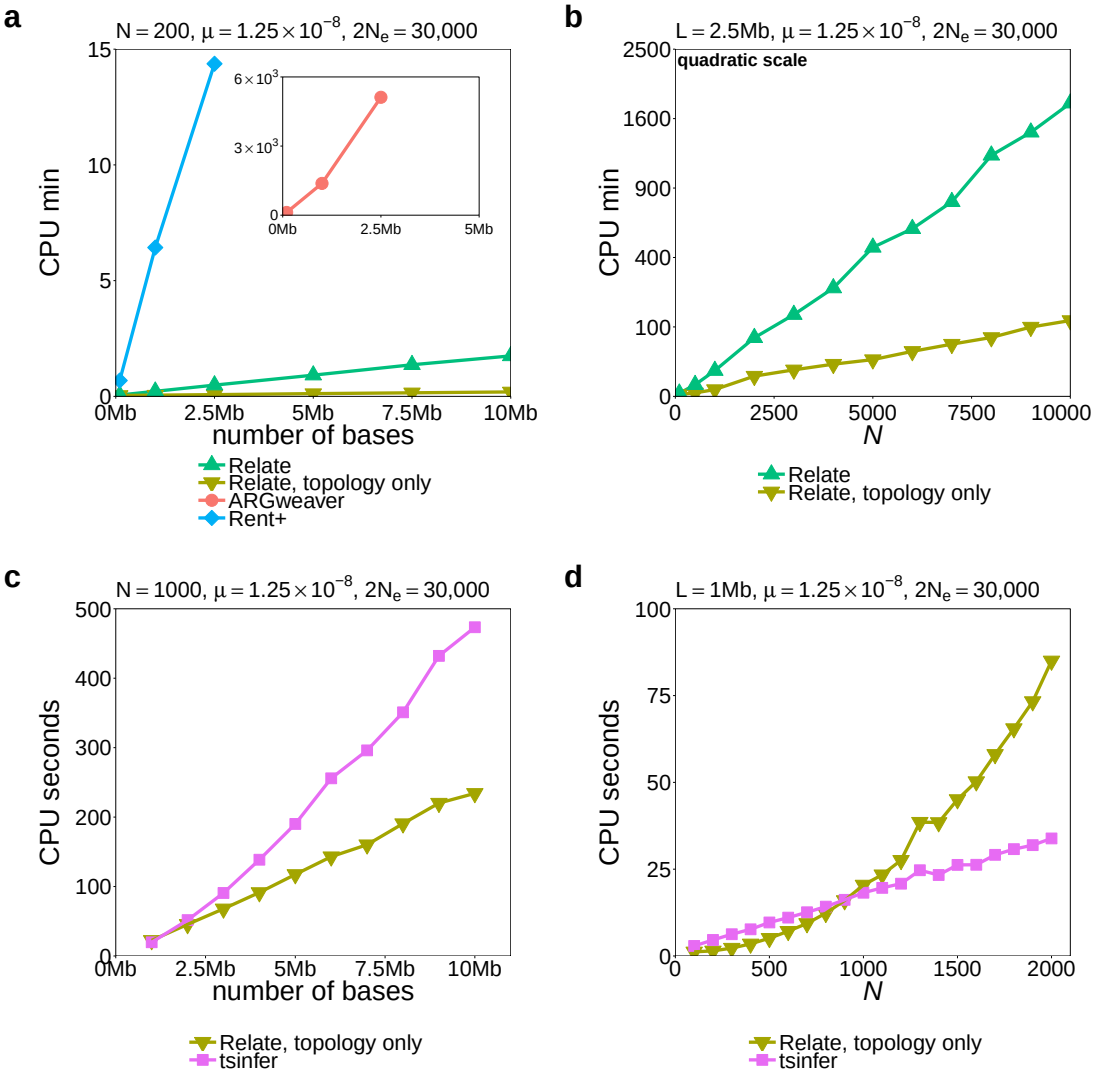


Figure 4.1: Runtime of Relate and alternative methods.
a, Runtimes of Relate and Rent+ in minutes as a function of the number of simulated bases with $N = 200$. The inset shows the runtime of ARGweaver, where runtime is also measured in minutes. We also show the runtime of Relate for only estimating tree topologies. **b**, Runtimes in minutes of Relate as a function of the number of haplotypes with $L = 2.5$Mb. **c**, Runtimes in seconds of tsinfer and Relate (topology only) as a function of the number of simulated bases with $N = 1000$. **d**, Runtimes in seconds of tsinfer and Relate (topology only) as a function of the number of haplotypes with $L = 1$Mb. In **a** and **b**, each point represents one simulation and in **c** and **d**, each point is the mean of 100 simulations. For all simulations, we used $\mu = 1.25 \times 10^{-8}$, $2N_e = 30,000$, and recombination rates taken from chromosome 1.

The branch length estimation step is computationally the most expensive step of the Relate algorithm. To compare runtime with tsinfer, which only estimates tree topologies, we use the runtime of Relate for inferring only the tree topologies. Tsinfer scales almost linearly with the number of simulated bases, but constants appear to be worse compared to Relate (Figure 4.3**c**). For instance, in a simulation with $N = 1000$ haplotypes, Relate is slower than tsinfer on a 1Mb sequence, but more than twice as fast on a 10Mb region. When varying the number of haplotypes ($L = 1$Mb), Relate is faster than tsinfer for $N < 800$, but is then outperformed by tsinfer which has an impressive linear scaling with sample size (Figure 4.3**d**).

4.2 Accuracy of TMRCAs and mutation ages

A quantity influenced by both the tree topology and branch lengths is the time to the most-recent common ancestor (TMRCA) of a pair of haplotypes. For a tree of N haplotypes, we calculate these TMRCAs for every pair of haplotypes which we store in a vector of length $\binom{N}{2}$. This vector uniquely defines a binary tree with branch lengths, such that there is a one-to-one mapping between binary trees and these vectors (notice that not all vectors in $\mathbb{R}_{\geq 0}^{\binom{N}{2}}$ define a tree).

As one indicator for accuracy of estimated tree topologies and branch lengths, we compute, at each site, this $\binom{N}{2}$ vector for the estimated tree and the true tree. In Figure 4.2, we visualise xy-plots of these vectors across all sites in a heatmap. We find that TMRCA estimates using Relate appear to be unbiased, whereas ARGweaver underestimates the TMRCA for recent coalescences and Rent+ exhibits a non-linear relationship to the true TMRCA (Figure 4.2(a)).

Next, we map mutations to trees as described in Section 2.2.2 to estimate mutation ages. We apply our code to do the same for ARGweaver and Rent+ because the output files of these methods do not provide age estimates and do not specify on which branches the mutations occurred. We find that across all

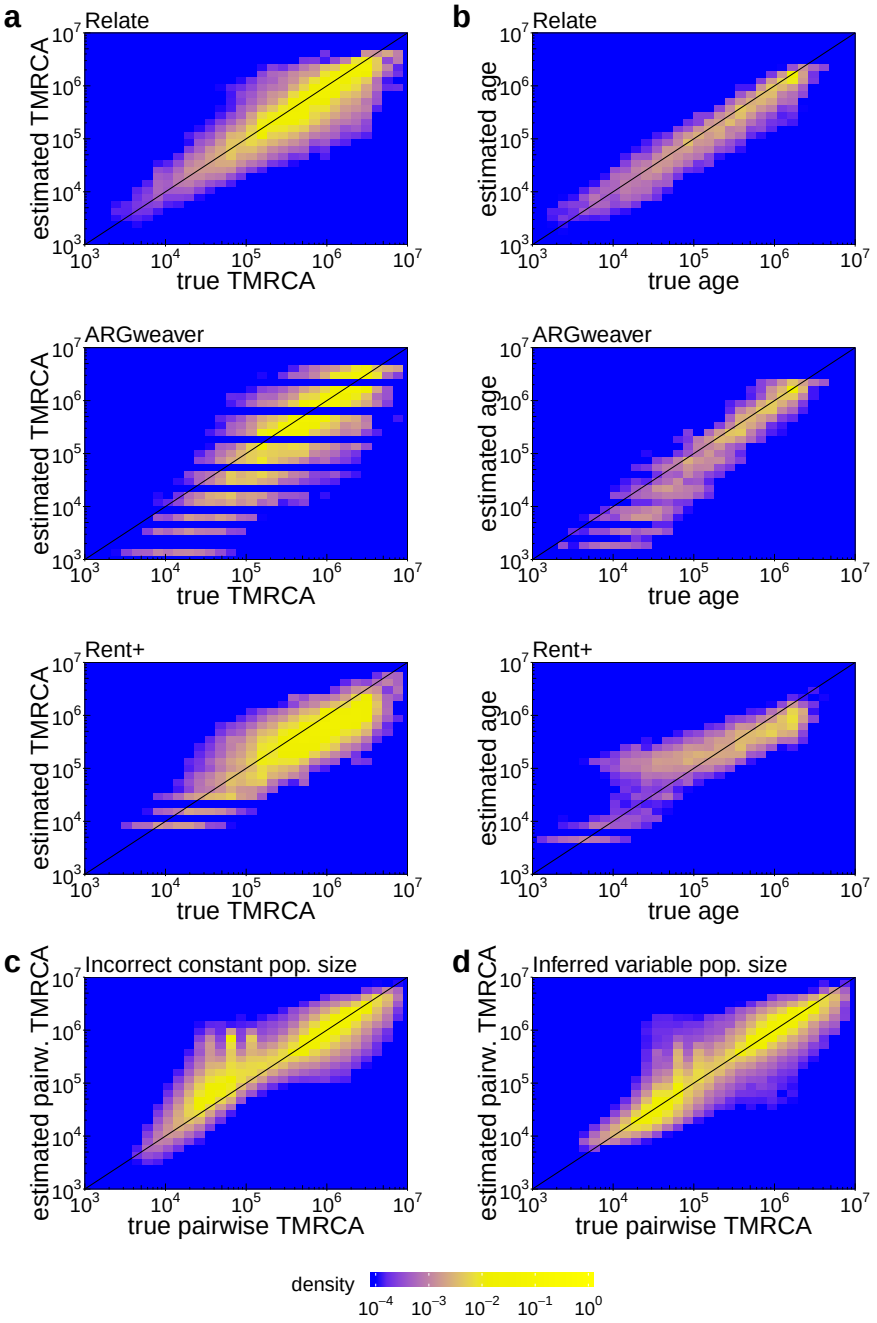


Figure 4.2: Accuracy of pairwise TMRCAs and mutation ages.
a, Time to most recent common ancestors (TMRCAs) between pairs of haplotypes in estimated trees compared to the truth for Relate, ARGweaver, and RENT+. **b**, Estimated ages of mutations plotted against the true age of mutations for Relate, ARGweaver, and RENT+. We determined the age of a mutation by placing it at the midpoint of the branch onto which it maps. The data was simulated with parameters $N = 100$, $2N_e = 40,000$, and $\mu = 1.25 \times 10^{-8}$, and recombination rates taken from a subregion of chromosome 1. **c**, TMRCAs between pairs of haplotypes compared to the truth for a simulated data set with $N = 200$ haplotypes and a population bottleneck resembling that of Europeans. Branch lengths are estimated using a constant population size of $2N_e = 30,000$. **d**, Estimated TMRCAs compared to the truth for the same example as in **c**, where branch lengths and population size history are now jointly inferred.

methods, age estimates of mutations appear to have a smaller deviation from the truth compared to TMRCAs. This may indicate that branches with at least one mutation have less uncertainty in their age estimates. In terms of accuracy, we observe similar trends as for TMRCAs, where our method appears to have the least bias (Figure 4.2**b**).

Additionally, we test Relate on a simulated data set with a variable population size history ($N = 200$). We simulated 200Mb using the population size inferred for 1000 Genomes project GBR individuals (British in England and Scotland) using Relate (see Table A.1 for population labels, see Chapter 5 for details). In Figure 4.2 **c**, we apply Relate assuming a constant population size of $2N_e = 30,000$. Here, we deliberately misspecified the population size history and we observe a clear bias in the estimated TMRCAs, e.g., a coalescence event dated at approximately 10^5 years before present may have occurred between 10^4 to 2×10^6 years before present in the true trees. In Figure 4.2**d**, we jointly infer branch lengths and population sizes for the same data set. We observe that this corrects for most of this bias, highlighting the importance of accounting for the demographic history of the sample.

4.3 Accuracy measured using tree metrics

In addition to measuring accuracy using pairwise TMRCAs, we use three distance metrics that capture differences between trees. For each distance metric, we calculate a genome-wide mean score by weighting the distance between an estimated genealogical tree and the truth according to how long these trees persist in the genome.

First, we compute the Robinson-Foulds metric d_{RF} adapted for rooted binary trees [130]. For each coalescence event, we find the set of present-day descendants, which we call a clade. We count the number of clades that exist in one tree but not the other. We then divide this number by $4N - 2$ such that $d_{\mathrm{RF}} = 1$ if trees are entirely different and $d_{\mathrm{RF}} = 0$ if they are exactly the same. Notice that this

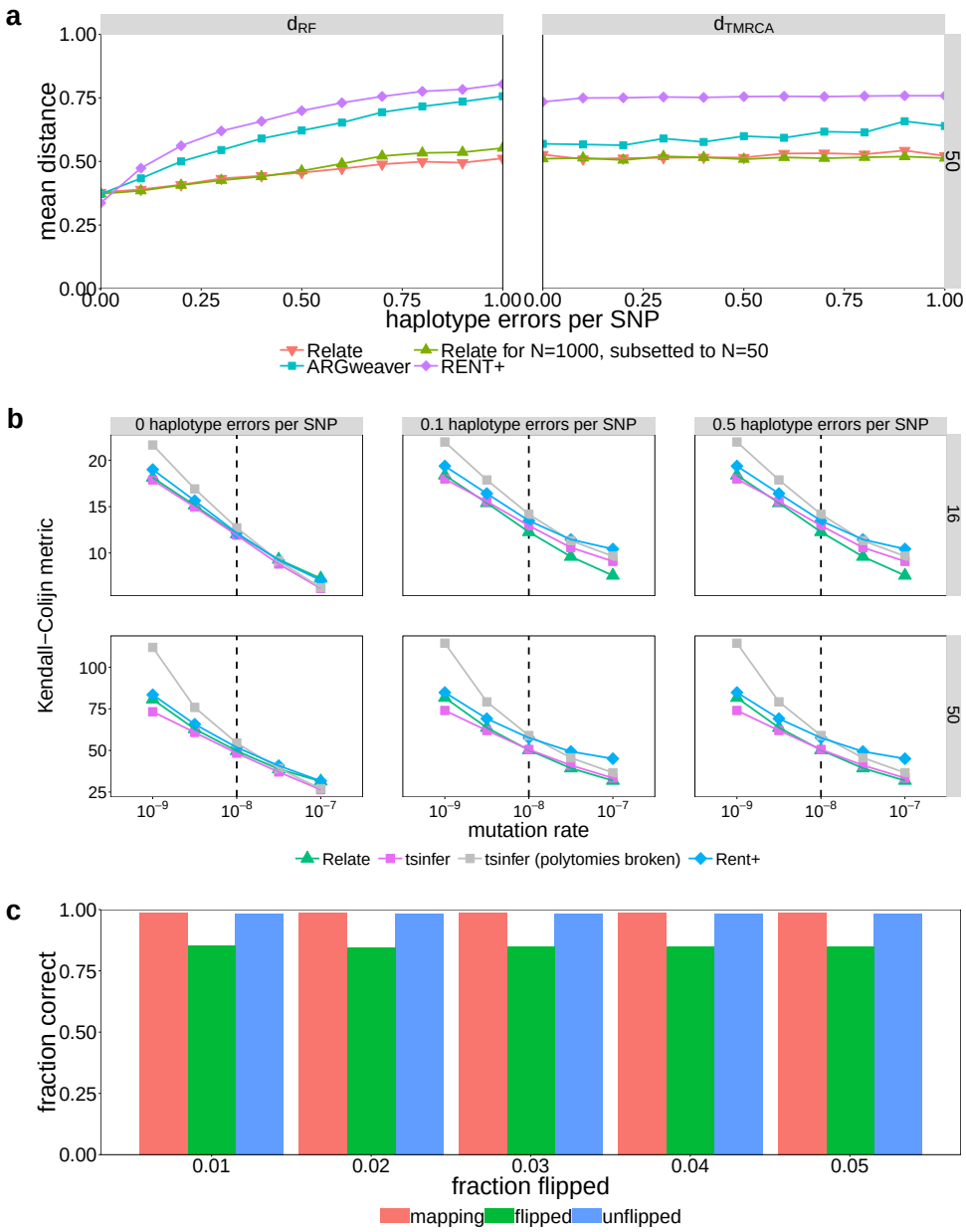


Figure 4.3: Accuracy using tree metrics and robustness to errors in the data. **a**, Robinson-Foulds distance and pairwise TMRCA distance averaged over 2.4Mb for Relate, ARGweaver, and RENT+. We estimate genealogies for $N = 50$ haplotypes at different number of errors. In addition, we show the accuracy of the genealogy corresponding to $N = 50$ haplotypes, embedded in an estimated genealogy for $N = 1000$ haplotypes (see Section 4.3.1 for details). **b** Kendall-Colijn distance averaged over 1Mb for Relate, tsinfer, tsinfer (with polytomies broken at random), and RENT+. We simulated 16 and 50 haplotypes with a constant recombination rate of 10^{-8} (dashed line) and different mutation rates and haplotype errors. For each combination of parameters, we simulated 100 realisations. **c**, Robustness of Relate with respect to randomly introduced flipped mutations. We show the fraction of SNPs mapping to a unique branch, fraction of correctly flipped SNPs, and fraction of correctly unflipped SNPs for Relate. We exclude SNPs at frequency 1, which always map to the tree. We simulate 2.5Mb for $N = 200$ haplotypes with $2N_e = 30,000$.

metric is only dependent on tree topology and not on branch lengths. For two uncorrelated trees drawn from the standard coalescent, the Robinson-Foulds metric equals 1 with probability 1 as $N \to \infty$.

To also compare the accuracy of estimated branch lengths, we define a second metric d_{PTMRCA}, in which we compare the time to the MRCA (TMRCA) of every pair of haplotypes. For each tree, we calculate a vector of lengths $\binom{N}{2}$ containing the TMRCAs between every pair of tips. We then calculate, at every SNP, the mean squared difference between the vectors corresponding to the estimated and true trees and divide the result by the diploid effective population size N_{e}. We note that d_{PTMRCA} inherits its metric properties from the mean square difference together with the fact that for trees T_1 and T_2, we have $d_{\mathrm{PTMRCA}}(T_1, T_2) = 0$ if and only if $T_1 = T_2$. In this metric, the expected score for two uncorrelated trees drawn at random from the standard coalescent model equals 1.

Finally, we calculate the Kendall-Colijn metric for comparison of tree topologies [82]. For this metric, we compute for each tree a vector of length $\binom{N}{2}$. Each entry of this vector stores the number of branches from the MRCA of two haplotypes to the root. We then compute the Eulidean distance between these vectors to define a metric between trees. We note that this metric can be applied to non-binary tree as well as binary trees, although a comparison between non-binary and binary trees is not necessarily meaningful because the trees exist in different subspaces with different constraints.

4.3.1 Impact of errors in the data set

We evaluate Relate at varying levels of errors introduced to the data. We simulate 2.5Mb and $N = 1000$ haplotypes with $2N_{\mathrm{e}} = 30,000$, $\mu = 1.25 \times 10^{-8}$, and recombination rates taken from the 1000 Genomes Project map for chromosome 1 [1]. We then subset this dataset to $N = 50$ haplotypes. We estimate genealogies for the same 50 haplotypes using Relate, RENT+, and ARGweaver. In addition, we estimate the genealogy for all 1000 haplotypes using Relate, and extract the embedded genealogy corresponding to the same 50 haplotypes.

We introduce errors to the 50 haplotypes by first choosing a SNP and a haplotype uniformly at random and then changing this haplotype. We find that in the absence of errors, accuracy across the three methods is similar by the d_{RF} metric, while Relate offers improvements for the d_{PTMRCA} metric. However when we introduce errors, Relate outperforms ARGweaver and RENT+ in both the Robinson-Foulds and PTMRCA metrics (Figures 4.3**a**).

4.3.2 Comparison to tsinfer

We compare accuracy of Relate inferred tree topologies to those of tsinfer [81]. Tsinfer estimates non-binary trees, which is why we use the Kendall-Colijn metric for comparison. In addition, we compare Relate-estimated genealogical trees to tsinfer trees in which any polytomies have been randomly broken to form binary trees. To emulate the simulation scenario in Ref. [81], we simulate 1Mb with a constant diploid population size of $2N_{\mathrm{e}} = 5000$ for 16 and 50 samples (Figure 4.3**b**). We use a constant recombination rate of 10^{-8} and vary the mutation rate, as well as the number of random haplotype errors introduced to the data. For each combination of mutation rate and sample size, we simulate 100 realisations.

We first note that tsinfer with polytomies broken attains substantially worse Kendall-Colijn metric scores than tsinfer particularly for low mutation rates, suggesting that Rent+ and Relate are able to break some polytomies better than random. Apart from this observation, in the case of no haplotype errors, all methods perform similarly, with tsinfer improving, and eventually slightly outperforming Relate and RENT+ at high mutation rates. However, already with the introduction of a small number of haplotype errors, we find that Relate estimates more accurate tree topologies compared to both alternatives, especially for high mutation rates. This suggests that Relate uses a less strict infinite sites assumption, making it more robust to errors. This effect is amplified for higher haplotype error rates, where RENT+ performs substantially worse than Relate or tsinfer.

4.3.3 Impact of incorrect ancestral alleles

We evaluate the robustness of Relate with respect to incorrect identification of ancestral and derived alleles. We simulate 2.5Mb for $N = 200$ haplotypes, with other parameters equal to $2N_e = 30,000$, $\mu = 1.25 \times 10^{-8}$ and human chromosome 1 recombination rates. We find that over 99% of SNPs can be mapped to a unique branch (Figure 4.3**d**). The fraction of correctly unflipped SNPs remains above 98% regardless of the fraction of flips introduced to the data. The fraction of correctly flipped SNPs decreases slightly but stays at around 85%. We note that for SNPs that map to a branch connected to the root of the tree, we cannot identify whether it is flipped or unflipped. We therefore excluded such SNPs from this analysis. In addition, we also excluded singleton mutations because these can always be mapped to a unique branch.

4.4 Accuracy of coalescence rates estimates

We compare our algorithm to MSMC [136] and SMC++ [155] across different simulated population size histories. We simulate 200Mb for $N = 200$ haplotypes and a constant mutation rate of 1.25×10^{-8}. The recombination rates taken from chromosome 1.

The command line used for MSMC is

```
msmc_2.0.0 \
     -t 6 \
     -p 1*2+15*1+1*2 \
     -o msprime.msmc2 \
     msprime.multihetsep.txt
```

where "msprime.multihetsep.txt" denotes the input filename and "msprime.msmc2" denotes the output filename. The command line used for SMC++ is

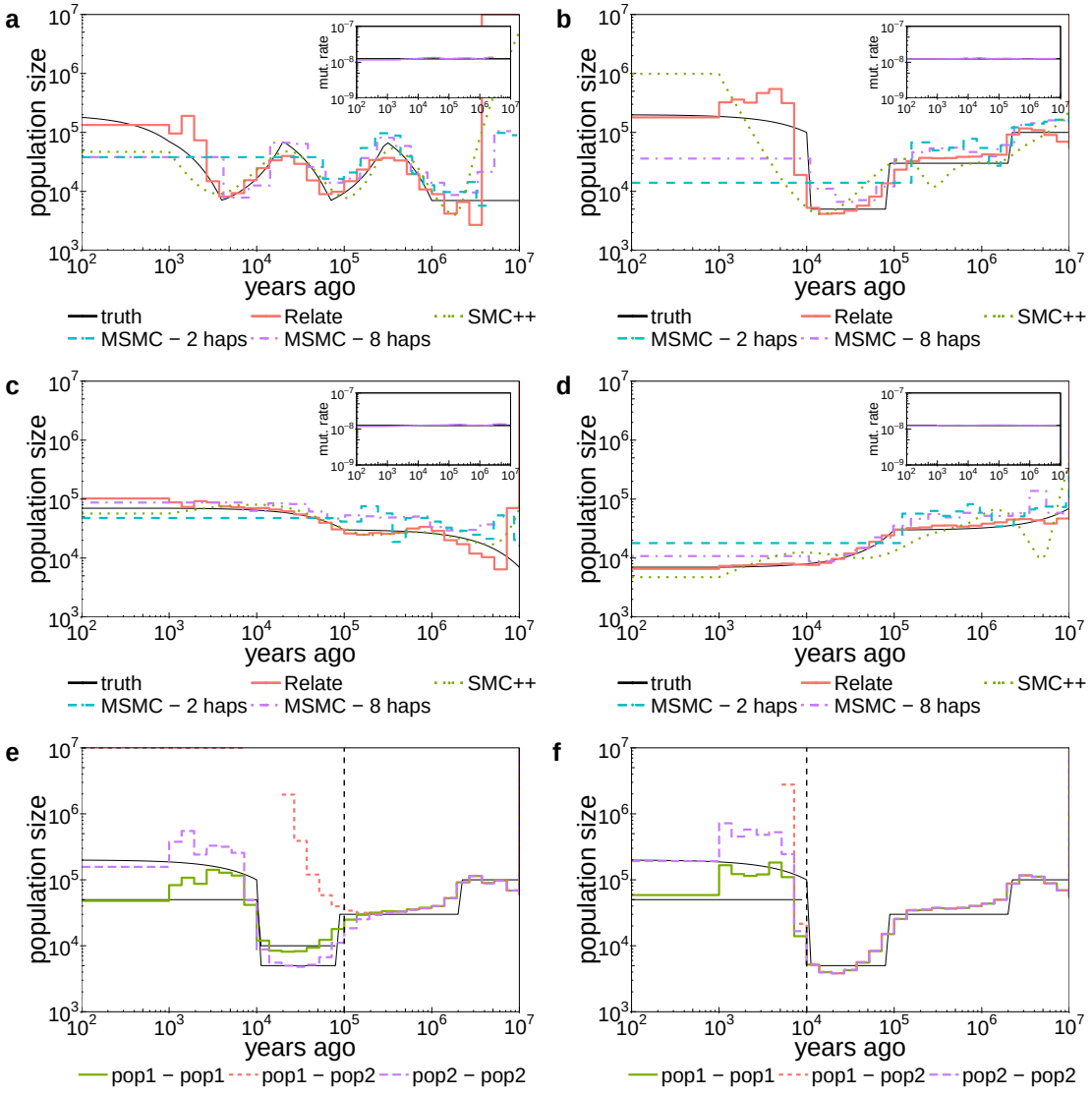


Figure 4.4: Accuracy of population size estimates.
Simulations of a single population with **a** a sawtooth history, **b** a discrete bottleneck, **c** an increasing trend, and **d** a decreasing trend in population size. Estimates using Relate are shown by the red solid line. We applied SMC++ to the same data set and MSMC2 on a subset of 2 and 8 haplotypes. In the insets, we show the mutation rate over time estimated by Relate. We also simulate two diverging populations that separated 80,000 years ago (in **e**) and 10,000 years ago (in **f**). In each case, we simulated one realization with $N = 200$ haplotypes and 200Mb. Recombination rates are taken from chromosome 1 and $\mu = 1.25 \times 10^{-8}$.

```
smc++ estimate \
      --regularization_penalty 5.0 \
      --knots 16 \
      --timepoints 35,100000 \
      1.25e-8 \
      -o analysis/ \
      out/msprime.smc.gz
```

where "analysis/" denotes the directory into which the output of SMC++ is saved and "out/msprime.smc.gz" denotes the input filename. We noticed that the accuracy of SMC++ is sensitive to the choice of parameters "regularization_penalty" and "knots". Our choice of these parameters is based on direct communication with authors of Ref. [155].

We note that Relate uses all phased haplotypes as input, while MSMC and SMC++ use a subset of the available data as input. In particular, SMC++ utilises haplotype information for one pair of phased haplotypes and summarises all remaining haplotypes using their (conditioned) site-frequency spectrum.

We simulate four scenarios: a sawtooth history, a discrete bottleneck, with a ten-fold change in population size ranging from $2N_{\mathrm{e}} = 7,000$ to $2N_{\mathrm{e}} = 70,000$, an increasing trend, and a decreasing trend (Figure 4.4 **a**-**d**). We find that Relate estimates population size with high accuracy in all four scenarios. The accuracy of MSMC depends on the number of haplotypes used, where its population size estimate is accurate up until $100,000$ years before present with 2 haplotypes and $10,000$ with 8 haplotypes. SMC++ has a comparable accuracy to MSMC, but detects trends in the more recent past as well.

A consequence of correctly estimated population sizes is a mutation rate that remains constant and close to $\mu = 1.25 \times 10^{-8}$ through time. The insets of Figure 4.4 show that indeed the mutation rate stays mostly constant and close to the truth.

We also simulate two diverging populations that separated 80,000 years ago or 10,000 years ago. In both cases, Relate estimates the split time accurately, and recovers the population size history of both populations (Figure 4.4**e**-**f**).

4.5 Comparison of data simulated with Relate-estimated demographic histories to 1000GP data

Relate, like other previous methods, estimates *effective* population sizes which may be influenced by unmodeled complexities not accounted for in the inference model [9]. To assess how well simulated data assuming panmixia and Relate-estimated effective population sizes emulate real data, we followed the analysis of Ref. [9] and compared three statistics capturing distinct aspects of variation patterns. Here, we used effective population sizes estimated for the 1000 Genomes Project populations CEU, CHB, and YRI (see Table A.1 for population labels, see Chapter 5 for details).

Our first measure is the expected heterozygosity, defined by

$$\pi = \frac{N}{N-1}\frac{\sum_{i=1}^{L} 2p_i(1-p_i)}{L}, \tag{4.1}$$

where p_i is the frequency of an allele, L is the total number of callable sites in the window (passing the Pilot mask for the 1000 Genomes Project data set), and N is the number of sampled haplotypes ($N = 20$). We calculated π for 20,000 randomly chosen 100kb windows using chromosome 1 of 10 individuals (20 haplotypes). We find in Figure 4.5**a** that panmictic simulations closely match the data, with better or equal accuracy to all methods included in Ref. [9]. In Figure 4.5**b**, we compare the site-frequency spectrum across all biallelic variants passing the Pilot mask on chromosome 1. We observe some discrepancies, particularly for rare variants, which may partially be explained by population structure and deep lineages that lead to more rare variants and are not accounted for in our panmictic simulations. In Figure 4.5**c**, we compare how LD patterns decay with physical distance from a focal SNP. While none of the methods closely matched the data in Ref. [9], we find that our demographic estimates appears to perform reasonably well at capturing the LD decay pattern in CEU. For CHB and YRI, we appear to match the data less well, although better than methods in Ref. [9], possibly because recombination maps are biased towards European hotspots.

4.6 Perturbations from infinite-sites, constant mutation rates, or perfect phase

We conduct additional simulations to assess the robustness of Relate to perturbations from the infinite-sites assumption, constant mutation rates, as well as perfect phasing of haplotypes, which are likely to be present in real data.

Our base-line simulation scenario, to which we will add these perturbations, is a simulation of 3000Mb for 1000 haplotypes with a population size history estimated by Relate for the GBR samples in the 1000 Genomes Project data set (see Chapter 5 for details). We use human chromosome 1 recombination rates, and assume a constant mutation rate of 1.25×10^{-8}. We subset 200 haplotypes used for inference and retain the remaining 800 haplotypes as a reference panel for rephasing (see below). This simulation assumes infinite-sites, such that every mutation occurs at a new base-pair position. We then add perturbations to this simulation scenario as follows.

First, we emulate the variable mutation rate observed for the triplet mutation TCC to TTC in Europeans (Figure 5.5**a**). We assume that with a probability of 1/96, a base-pair position mutates with a varying rate through time and otherwise, it mutates with a constant rate of 1.25×10^{-8}, independently of other mutations. To achieve this, we first classify a mutation in the base-line simulation as belonging to the variable mutation rate category with probability 1/96, independently of other mutations. We then add novel mutations occurring at rate $1/96 \times 1.25 \times 10^{-8}$ between 10,000 and 50,000 YBP. This is equivalent to assigning a mutation rate of 2.5×10^{-8} for this mutation category between 10,000 and 50,000 YBP, and a mutation rate of 1.25×10^{-8} otherwise. We find that age estimates of mutations, as well as population size estimates, remain accurate (Figure 4.5 **d**, **e**). We calculate the normalised mutation rate, where we first eliminate any remaining temporal trends in the average mutation rate by dividing by the average mutation rate in each epoch. For each mutation category, we then normalise the mutation rates such that the average rate over time equals 1. The elevation in mutation rate is detected with

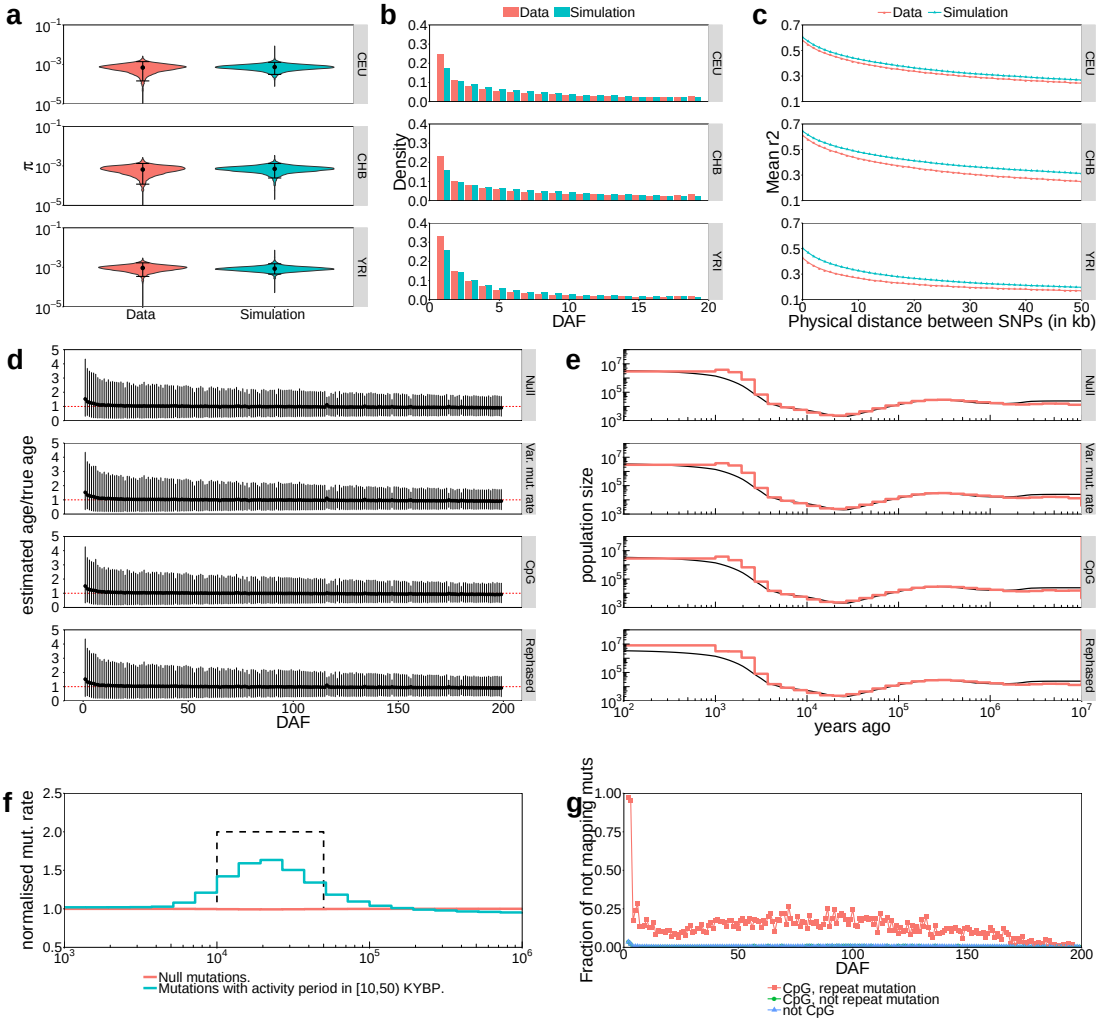


Figure 4.5: Accuracy under perturbations from infinite-sites, constant mutation rate, or perfect phase.
a, Expected heterozygosity (π) calculated for 20,000 randomly chosen 100kb windows. Circles show the mean and bars indicate the 2.5th and 97.5th percentiles. **b**, Derived allele frequencies, and **c**, LD decay patterns. For **a**, **b**, and **c**, we used ten 1000 Genomes Project individuals, and simulated 20 haplotypes using the demographic histories estimated by Relate. Each statistic is calculated using chromosome 1 (see Section 4.4 for details). **d**, Ratio of estimated and true age of a mutation, estimated as the mean of the lower and upper ends of the branch onto which the mutation maps, as a function of DAF. Circles show the mean ratio and bars indicate the 2.5th and 97.5th percentiles. Baseline simulation assumes infinite-sites and a constant mutation rate of 1.25×10^{-8}. We introduce perturbations, such as a variable mutation rate to a subset of sites, hypermutable base-pair positions emulating CpG dinucleotides, and inferred phase (see Sec. 4.6 for details). **e**, Accuracy of Relate-estimated population sizes on the same simulations as in (**d**). **f**, Normalised mutation rate for null mutations with a constant mutation rate of 1.25×10^{-8} and a mutation category with an activity period in $[10, 50)$ YBP during which the mutation rate doubled (dashed lines). **g**, Fraction of not mapping mutations as a function of DAF for the simulation with CpG-like mutations, categorised by whether the CpG-like site mutated once or more than once.

reasonable accuracy; however, we slightly underestimate the absolute elevation in mutation rate and the activity period appears longer than the truth (Figure 4.5 **f**).

Second, we consider a scenario in which 1% of all base-pair positions have a 20 times higher mutation rate, emulating CpG dinucleotides in the human genome [14]. We choose these CpG-like base-pair positions uniformly at random and remove any mutation in the base-line simulation that occurred at such a CpG-like position. At these CpG-like positions, multiple mutations may occur, where the mutation rate returns to its usual rate (1.25×10^{-8}) on any lineages below the first mutation. To account for the elevated average mutation rate caused by the introduction of CpG-like mutations, we specified a mutation rate of 1.49×10^{-8} in Relate. We find that age estimates of mutations, as well as population size estimates remain accurate (Figure 4.5**d**, **e**). At CpG-like sites that only mutated once, $> 99.5\%$ of mutations map to a unique branch, which is identical to the mapping rate in the base-line simulation (Figure 4.5**g**). However, at CpG-like sites with more than one mutation, we observe a substantially reduced mapping rate of 82.4%. In particular, of CpG doubletons at sites with more than one mutation, >97% did not map to a unique branch. We note that partial mapping of doubly mutated sites is expected. For example, in a case where two mutations have 1 and 20 descendants, respectively, it is possible to map the mutation to a branch with 20 descendants from the second mutation, if such a branch exists, because our mapping of mutations to trees allows for some noise/error.

Finally, we evaluate Relate on haplotypes that have been rephased using SHAPEIT2 [33]. We rephase 200 haplotypes (100 genotypes) using the remaining 800 haplotypes as a reference panel. Switch errors complicate the matching of rephased haplotypes to true haplotypes and make comparisons of estimated and true genealogies difficult. We therefore evaluate the accuracy of Relate using the accuracy of age estimates and population size histories. Both are almost unchanged in accuracy, with a slight elevation in recent population size estimates visible after rephasing (Figure 4.5 **d**, **e**). This elevation in estimated recent population sizes can,

at least in part, be explained by the random phase assigned to singletons, since singletons are on average moved from longer lineages to shorter lineages.

5

Reconstructing the genealogy of the 1000 Genomes Project data set

Contents

The 1000 Genomes Project was established to catalogue human genetic diversity across ethnic groups globally. In its final Phase 3 release, it records whole-genome sequencing data of 2504 individuals from 26 populations (see Table A.1) [1]. We apply Relate to this data set and obtain a joint genealogy for African, American, East Asian, European, and South Asian individuals.

After assessing the quality of the constructed genealogy, revealing for instance that more error-prone mutations map less frequently to a unique branch than other mutations (Section 5.4), we estimate historical population sizes for all 26 populations and study their separation histories (Section 5.5). We then discuss a remarkable increase in the TTC to TCC mutation rate in Europeans between 5,000 and 30,000 YBP (Section 5.6) and demonstrate that we find evidence for introgression with

	CPU days	Max. memory usage (Mb)
Chromosome painting	6.9	2,802
Tree building	132.2	14,632
Finding Equivalent branches	1.0	7,435
Estimating branch lengths	482.2	83
Other	0.2	6310
Total	622.4	

Table 5.1: CPU time spent in days (second column) and maximum memory usage in Mb (third column) in each stage of the algorithm applied to 4956 haplotypes of the 1000 Genomes Project dataset.

Neanderthals and Denisovan-like populations (Section 5.7).

These analyses reveal multiple instances of evolutionary processes that are themselves evolving through time: "evolution of evolution". We simultaneously infer multiple population specific signals of adaptation using the same set of genealogical trees, which should enhance interpretability compared to separate analyses conducted with conventional methods. We believe that the analyses presented here can be further extended to infer more complex events, including directional migration or ancient admixture, and motivate other applications, including using genealogical trees for quality control (QC), or tree-based phasing and imputations (see Discussion).

5.1 Pre-processing the data

We obtained a phased version of the data set from Ref. [2]. We pre-process the data as follows. First, we exclude one individual (two haplotypes) from each population for future applications. The total number of individuals included in the analysis is 2478. A breakdown into populations is shown in Table A.1. Second, we exclude all SNPs that are not marked as biallelic. Third, we use a genomic mask provided with the 1000 Genomes Project dataset to identify regions in the genome with low certainty of genotypes [1, 48], excluding any base marked as "not

passing" in the pilot mask. We also exclude any bases for which the fraction of "not passing" bases within 1000 bases to either side exceeds 0.9. This procedure excludes some SNPs and also readjusts the number of bases between SNPs at which we could have potentially observed a SNP. Finally, we use an estimate of the human ancestral genome to identify the ancestral allele of each SNP [77]. In total, we include 73,146,033 SNPs in the analysis.

5.2 Runtime

Relate terminated after less than 2 CPU years, or 4 days when run on a high-performance cluster, using up to 300 cores in parallel (see Table 5.1). Each core had a maximum memory allowance of 16GB and was equipped with an Intel Ivybridge 2.4 GHz or Intel Haswell 2.6 GHz processor.

5.3 Number of trees built

As a first indication of good accuracy of the inferred genealogy, the number of trees constructed in a genomic subregion comprising $10,000$ SNPs is correlated to recombination distance ($r^2 = 0.63$, Fig. 5.1**a**). In regions with close to zero rates of recombination, we estimate that we construct approximately one tree every 125 SNPs. The number of trees we build is slightly inflated by the fact that in our implementation of Relate, we rebuild a tree after 200 to 1000 SNPs for computational reasons. Consequently, not every new tree represents a recombination event. We construct 1,414,626 trees for the whole genome amounting to a new tree every 2,121 bases. The average tree has 3,883 SNPs mapped to it, reflecting block-like structures of human haplotypes between recombination hotspots (Fig. 5.1**b**). Any

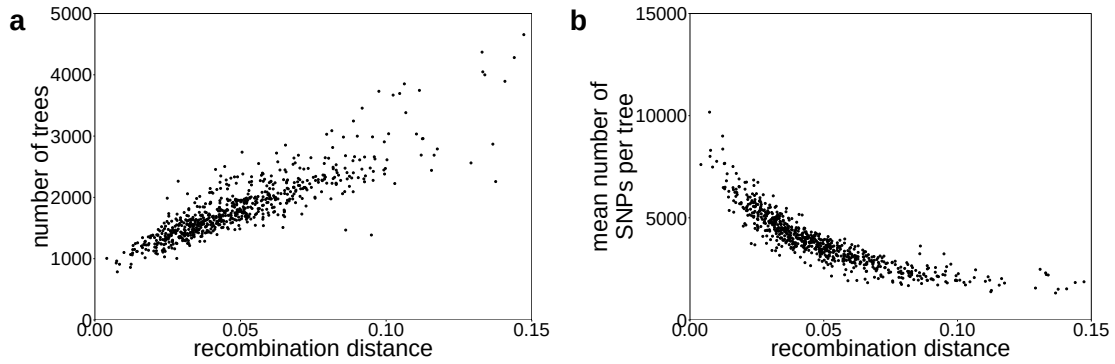


Figure 5.1: Number of trees built for the 1000 Genomes Project data set. **a**, Number of trees built versus the recombination distance for all 22 chromosomes. **b**, Mean number of SNPs that map to a unique branch versus the recombination distance in that bin. Every point represents a subregion of 10^5 SNPs. For each region, we calculate the total recombination distance (in units of centi-Morgans) from the first to the last SNP, number of trees built, and mean number of SNPs per tree in that region.

mutation mapped to one branch of its respective tree is propagated to branches of on average 892 neighbouring trees.

5.4 CpG mutations map less frequently than other mutations

Another way of assessing the accuracy of inferred genealogies is to look at the rate at which SNPs map to unique branches of marginal trees. Singleton mutations always map to a unique branch and we therefore exclude singletons from the following analysis. For any non-singleton, 14.3 % of SNPs do not map to a unique branch (Fig. 5.1). In Figure 5.2**a**, we show how the rate of mapping a mutation depends on the derived allele frequency of a mutation. We can see that the fraction of non-mapping SNPs is relatively high for rare SNPs and rapidly decreases to around 5% for more frequent SNPs. If we exclude any SNP with a derived allele frequency of less or equal to 10, only 4% of SNPs do not map to a unique branch (Fig. 5.2 **a**). A possible explanation is that common SNPs can be mapped to a branch even with some errors, whereas the criteria for mapping rarer mutations is more stringent (Fig. 2.3**b**).

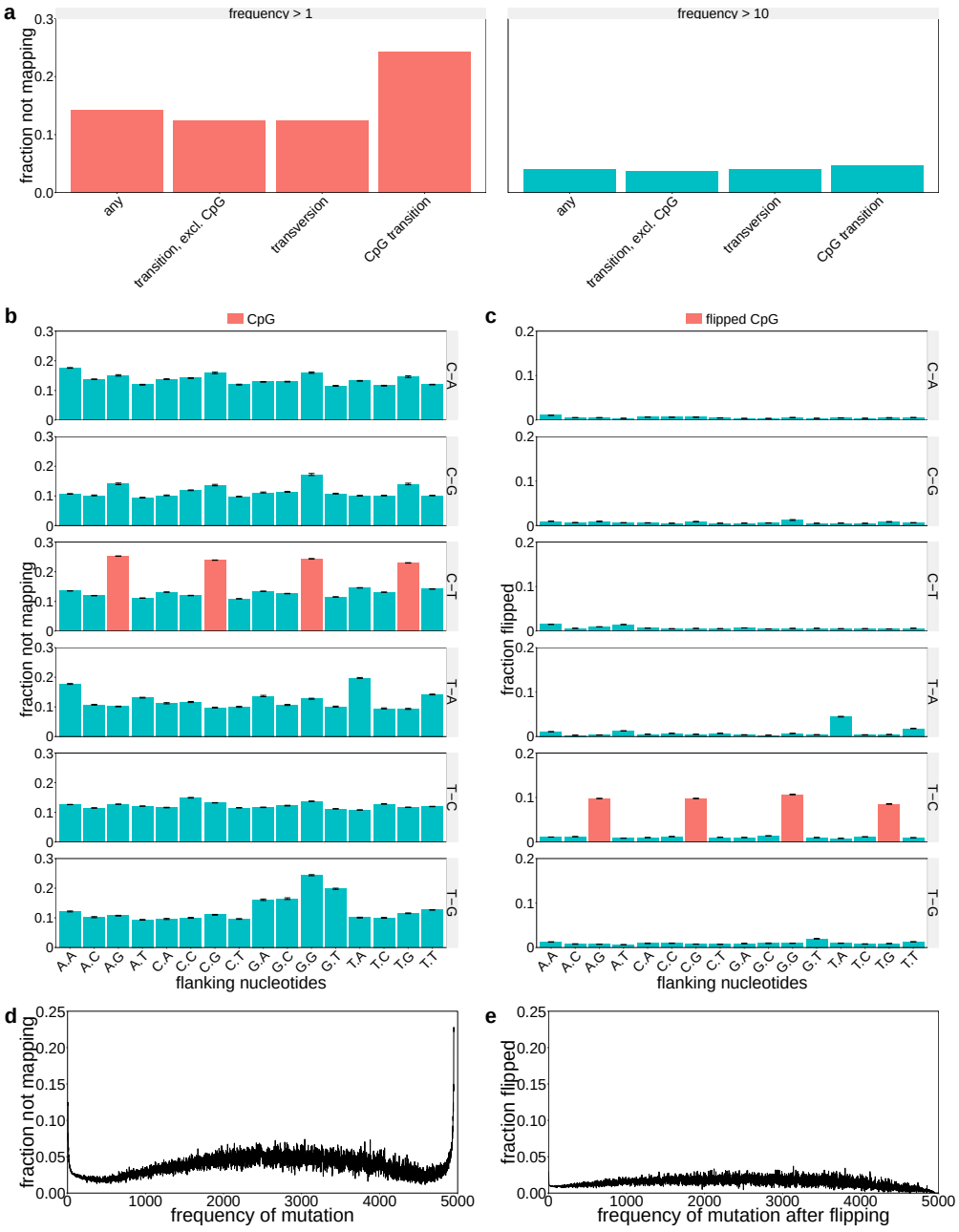


Figure 5.2: Fraction of SNPs that could not be mapped in the 1000 Genomes Project data set.
a Fraction of SNPs that could not be mapped to a unique branch for SNPs excluding singletons (left) and SNPs with derived allele frequencies larger than 10 (right). **b** Fraction of SNPs that could not be mapped to a unique branch for all 96 possible triplet mutations, excluding singletons **c** Fraction of SNPs that were flipped for all 96 possible triplet mutations, excluding singletons. In **b** and **c**, CpG transitions are indicated in red. The 95% confidence intervals are indicated by black brackets. **d** Fraction of non-mapping SNPs by derived allele frequency of the mutation in the sample. For each frequency, we divide the number of non-mapping mutations of that frequency by the number of mutations of that frequency. **e** Fraction of flipped SNPs by derived allele frequency of the mutation after flipping. For each frequency, we divide the number of flipped SNPs of that frequency (after flipping) by the number of SNPs of that frequency.

Next, we investigate whether SNPs that could not be mapped to a unique branch are enriched with certain mutation types that are known to be more error prone or have a higher mutation rate. For instance, CpG sites are known to have a substantially higher rate for a mutation $CG \rightarrow TG$ [153]. We should therefore expect a higher probability of observing two or more mutations at the same genomic position. Such SNPs usually cannot be mapped to a unique branch. Indeed, we observe that the fraction of not mapping CpG transitions is 24.3 % which is significantly higher than the overall average (Figure 5.2**a**).

To further study the effect of adjacent nucleotides, we group mutations by their two neighbouring nucleotides in sequence. This yields 96 categories after accounting for equivalent mutations due to symmetry on the complementary strands. We observe that the fraction of non-mapping SNPs is highly variable with respect to the mutation category (Fig. 5.2**b**), with some categories exhibiting resembling the CpG signal. We observe a clear elevation in the fraction of non-mapping mutation for GT $\rightarrow$ GG mutations, which have recently been suggested to be prone to sequencing artefacts in a subset of 1000 Genomes Project samples [4].

Mutations that occur at the same genomic position both in humans and other primates can cause confusion of the ancestral and derived alleles which we can detect as flipped SNPs. A higher mutation rate can be one cause for incorrect ancestral/derived alleles and we again find a clear signal for CpG transitions $CG \rightarrow TG$. The fraction of flipped SNPs appears to be less dependent on the mutation category (Fig. 5.2**c**) and is around 1% on average. The fraction of flipped SNPs is moderately dependent on the frequency of the alternative allele in the sample (Fig. 5.2(f)).

Currently, we cannot explain spikes in other categories and future studies correlating these rates of mapping mutations with QC measures could reveal other error-prone categories. We note that some categories, including ATA $\rightarrow$ AAA, or TTA $\rightarrow$ TAA have an elevated rate of not mapping to a unique branch as well as being flipped, suggesting that these underlie some biological cause. In particular, the TTA $\rightarrow$ TAA signal may be driven by the hexamer TTAAA which has previously

been reported to have an elevated mutation rate due to being common targets for transposons that induce T to A mutations [21].

5.5 Historical population sizes

We estimate historical population sizes for all 26 populations in the 1000 Genomes Project data set (Fig. 5.3). We extract the embedded genealogy corresponding to each population within the full genealogy of 2478 samples and jointly fit branch lengths and population size histories using the method described in Chapter 3.

Multiple lines of evidence suggest that *homo sapiens* evolved around 150,000 - 200,000 YBP in parts of the African continent [46]. Consistent with this estimate, population sizes of all 26 populations converge in the ancient past, around 200,000 YBP. All non-African groups show a severe bottleneck following their out-of-Adrica migration. We find a second bottleneck in FIN around 3,000 to 9,000 YBP after separation from GBR. This bottleneck has been reported previously and is thought to have caused enrichment of certain disease-causing gene variants, commonly classified as Finnish heritage diseases [27, 91]. Indications of a second bottleneck can also be observed in Indian Telugu in the UK (ITU) and a severe second bottleneck can be observed in the Peruvian population (PEL). All populations achieve a remarkable increase in population size in the recent past, often to >1,000,000, however we note possible inaccuracies due to incomplete power in the 1000 Genomes Project to detect rare variant [1] leading to underestimation, and computational phasing leading to overestimation (Figure 4.5**e**).

In Fig. 5.4, we estimate cross-coalescence rates for pairs of populations. In addition to population-wide averages of coalescence rates, we also show matrices containing the coalescence rate of pairs of haplotypes at specific times. This demonstrates that we can directly observe population structure, where some cases indicate finer-scale structures not captured by the labels assigned to samples.

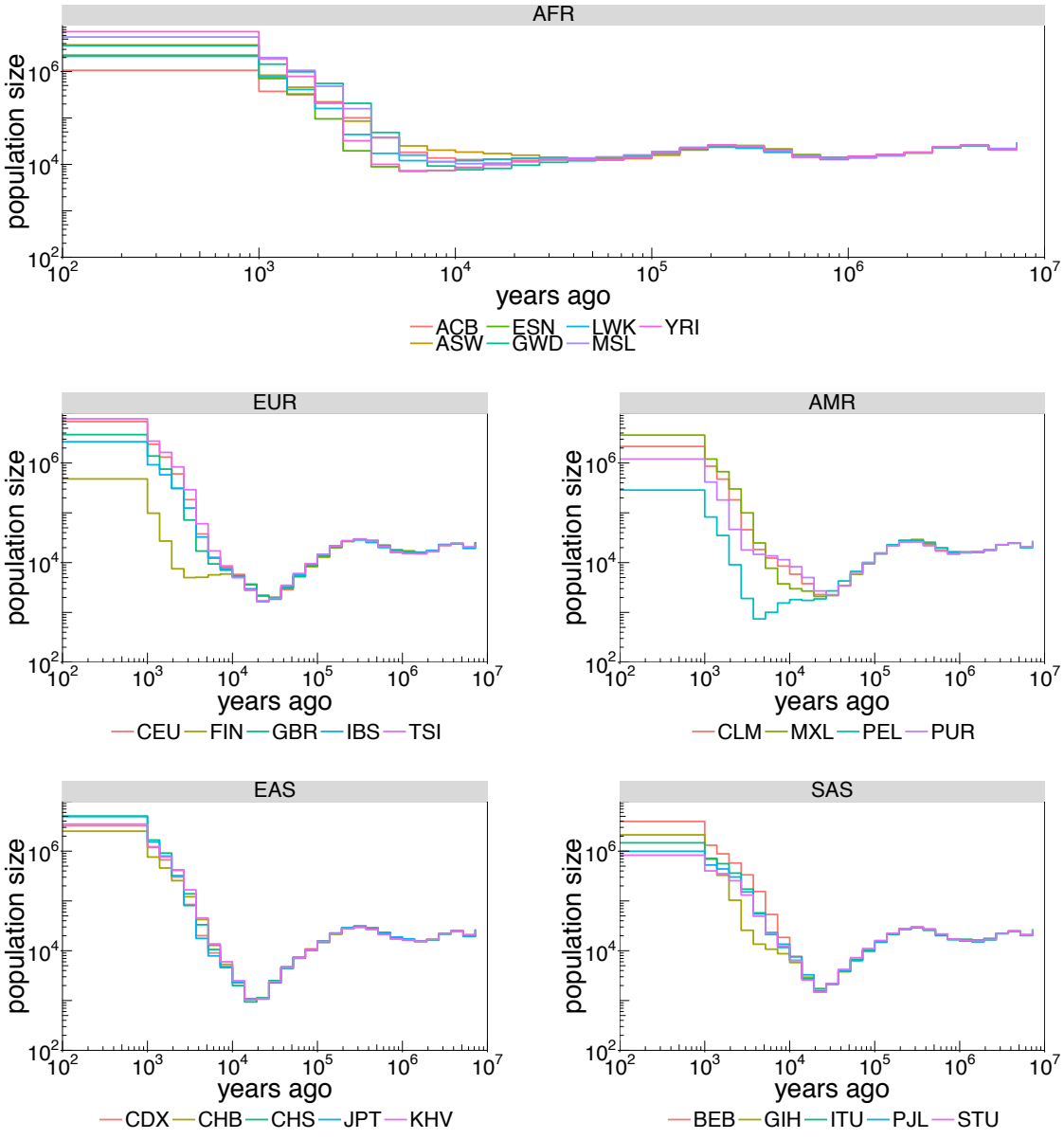


Figure 5.3: Historical population sizes of all 26 populations of the 1000 Genomes Project data set.
Historical population sizes of all 26 populations of the 1000 Genomes Project data set. For each population, we first extracted the genealogy corresponding to that population. We then jointly fitted branch lengths and population size history for this genealogy.

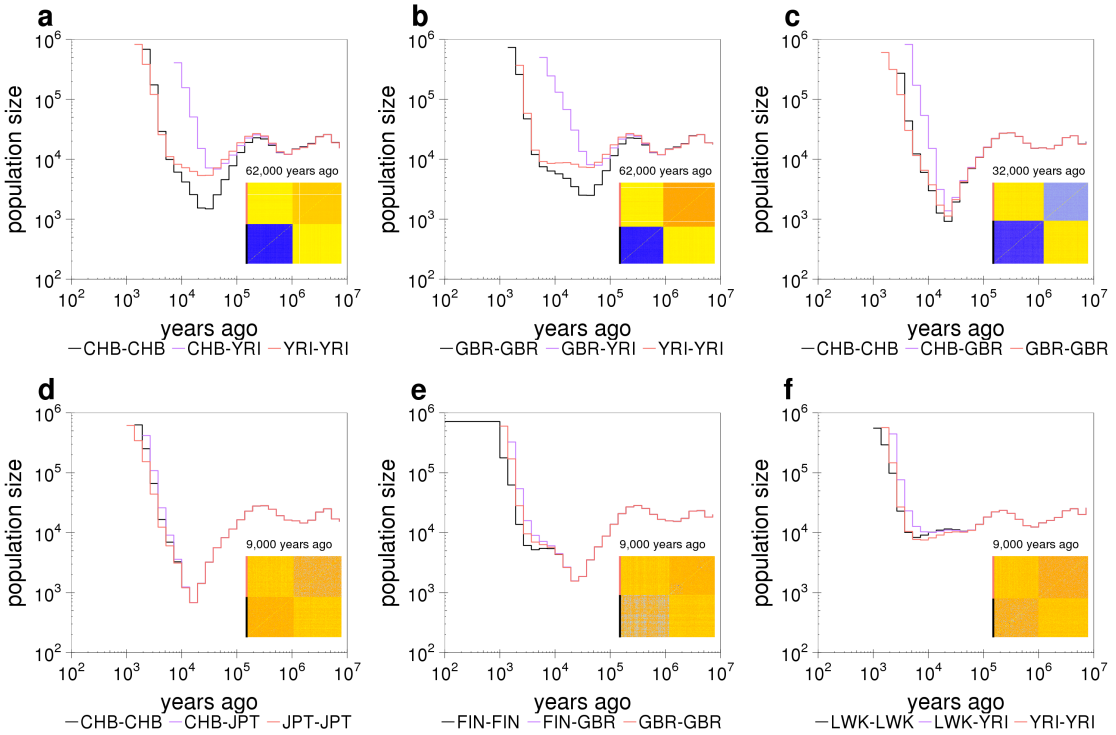


Figure 5.4: Cross-coalescence rates for pairs of populations of the 1000 Genomes Project data set.
Population size and cross-coalescence rate estimates using the genome-wide genealogy for **a**, CHB and YRI, **b**, GBR and YRI, **c**, CHB and GBR, **d**, CHB and JPT, **e**, FIN and GBR, and **f**, LWK and YRI. Insets show the matrices of coalescent rates between pairs of haplotypes at the indicated time. Rows and columns are sorted by population labels of haplotypes, as indicated by the colour on the left of each matrix.

A split of Asian and European populations (CHB: Chinese in Beijing and GBR: British in England and Scotland shown) from African populations (YRI: Yoruba in Ibadan, Nigeria shown) is already visible at 200,000 years before present (YBP). This split appears to be gradual, with cross-coalescence rates remaining relatively high until around 60,000 YBP (Figure 5.4 **a,b**). This is consistent with previous estimates, where a major out-of-Africa event has been dated at around 60,000 YBP, and initial migrations may date back to more than 150,000 YBP [6, 88]. Following this, Asian (CHB shown) and European (GBR shown) populations begin to separate, with a clear separation visible at around 30,000YBP (Figure 5.4 **c**). We also detect more recent separations, such as between CHB-JPT (JPT: Japanese in Tokyo) or FIN-GBR (FIN: Finnish in Finland) (Figure 5.4 **d,e**). We also observe separations of populations that remained on the African continent (LWK:

Luhya in Webuye, Kenya and YRI shown) following the departure of European and Asian populations (Figure 5.4 **f**).

5.6 Rapid evolution of mutation rates

We estimate the mutation rate through time for all 96 triplet mutations. To estimate mutation rates for a mutation category of interest, we calculate, for each epoch, the quotient of the number of mutations in that category by the total branch length over bases at which such a mutation may have occurred. When estimating branch lengths, we fix the average mutation rate to a constant value through time, such that any change in average mutation rate should in theory be absorbed in our population size estimate. We therefore first eliminate any remaining temporal trends in the average mutation rate by dividing by the average mutation rate in each epoch. For each population, we then normalise the mutation rates such that the average rate over time equals 1. In simulations we confirmed that this procedure can detect and approximately date variable mutation rates in subsets of mutations (see Fig. 4.5**f**).

The strongest signal is observed for the TCC to TTC category (Fig. 5.5**a**), with similar but weaker signatures for ACC to ATC and TCT to TTT. This remarkable increase in the mutation rate in these three categories is in agreement with previous studies [65, 66] and primarily observed in European groups around 5,000 - 30,000 YBP, but weak or absent in the present day. This elevation in mutation rate is absent in populations from East Asia and Africa.

Other mutation types show more subtle temporal biases and signatures consistent with GC-biased gene conversion, a process known to affect large parts of the human genome [50, 121] (Fig. 5.5**b**). GC-biased gene conversion biases the conversion rate of heterozygotes during repair of double-strand breaks towards C or G nucleotides [34]. This effect leads to a faster-than-expected spread of mutations towards C and G. A consequence is that mutations towards C or G are overrepresented, whereas

mutations from C and G are underrepresented in the past. Consistent with this idea, we observe that mutation rates decrease towards the present in T to C and T to G mutations, whereas we observe a small increasing trend in C to A and C to T mutations.

In African populations, we observe a recent increased mutation rate of GCA to GGA and GCT to GGT. We note that the apparent decrease in mutation rate for categories involving a CpG dinucleotide is likely an artefact, caused by many rare mutations at these sites being not mappable to the tree due to repeat mutations.

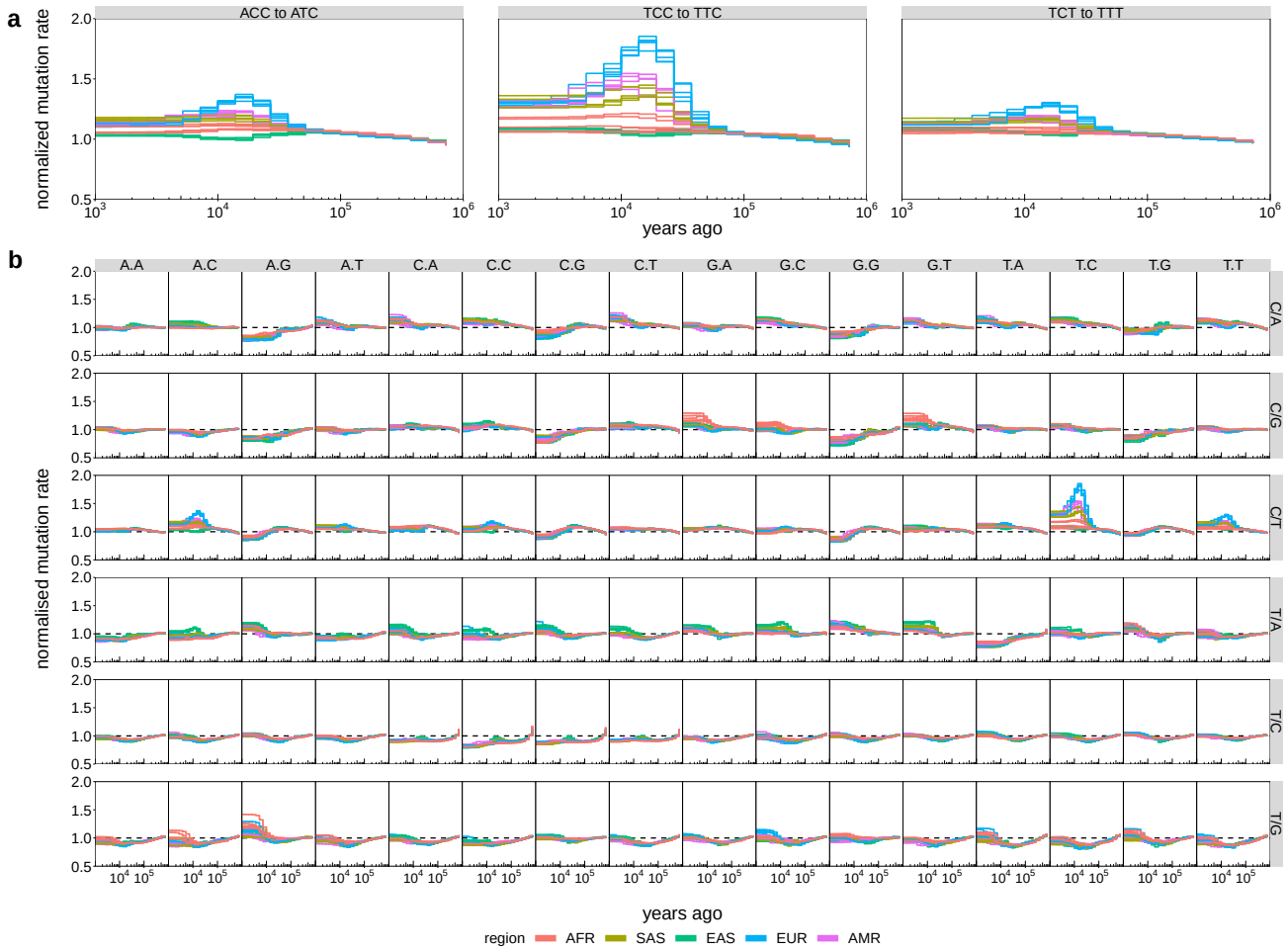


Figure 5.5: Mutation rate trends for 96 triplet mutations.
a, Evolution of mutation rates for three triplet mutations ACC to ATC, TCC to TTC, and TCT to TTT. **b**, Evolution of mutation rates of triplet mutations for all 96 possible categories. See Section 5.6 for how mutation rates are normalised.

5.7 Evidence for introgression with archaic humans

Neanderthals and Denisovans are an extinct subspecies of humans that diverged from the lineage that would later evolve into (among others) *homo sapiens* around 800,000 YBP and subsequently migrated out of Africa [73, 127]. Neanderthals were wide-spread in Eurasia, while Denisovans are believed to have primarily inhabited a region ranging from Siberia to South Asia and Oceania. Following the out-of-Africa migration of modern humans around 100,000 YBP, modern humans introgressed with Neanderthals and Denisovans who subsequently went extinct [51, 101, 129]. Today, it is established that all non-African human groups possess similar levels of Neanderthal introgression, and specific Asian and Australasian groups possess admixture from a group related to Denisovans. Such or similar events with distantly related groups introgressing into the modern human lineage are expected to appear in form of branches that remain separated from other human lineages for long times. The tree depicted in Figure 5.6**a** is consistent with an out-of-Africa migration of Neanderthals around 800,000 YBP and subsequent introgression into GBR as suggested previously.

To find out whether long branches in non-African genealogies are enriched with Neanderthal and Denisovan mutations and to potentially identify separate, unknown events, we first identify deep branches with an upper coalescence age >1 million years (MY) in age, binning them by the age of the lower coalesence event. Because the same branch may persist over multiple trees, we identify equivalent branches (Section 2.4.1) and average ages of lower and upper ends across these equivalent branches. We annotate branches possessing at least two derived mutations by whether at least one of the mutations is shared with Denisova or Neanderthal genomes, leveraging complete genome sequences of the Altai [124] and Vindija Neanderthals (NEA) [51], and a Denisovan (DEN) [101]. We then calculate the fraction of deep branches attributable to Neanderthal or Denisovan introgression as

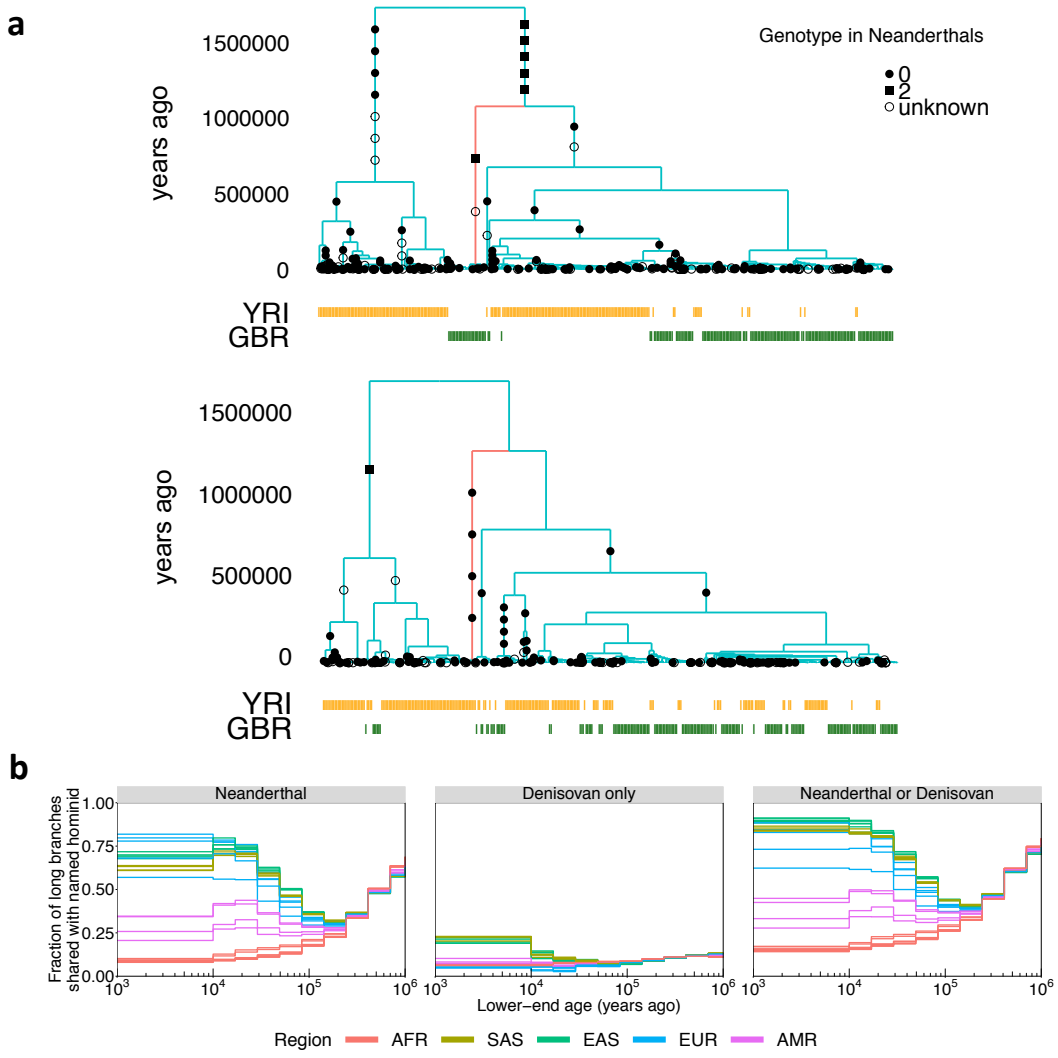


Figure 5.6: Evidence for introgression with Neanderthals and Denisovans. **a**, Marginal trees for a subregion on chromosome 14 (top) and chromosome 11 (bottom). The top tree contains a long branch with descendants only in GBR (red) consistent with Neanderthal introgression into GBR. The bottom tree contains a long branch with descendants only in YRI (red) consistent with introgression in YRI involving a hominid not closely related to Neanderthals. **b**, Fraction of branches with an upper-end age >1M YBP that are shared with Neanderthals (left), Denisovans and not Neanderthals (center), or Neanderthals or Denisovans (right). Colours encode geographic regions (AFR: Africa, EAS: East Asia, EUR: Europe, SAS: South Asia, AMR: Americas).

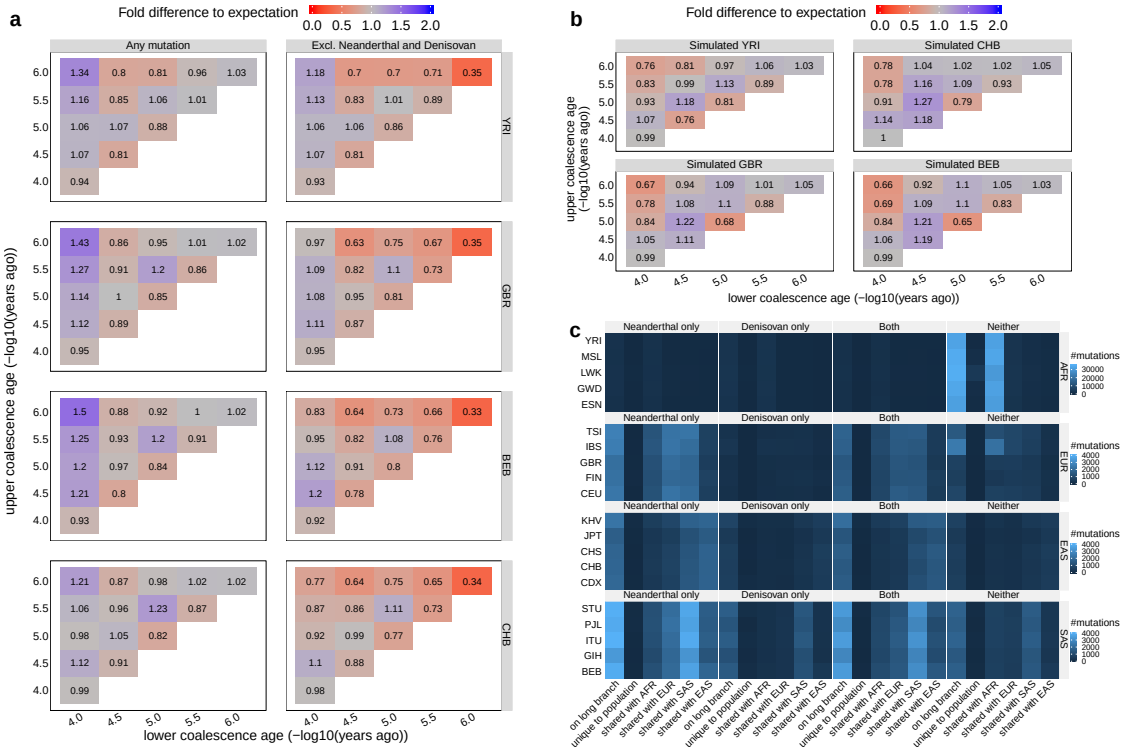


Figure 5.7: Evidence for introgression in African populations.
a, Number of mutations binned by age of upper and lower coalescence event, relative to the expected number of mutations when randomising topology while fixing ages of coalescence events. Right column shows mutations not present in Neanderthal or Denisovan samples. **b**, Same as **a** for four simulated data sets. We simulated 3×10^9 bases with Relate-estimated population size histories of YRI, CHB, GBR, and BEB. **c**, Number of mutations on branches with an upper end older than 1M YBP and lower end younger than 30,000 YBP, categorised by whether the mutation is additionally found only in Neanderthals, only in Denisovans, both, or neither. For each category, we also distinguish whether the mutation is unique to the population of interest or shared with other populations in AFR, EUR, SAS, or EAS.

a function of the lower age of the branches (Fig. 5.6**c**). We assign a branch to a population if at least one descendant of that branch is in the population.

Let us first understand how we expect the fraction of such branches to change with lower end age in a scenario where Neanderthals and Denisovans split from the human lineage around 800,000 YBP and do not introgress back subsequently. In this scenario, any branch with an upper end age extending to before the Neanderthal divergence time has a fixed probability of carrying mutations shared between Neanderthals and modern humans. This is because the Neanderthal lineage has equal probability of coalescing into any lineage ancestral to modern humans and

will carry any mutation that happened on that branch prior to the Neanderthal divergence. In particular, this probability does not depend on the lower-end age, as long as the lower end age is younger than the divergence time and therefore we expect that the fraction of branches shared with Neanderthals is constant and independent of the lower time age (for a sufficiently young lower-end age, e.g., <200,000 YBP).

In an alternative scenario, in which Neanderthals and Denisovans recently introgressed back into the modern human lineage, we expect the fraction of shared branches to increase. We note that any lineages from recent introgression events will show a lower-end age younger than the time of introgression, and upper-end older than the split time of the introgressing group, so we expect branches with a younger lower-end to be most enriched with lineages that came from distantly diverged introgressing groups. For branches originating within the last 10,000YBP, 85-90% are shared with Neanderthal or Denisovan for most Eurasian groups (Figure 5.6**c**). We therefore hypothesise that aside from groups closely related to Neanderthals and Denisovans, no strongly diverged hominid has left a major, recent impact in non-African populations studied here. An exception is IBS, which has more long branches shared with African populations (Fig. 5.7**c**), as well as recently admixed populations in the Americas. In East and South Asian groups, the data suggest a very recent arrival of Denisovan DNA (mainly <15,000YBP), which would be consistent with introgression of Denisovan DNA into a third modern human group possibly in South East Asia or Oceania [129] and subsequent admixture with East and South Asian groups. In non-Africans, Neanderthal sharing remains high for branches with lower-end age younger than ∼30,000YBP. These dates are only lower bounds on the introgression time, and an accurate arrival date of Neanderthal DNA would require estimating a joint genealogy which requires further work. Nevertheless, they are consistent with previous estimates based on LD [135], and of direct evidence of hybrids [44, 135] around 40,000 YBP. Moreover, elevation in the sharing of quite deep haplotypes with Neanderthals steadily increases for branches with lower-end age of 100,000 YBP towards the present, which is suggestive of introgression beginning

from this time in non-African individuals, although it is important to note that our date estimates for individual events might be over- or under-estimates in some cases.

In contrast to non-African groups, sharing with Neanderthal/Denisovans is lower (<20%, Fig. 5.6**c**) in African populations, and declines towards the present, suggesting minimal recent interactions [51, 101]. Instead, the decline in branches shared with Neanderthals and Denisovans towards the present is indicative of an excess in long branches with different origin. In fact, the largest number of long branches observed come from African populations (on average, on deep branches with lower coalescence age <30,000 YBP; 42,434 vs. 7,012 mutations occur in African vs. non-African populations). Of mutations on long branches, 98% are unique to populations in Africa, indicative of separate events occurring in non-African and African populations (Fig. 5.7**c**).

To test whether the excess of long branches and mutations on top of them in African groups (as well as non-African groups) can be explained by branches extending to >1MY by chance, we calculate the expected number of mutations for bins divided by upper and lower end times assuming panmixia. The expected number of mutations by lower and upper end depends on the demographic history. To match the demographic history, and potential other factors influencing coalescence times including potential biases in the tree building method, we fix the ages of coalescence events in each tree, but randomise the topology assuming a single panmictic population. To derive a formula for the expected number of mutations in each bin, we first note that the probability that a branch with lower end reducing the number of lineages from $\ell+1$ to ℓ has an upper end that reduces the number of lineages from $h+1$ to h is given by

$$\frac{\binom{\ell-1}{2}}{\binom{\ell}{2}}\frac{\binom{\ell-2}{2}}{\binom{\ell-1}{2}}\cdots\frac{h}{\binom{h+1}{2}}=\frac{2h}{\ell(\ell-1)}. \tag{5.1}$$

The probability of upper and lower coalescence ages falling into time intervals (bins) s and r for any particular branch that exists while k lineages remain in the tree is given by the sum of Eq. (5.1) over all $\ell > k$, $h \leq k$, such that the time of coalescences t_ℓ and t_h fall into s and r, respectively. The probability that a

mutation arising while k lineages remain has upper and lower end ages in s and r requires multiplication of a factor of $1/k$ and we obtain

$$P(r, s|k) = \sum_{\ell \geq k, h < k} \mathbb{1}_{t_\ell \in s} \mathbb{1}_{t_h \in r} \frac{2h}{\ell(\ell-1)} \frac{1}{k}, \tag{5.2}$$

where $\mathbb{1}$ denotes the indicator function. While the number of lineages k remaining in the observed tree when the mutation arose is unknown, a mutation is equally likely to have arisen anywhere on the branch it maps to assuming neutrality. Therefore, we calculate the weighted average $\sum_{k=2}^{N} w_k P(r, s|k)$, with weights w_k defined as the proportion of the observed branch while k lineages remain. Summing this over all SNPs yields the expected number of mutations with upper and lower coalescence age falling, respectively, into bins s and r. In Figs. 5.7**a** and **b**, the $\log_{10}$ transformed age bins are defined by $[-\infty, 4.25)$, $[4.25, 4.75)$, $[4.75, 5.25)$, $[5.25, 5.75)$, $[5.75, \infty)$.

We observe a strong enrichment of mutations on deep branches with upper end >1MYBP and lower end <40,000 YBP in YRI, GBR, BEB, and CHB, which can be almost entirely explained by Neanderthals and Denisovans in the non-African populations, but not in YRI (Fig. 5.7**a**). In panmictic simulations with matched population size histories, we observe no such enrichment (Fig. 5.7**b**). This may be consistent with ancient but uncharacterised population structure within Africa, for which there is increasing evidence [63, 119, 125]. Fig. 5.6**b** shows one example consistent with an introgression event in YRI with a hominid not closely related to Neanderthals.

6

Detecting evidence for natural selection

Contents

Some instances of genetic adaptation to changing environments, diet, and lifestyles have been well documented, driven by selective pressures on mutations that influence favourable traits. One of the strongest signals of recent positive selection in the genome is a mutation that has led to Lactose tolerance in some Europeans [13], while similarly strong signals have been observed for changes in skin and hair pigmentation (e.g., Ref. [112]) and protection from infectious diseases such as malaria (e.g., Ref. [72]). Interestingly however, there are only a handful of instances showing such strong evidence of positive selection on single loci. While

increased sample sizes and more powerful statistical approaches will reveal further cases where phenotypes are drastically changed by one mutation, the understanding that most traits are governed by a large number of mutations, each typically with a small effect, predicts that selection more commonly should be dispersed over many loci (see e.g., Ref. [18, 138, 140]). Understanding how selection acts on such polygenic traits has been a topic of much discussion, and some studies have worked to catalogue traits that may have been selected (see e.g., Refs. [12, 169]).

In this chapter, we propose a tree-based statistic for identifying positive selection. Positive selection can appear in the form of favourable mutations spreading rapidly in a population. This should appear as a burst of coalescence events following the emergence of a beneficial mutation. Our statistic will measure, for each mutation, to what extent lineages carrying this mutation have out-competed other lineages. We apply this statistic to variants recorded in the 1000 Genomes Project data set and find known, as well as previously unreported loci under strong positive selection. Using SNP-trait associations documented in genome-wide association studies (GWAS), we then investigate selection acting on polygenic traits, revealing a complex picture of directional adaptation.

6.1 Existing methods

Numerous different approaches exist, often levering summary statistics of the data that capture some intuition about expected diversity patterns or population differentation under different modes of selection [159].

One of the most commonly used approaches is the F_{st} statistic [71], which measures population differentiation of a single locus by comparing the variance of allele frequencies within and across populations. This statistic requires classification of samples into distinct populations and can be used to detect genomic regions

that are highly differentiated across these populations, indicative of directional selection in one population but not the other.

Other methods quantify selection evidence using changes in diversity patterns in genomic regions around a selected focal SNP. Tajima's D [154] is one of the most influential methods building on this concept and compares two measures of genetic diversity: the average number of pairwise differences observed in a genomic region, given by

$$\Pi_N = \frac{\sum_{i<j} d_{ij}}{N(N-1)/2}, \tag{6.1}$$

where d_{ij} denotes the number of differences between haplotypes i and j, and Watterson's estimator of the population-scaled mutation rate, defined as

$$\hat{\theta}_W = \frac{S}{\sum_{k=1}^{N-1} 1/k}, \tag{6.2}$$

where S denotes the number of segregating sites. Using these quantities, Tajima's D is defined as

$$D = \frac{\Pi_N - \hat{\theta}_W}{sd\left(\Pi_N - \hat{\theta}_W\right)}, \tag{6.3}$$

where *sd* denotes the standard deviation. Under neutrality and assuming a constant population size, Π_N and $\hat{\theta}_W$ both equal the population-scaled mutation rate in expectation, such that $E[D] = 0$. The average pairwise heterozygosity Π_N depends on allele frequencies, with an excess of rare minor alleles implying a smaller inferred mutation rate, while Watterson's estimator is unaffected by allele frequencies. Positive selection is therefore expected to result in a negative D. It should be noted that recent population expansions and other mechanisms diverting from idealistic random mating scenarios can also result in a negative, non-zero D statistic and disentangling these different scenarios is not always easy. Many variants of Tajima's D exist, based on the concept of comparing different measures of genetic diversity to detect selection evidence. For instance, Fay and Wu's H statistic [38] is defined as the difference of Π_N (Eq. (6.1)) and

$$\hat{\theta}_{\mathrm{H}} = \frac{2N}{N-1} \sum_{k=1}^{S} p_k^2, \tag{6.4}$$

where p_k denotes the derived allele frequency of SNP k. As for Tajima's D, the expectation of H is zero under neutrality, and a negative H is indicative of an excess in common variants compared to intermediate variants which can be suggestive of positive selection. Common to all such methods is the necessity of analysing genomic regions rather than individual SNPs, leading to poor resolution and poor power on weaker selective sweeps with small or intermediate present-day frequency.

A third group of methods leverage LD patterns to identify SNPs undergoing positive selection. These methods, including the widely used iHS method [133, 160], rely on the intuition that haplotypes that have swept rapidly to high frequencies have not had the time to be broken down by recombination, and hence should extend for longer than expected.

Lastly, in some cases where ancient genomes are available, allele frequency changes have been observed directly and tested for non-neutral evolution [95]. Using ancient DNA samples, frequency trajectories and selection coefficients can be estimated by fitting an HMM to the observed frequencies, with hidden states given by "true" population-wide allele frequency trajectories and transition probabilities of frequencies between time points depending on the selection coefficient [96, 97].

Using genealogical trees, we can directly track frequency changes through time. This opens up the possibility of calculating a range of statistics tailored to different selection scenarios. A method relying partially on information captured by genealogies is the singleton density score (SDS), which investigates positive selection based on the density of singletons around a focal SNP [40]. This method indirectly measures changes in tip-branch lengths of the genealogy, capturing the signal that tip branches inheriting a strongly advantageous allele are expected to be shorter compared to remaining tip branches, making it suitable for detecting very recent adaptation on standing variants. One can adapt the concept of SDS by directly comparing tip branch lengths of carriers and non-carriers of a mutation [35]. To define this tree-based SDS statistic (trSDS), we first define the raw trSDS (rtrSDS) by

$$\text{rtrSDS} = t_{\text{non-carriers}} - t_{\text{carriers}}, \tag{6.5}$$

where $t_{\text{non-carriers}}$ denotes the mean tip branch length of non-carriers, and t_{carriers} denotes the mean tip branch length of carriers. If a mutation has experienced recent positive selection, we expect rtrSDS > 0. We define trSDS by standardising rtrSDS using the rtrSDS distribution for a given present-day DAF under the null.

A number of methods fully leverage genealogical trees. For instance, Ref. [30] employs an importance sampling scheme that computes the likelihood of a selection coefficient by integrating over possible allele frequency trajectories and coalescent histories in genomic windows without recombination. More recently, a new method CLUES adapted this idea by integrating over possible allele frequency trajectories using an HMM, and sampling marginal trees using methods such as Relate [149]. Another statistic, DRC_T, quantifies bursts in coalescence events as signals of positive selection, and is suited particularly well to very recent selection potentially acting on standing variants [117].

6.2 A tree-based statistic for detecting selection

We propose a new statistic that examines whether the observed frequency change is unusual under the standard coalescent model. For this test, we utilise a classical result in coalescent theory describing the shape of the coalescent without recombination. We let k be the number of lineages remaining in the tree and define $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)$ as the vector containing the number of present-day offspring of each of the k lineages. Then, the probability distribution of $\boldsymbol{Z}$ is the uniform distribution on the possible partitions of N lineages to k ancestors (see e.g., Ref. [54]), i.e. it is given by

$$P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k) = \frac{1}{\binom{N-1}{k-1}}. \tag{6.6}$$

Let us assume that the mutation of interest had a frequency of f_k when there were k lineages remaining in the tree. In particular, f_N is the present-day derived allele

frequency of the mutation. Using Eq. (6.6), we obtain that the probability that a mutation of frequency f_k when k lineages remained in the coalescent spread to f_N lineages when N lineages remained in the coalescent is given by

$$P\left(\sum_{\ell=1}^{f_k} Z_\ell = f_N, \sum_{\ell=f_k+1}^{k} Z_\ell = N - f_N\right) \tag{6.7}$$

$$= \sum_{\substack{z_1,\ldots,z_{f_k}: \\ \sum_{\ell=1}^{f_k} z_\ell = f_N}} \sum_{\substack{z_{f_k+1},\ldots,z_k: \\ \sum_{\ell=f_k+1}^{k} z_\ell = N-f_k}} P(Z_1 = z_1, Z_2 = z_2, \ldots, Z_k = z_k)$$

$$= \frac{\binom{f_N-1}{f_k-1}\binom{N-f_N-1}{k-f_k-1}}{\binom{N-1}{k-1}}. \tag{6.8}$$

It follows that the probability that the mutation spreads to *at least* f_N haplotypes is given by

$$p_{\mathrm{R}} = \sum_{f=f_N}^{N-k+f_k} P\left(\sum_{\ell=1}^{f_k} Z_\ell = f, \sum_{\ell=f_k+1}^{k} Z_\ell = N - f\right)$$

$$= \sum_{f=f_N}^{N-k+f_k} \frac{\binom{f_N-1}{f-1}\binom{N-f_N-1}{k-f-1}}{\binom{N-1}{k-1}}. \tag{6.9}$$

This p-value is (in theory) uniformly distributed if the mutation, as well as all other mutations on the same tree, are neutral. We can therefore reject the null-hypothesis of no selective pressure acting on the mutation (or any other mutation on the same tree) if this p-value is sufficiently small.

In this thesis, we only calculate this p-value for $f_k = 2$, where k is the number of lineages remaining when the mutation increased from $f_k = 1$ to $f_k = 2$. However, there are cases where it may be beneficial to use other values for f_k and k. For instance, if a mutation had an initial phase of neutrality and became beneficial only recently, it may be beneficial to set k according to the time when selection was turned on.

6.3 Simulated data

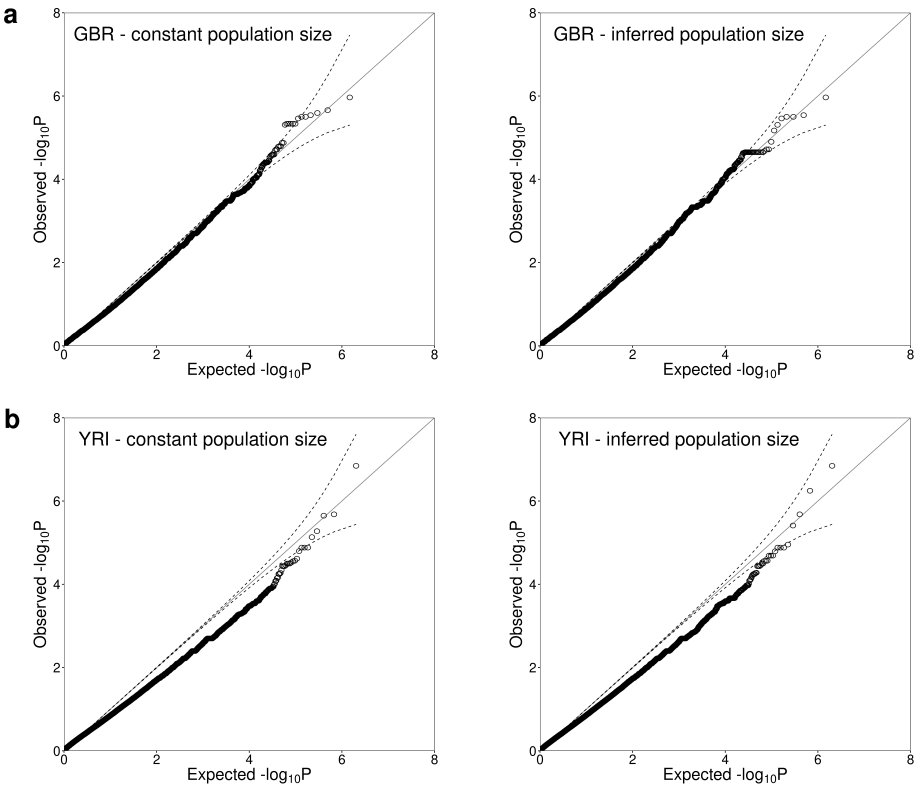


Figure 6.1: QQ-plot for p-values of selection test for neutral mutations. QQ-plot of p-values for selection evidence of SNPs. We simulated 250Mb for $N = 1000$ haplotypes using the genetic recombination map of chromosome 1 and a bottleneck population size resembling that of non-African populations.

6.3.1 Distribution of p-values under neutrality

We first investigate, whether the p-value p_{R} defined in Eq. (6.9) is uniformly distributed under neutrality. We simulate 250Mb for 1000 haplotypes with a hotspot recombination rate and population size history resembling that of GBR (Fig. 6.1**a**) or YRI (Fig. 6.1**b**). We estimate the genealogy either using a constant population size, or jointly infer branch lengths and population size histories. In both cases, a QQ-plot confirms that p-values are approximately uniformly distributed, confirming robustness of this approach with respect to the underlying demographic history.

6.3.2 Simulating positive natural selection

To simulate positive natural selection we adopt the pipeline outlined in Ref. [40]. We first simulate the trajectory of the derived-allele frequency using simuPOP [118]. Here, we vary the selection coefficient between $s = 0.001$ and $s = 0.01$ and assume

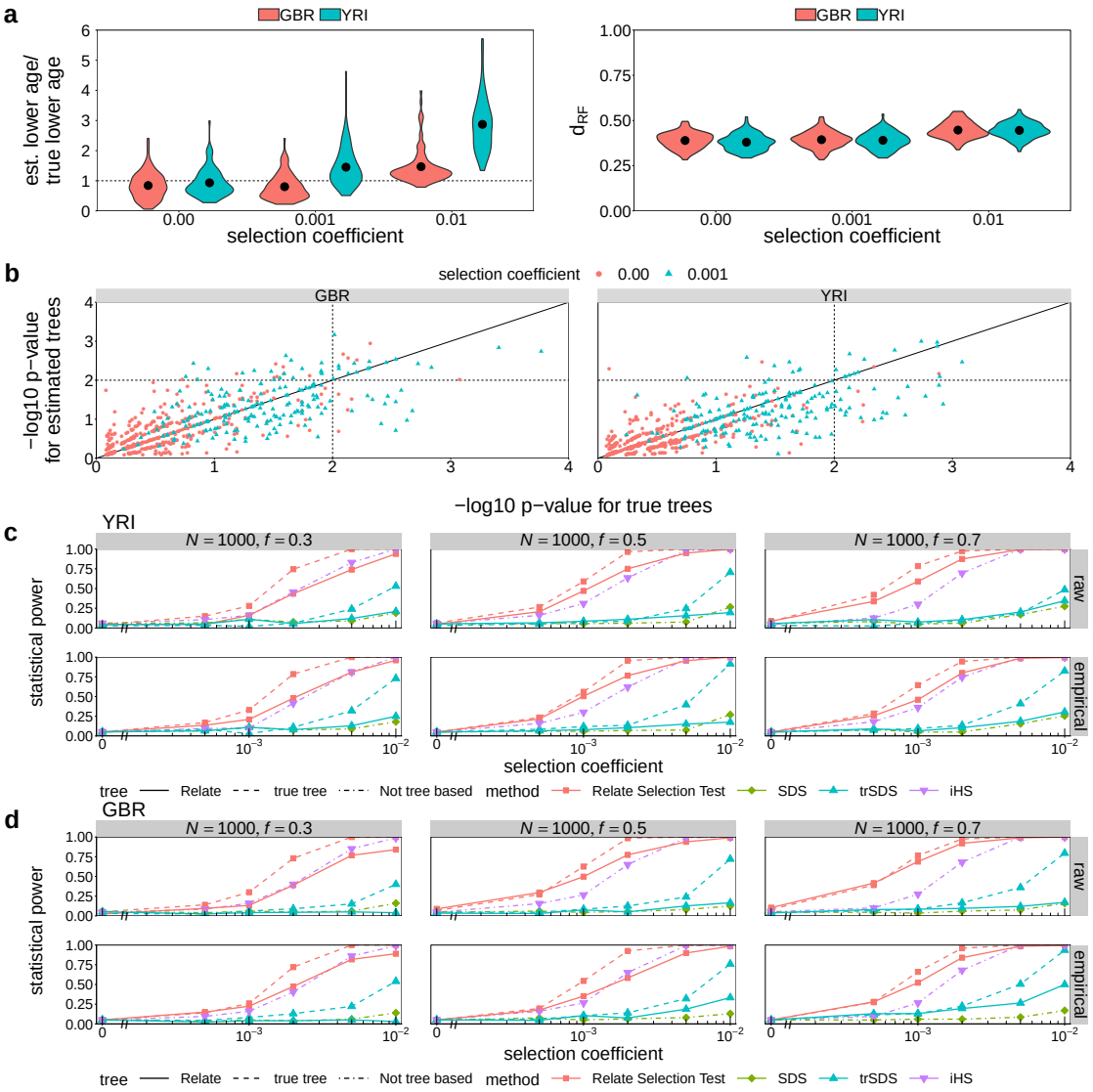


Figure 6.2: Statistical power of the selection test.
a, Ratio of estimated and true lower-end ages of the branches onto which a mutation with present-day DAF of 0.5 maps. This mutation has a selection coefficient of $0, 0.001$, or 0.01 and is positioned at 10Mb of a 20Mb simulated genomic region with Relate-estimated population size histories for GBR and YRI. We simulated 100 realisations of $N = 200$ haplotypes. Circles indicate the mean ratios. **b**, P-values for selection evidence in simulations calculated using true trees (horizontal axis) and estimated trees (vertical axis) for the same simulation scenario as in **a**. We plot p-values for loci under no selection (circles) and loci under weak selection (triangles). **c**, **d**, Power simulations with $N = 1000$ haplotypes and present-day derived allele frequencies of 0.3, 0.5, and 0.7. We assume a population size history estimated for YRI (**c**), and GBR (**d**), respectively. The significance threshold is 0.05. We show power estimates using the p-values for trees estimated by Relate, as well as those for the true trees. In both cases, we estimate power using raw p-values of our test statistic (top row) and empirical p-values given the distribution of raw p-values in the neutral case (bottom row). For iHS, SDS, and trSDS, power is estimated by standardising raw scores by the frequency specific mean and standard deviation under the null. In the top row, we assume a standard normal distribution of the standardised score and in the bottom row, we calculate empirical p-values by determining a critical score corresponding to the 0.05 significance level in the neutral case.

that the selected allele is beneficial throughout its history. We fix the present-day derived allele frequency to 0.3, 0.5, or 0.7. We then use mbs2 [156] to simulate a region of 20Mb given the derived allele frequency trajectory for the selected loci. We place the loci under selection at 10Mb. We use a mutation rate of $\mu = 1.25 \times 10^{-8}$ and a constant recombination rate of 5×10^{-9}. For each non-zero selection coefficient, we perform 200 simulations and we perform 500 simulations for the neutral case. We simulate with a population size history estimated for YRI and GBR.

To assess the accuracy of genealogies estimated by Relate at loci under selection, we calculate Robinson-Foulds distance (d_{RF}) between the true and estimated tree at the selected locus, as well as the ratio of estimated and true lower-end age of the branch onto which a mutation under selection maps (Fig. 6.2**a**). We find that tree topology (d_{RF}) remains accurate, but decreases slightly, potentially due to fewer mutations mapping underneath the selected mutation, such that tree topology is less accurate in this part of the tree. The age of a mutation under selection, approximated by the lower-end of the corresponding branch, is over-estimated when the mutation is selected, which is expected because our model assumes neutrality.

In Fig. 6.2**b**, we illustrate the distribution of p-values for neutral mutations ($s = 0.00$), as well as weak positive selection $s = 0.001$ by plotting p-values calculated with the true trees against p-values calculated using estimated trees. In this plot, the present-day derived allele frequency is 0.5. We observe a high correlation and a clear shift in p-values when selection is turned on. While many p-values remain above 0.01 even when $s = 0.001$, we note that the clear shift in the distribution of p-values can be utilised to detect selection on polygenic traits in Section 6.5.

6.3.3 Statistical power

In Fig. 6.2**c**, we estimate the statistical power of our selection test for varying present-day derived allele frequencies. We compare our method to the integrated haplotype score (iHS) [160] and singleton density score (SDS) [40], as well as the tree-based SDS (see Section 6.1). For iHS, SDS, and trSDS, we standardise the raw scores using the frequency specific empirical mean and standard deviation for the

neutral case. Because it is usually unknown which mutations are neutral, this is an idealised setting that should favour power estimates of these methods.

Across a range of selective advantages and SNP frequencies (Fig. 6.2**c**), our approach increases power relative to (tr)SDS, as well as iHS for weaker selection in particular. trSDS is more powerful than SDS, while applying the Relate Selection Test to true genealogical trees yields a test that is uniformly more powerful than other approaches, indicating the strength of tree-based approaches. In practice, there is some decrease in power from the need to infer trees via Relate. The increased power of our statistic for detecting weak selection might be particularly beneficial when investigating selection on complex, polygenic traits, where small effect sizes mean the selection coefficients on single loci are expected to be small [140].

6.4 Evidence for positive selection acting on the human genome

We calculate p_{R} for each bi-allelic SNP in the 1000 Genomes Project data set that pass quality filters across 20 populations, excluding populations in the Americas as well as African American groups. We apply quality filters to reduce the false-positive rate in detecting selection.

6.4.1 Filtering of SNPs based on the quality its tree

First, we calculate for each SNP, using the world-wide tree onto which the SNP maps, the number of mutations mapping to the world-wide tree and the fraction of branches (excluding tip branches) with at least one SNP. We then exclude any SNP for which either of these two quantifies falls within the bottom 5th percentile. This excludes approximately 5% $\sim$ 7% of SNPs.

6.4.2 Genome-wide significant hits for positive selection

We identify 35 regions containing genome-wide significant signals ($p_{\mathrm{R}} < 5 \times 10^{-8}$), using the stringent criterion that this threshold is reached separately in each of at least three populations. These regions are determined by clustering genome-wide significant hits into blocks, such that any two SNPs in different blocks have a squared Pearson correlation coefficient $r^2 < 0.5$, and for any SNP, there always another SNPs in the same block with $r^2 \geq 0.5$. We then extend these identified genomic regions, by including any SNP within 2Mb with $r^2 \geq 0.5$ to any of the genome-wide significant hits in a block.

In Table A.2, we list all identified regions, where we choose the SNP with the lowest p-value in that genomic region for each population. We additionally show genes, eQTLs (GTEx [55]), GWAS hits, and non-synonymous substitutions at SNPs with highest r^2 to the listed mutation.

We observe that all considered geographic regions (AFR, EUR, SAS, EAS) are represented in this table. Of all identified regions, 11 have been previously reported, including the LCT region associated with Lactose tolerance in Europeans, and a mutation in the EDAR gene in East Asian populations [17, 36]. In both cases, the causal variant is in high linkage disequilibrium (LD) to the mutation with lowest p_{R} ($r^2 \geq 0.8$).

Among unreported regions, we identify the EDARADD gene – which interacts with the EDAR gene [134] in the formation of hair follicles, sweat glands, and teeth [17] – as exhibiting selection evidence in all South Asian populations, as well as the Finnish population and reaching $p_{\mathrm{R}} < 10^{-6}$ in all European populations. In 16 of 35 regions, we identify GWAS catalogue hits (OR = 6.44; $p = 0.01$), non-synonymous mutations (OR = 2.49; $p = 0.16$), or eQTLs (OR = 1.74; $p = 0.1$), in LD with the mutation with strongest selection evidence ($r^2 \geq 0.8$, Methods), suggesting functional effects, reaching statistical significance for the case of GWAS hits despite the small number of cases tested. Only 8 of the 35 regions are attributed to European populations, and 18 regions are found only for African populations.

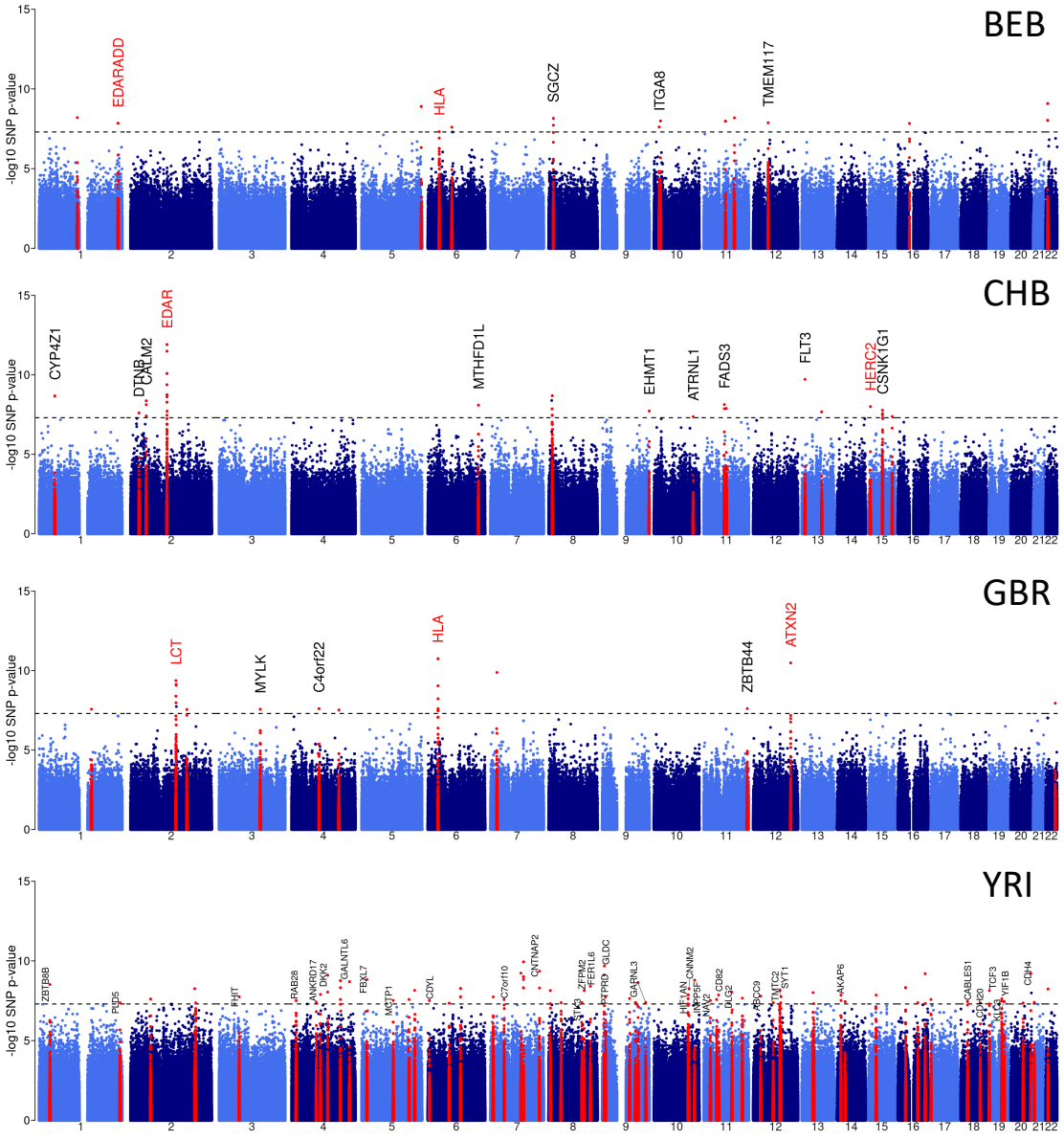


Figure 6.3: Manhattan plots for selection evidence.
Manhattan plots showing p-values p_{R} for selection evidence genome-wide for BEB, CHB, GBR, and YRI. Dashed line indicates $p_{\mathrm{R}} = 5 \times 10^{-8}$, which is the Bonferroni-corrected threshold for genome-wide significant signals of positive selection. Regions with at least one genome-wide significant p-value are highlighted in red.

We note that the number of genome-wide significant hits is much larger than Table A.2. For instance, in BEB, CEU, GBR, KHV, and LWK, we detect genome-wide significant signals in the MHC region on chromosome 6 (also see Fig. 6.3). This region is known to contain proteins that recognise foreign substances and is directly linked to immune response and autoimmune or inflammatory diseases [102]. We observe a large number of genome-wide significant selection signals in African populations, as well as an excess of nominally significant signals. For instance in YRI, 1,022,871 SNP have a selection p-value $p_{\mathrm{R}} < 0.05$, which is about 8.6% of SNPs for which a selection p-value was calculated. In comparison, 482,333 SNPs or 7.1% of all tested SNPs reach $p_{\mathrm{R}} < 0.05$ in GBR. In YRI, 10^{-5} of all tested SNPs reach $p_{\mathrm{R}} < 5 \times 10^{-8}$ and 3.5×10^{-6} of all tested SNPs reach the same threshold in GBR.

Population bottlenecks are expected to strengthen the effect of genetic drift and weaken selection signals leading to decreased power for detecting selection in non-African groups. It is however also possible that p-values are slightly inflated in African groups, due to increased levels of admixture and introgression for which we have presented some evidence in Chapter 5. Lineages attributed to distinct subspecies or subgroups can increase the number of lineages remaining when a mutation arises because these lineages cannot coalesce with the remaining population back in time. They also have fewer than expected descendants, because they remain separated and unable to branch forwards in time until after the introgression event. Such events are therefore likely confounders for our selection statistic; mutations in the remaining human lineages may appear to be out-competing these introgressing lineages which can lead to a signal resembling that of positive selection.

6.4.3 Enrichment of SNPs with functional annotation among targets of positive selection

Next, we test for enrichment of SNPs with functional annotation among targets of positive selection.

We merge selection evidence for SNPs by region (AFR: Africans, EAS: East Asians, EUR: Europeans, SAS: Southern Asians) by first calculating zscores of

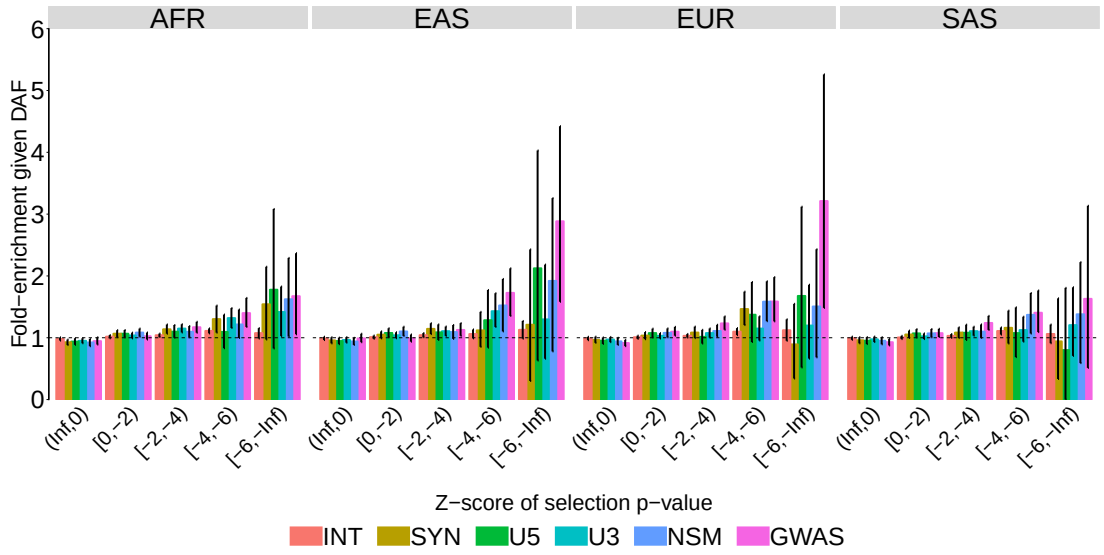


Figure 6.4: Enrichment analysis for SNPs with functional annotation among targets of selection.
Mean enrichment of functional annotation among targets of selection, conditional on allele frequency. Error bars show 95% confidence intervals estimated from 1000 iterations of a block bootstrap resampling. We group SNPs by mean regional Z-score corresponding to the log p-value for selection evidence, where a smaller Z-score indicates stronger selection evidence. SNPs are binned by partially overlapping functional annotations: intronic mutations (INT), synonymous mutations (SYN), mutations at the 5' end and 3' end of a gene (U5, U3), non-synonymous mutations (NSM), and GWAS hits (GWAS).

logarithm of selection p-values within populations, and then averaging these zscores across populations. We also average derived allele frequencies across populations. We exclude highly admixed groups, such as African-American groups and groups from the Americas, because recent admixture may confound selection signals. We further exclude SNPs with a DAF $<5\%$ in the region of interest.

To assess statistical significance for the observed enrichment of GWAS hits and functional mutations in groups of SNPs showing evidence of selection, we use a block bootstrap with a block size of 1Mb. This will account for LD at scales below this threshold. In each bootstrap iteration, we resample blocks containing SNPs with a selection Z-score within the range of interest, with replacement, and calculate the ratio of the number of SNPs with functional annotation obtained using the HaploReg database [64, 162] and the GWAS catalogue [59, 92] to the expected number of such SNPs, conditional on DAF. We condition on frequency, to account for the possibility that skewed frequency spectra in functional SNPs could be driving

the signal. We only consider GWAS hits with a p-value of less than 5×10^{-8}.

After controlling for these factors, we observe enrichment above expectation in most annotation categories, with significance observed particularly often for intronic, synonymous, and non-synonymous, and GWAS SNPs (Figure 6.4). Importantly, the strongest enrichment is for GWAS SNPs, where we observe a trend of increased enrichment among increasingly likely targets for positive selection. These GWAS SNPs have known associations with function in human groups and so encouragingly support a link between evidence of selection and SNPs with detectable influences on phenotypes at the organism level.

6.5 Evidence for polygenic adaptation

Selection acting through a handful of mutations with large effects is rare, partly because mutations with large effects are often deleterious and therefore rare themselves, and partly because traits more commonly depend on a complex architecture of many genes. Selection on complex traits therefore is expected to happen mainly via weak selection on many loci of smaller effects, driving phenotypic change by small changes in allele frequencies at many loci [140].

Using the SNP-trait associations documented in large-scale GWAS, including those recorded in the GWAS catalog [92], and GWAS conducted with the UK Biobank [19], we test for directional positive selection acting on individual traits. More specifically, we test whether derived mutations that increase (or decrease) a trait show increased evidence of directional selection relative to randomly sampled control mutations of the same frequency.

6.5.1 Pre-processing of GWAS hits

We use only associations with GWAS p-value smaller than 5×10^{-8}, because confounding due to population stratification is thought to operate through relatively

small - but systematic - biases in effect size estimates [11, 142], but is not known to produce false-positives that are genome-wide significant.

For every combination of population, trait, and effect direction, we thin genome-wide significant GWAS hits to account for LD. Using GWAS documented in the GWAS catalogue [92], we first group SNPs into approximately independent blocks, defined such that any two GWAS hits in separate blocks are separated by at least 100kb and there are no intervals larger than 100kb with no GWAS hit inside a block. We note that the union of these blocks do not cover the entire genome. We then choose one GWAS hit from each block uniformly at random. We remove any SNP with a DAF <5%. To determine the effect direction of a SNP, we use the annotation in column "95% CI (TEXT)" combined with the indicated risk allele. We then realign the effect direction to the derived allele. We only consider SNPs for which an effect direction can be determined with this procedure.

At each such SNP retained, we use only the association direction, rather than its strength, to offer additional robustness to potential confounding. We restrict our analysis to traits with at least 10 independent hits in both effect directions in all populations. This results in 76 traits and a total of 7302 GWAS hits (before filtering for SNPs in close proximity in each population).

For Schizophrenia, we are unable to obtain an effect direction using the procedure described above. Instead, we download results for a large-scale GWAS conducted by the Psychiatric Genomics Consortium [89, 137]. As before, we only consider SNPs reaching a GWAS p-value of 5×10^{-8}, of which there are 9138. We intersect this set of SNPs with SNPs segregating in each of the considered populations. As for the GWAS catalogue, we identify approximately independent blocks. We then choose the SNP with lowest GWAS p-value in each block, resulting in 81 to 89 hits per population.

In addition, we use GWAS conducted as part of the UK Biobank [19, 158], focussing on highly polygenic physical traits. Our pre-processing protocol is analogous to that for schizophrenia detailed above. The number of approximately independent hits per population range from 272 hits for waist circumference to 989 hits for standing height.

6.5.2 Trait selection test

For every combination of population, trait, and effect direction, we test whether p-values are smaller than expected. For this test, we first sample SNPs that we use for comparison. For each SNP associated with the population, trait, and effect direction tuple of interest, we sample 20 SNPs uniformly at random with replacement from SNPs, with the same present-day DAF in the population of interest. We then use a one-sided Wilcoxon rank-sum test to test whether the p-values of SNPs associated with the tuple of interest tend to be smaller than those for the frequency-matched set of SNPs. We repeat this test 20 times and report the mean p-value of the Wilcoxon rank-sum test.

If positive selection occurs so as to advantage SNPs influencing a trait in a certain direction, e.g. trait-increasing, we would expect positive selection on trait-increasing mutations, and negative selection on trait-decreasing mutations. In general, we expect our test to be sensitive mainly to the former, because selection will increase frequencies of such SNPs, and the Relate Selection Test has reduced power to identify selection at rarer markers (Figure 6.2**b**). However, for traits with a large number of hits, and strong selection, it is theoretically possible for our approach to observe some selection evidence in both directions [83, 94], because to avoid ascertainment effects we condition on SNP allele frequencies at trait-influencing sites. Therefore, we additionally test for differences in present-day DAFs between trait-increasing and trait-decreasing mutations. This test can provide orthogonal evidence of polygenic adaptation because we condition on SNP frequency in our first test, aiding interpretation of results.

To test for shifts in DAF, we conduct a one-sided Wilcoxon rank-sum test to test whether DAFs of SNPs associated with the effect direction with selection evidence tend to exceed those associated with the opposing effect direction. We note that we expect to lack power to reliably detect selection with this test, given that there are typically only tens of SNPs independently associating with each trait. In addition, the relationship between selection and SNP frequencies can be complex if selection strength varies through time and/or geographic locations.

6.5.3 Interpretation of polygenic selection

Before discussing specific selection signals detected by our test, we note that some caution needs to be exerted when interpreting polygenic selection [114]. An important issue is the assignment of selection to specific phenotypes. This remains challenging for multiple reasons. For example, a directional signal might be partly driven by selection on other phenotypes correlated to those studied. Moreover, even if mutations e.g. increasing WBC counts have been generally favoured in a group, this does not imply that WBC count itself has increased evolutionarily; if for example a selective sweep has fixed a single SNP of major effect on this phenotype (such as Duffy negativity in Africa, associated both with malaria resistance and decreased WBC count [72]), then selection might be acting on other SNPs to compensate this change. Environmental influences might have similar impacts.

Differences between populations must also be interpreted carefully: aside from impacts of demographic history, most human GWAS's to date have been conducted in European populations, so that recently arisen phenotype-influencing mutations in other groups might not have been observed, reducing power in those populations.

Finally, we note that we only utilise the direction of association signals in testing for selection evidence, and test derived mutations, in order to increase robustness to residual population stratification still present in a GWAS, even after attempts to correct for such stratification. We believe that this is likely to resolve the most serious known issues, except in a setting where residual stratification (which can correlate with selection evidence [122]) improves power to observe effects that are genome-wide significant in one direction vs. another. Implicit in our approach is the idea that stratification issues are relatively far weaker for potentially genome-wide significant SNPs (of relatively large effect size) compared to directly using effect size estimates - which may be comparable to the strength of bias – across many or all SNPs genome-wide.

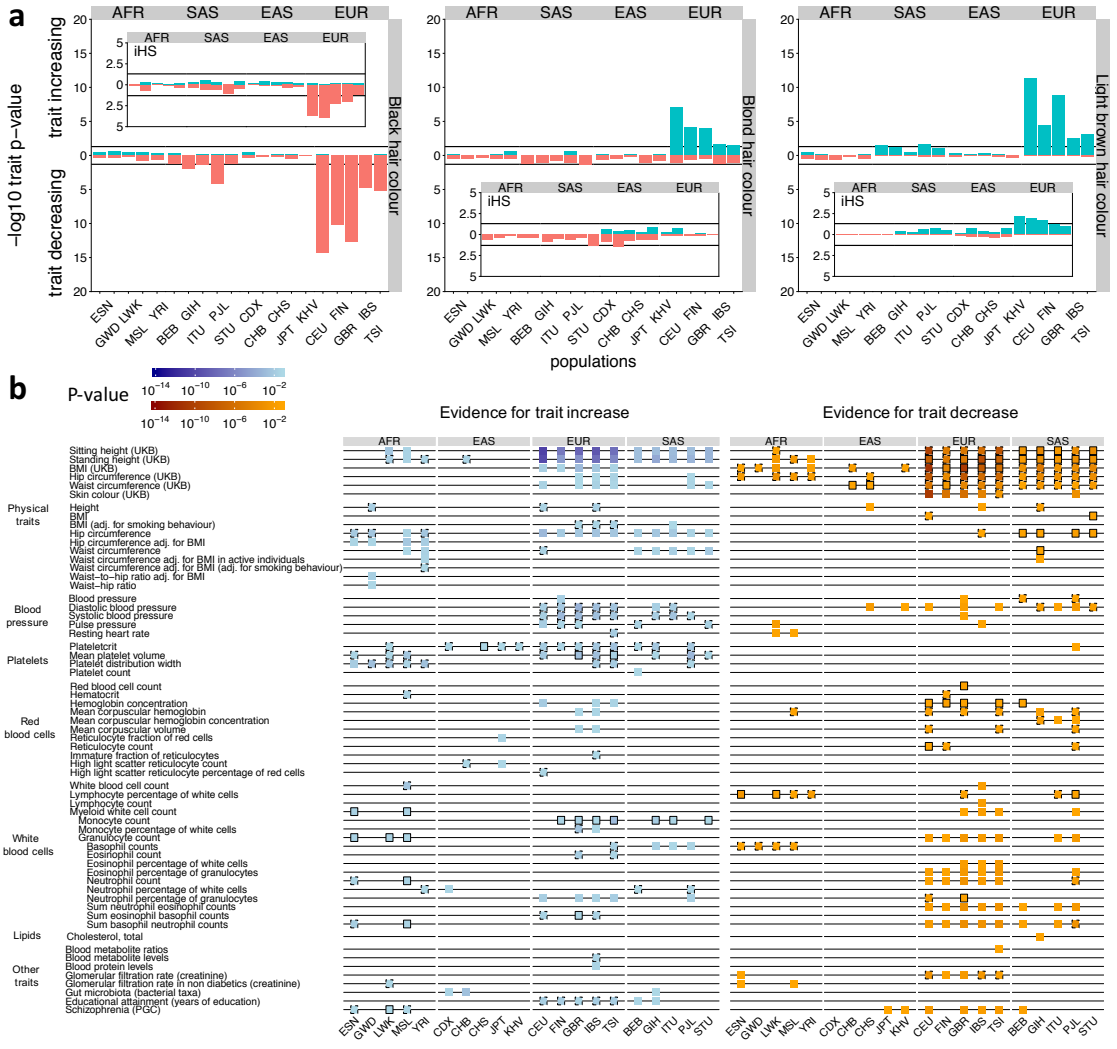


Figure 6.5: Evidence of selection on traits.
a, P-values for evidence of directional selection of black, blond, and light brown hair colour (see Section 6.5.2 for calculation of p-values). Insets show p-values for the same test but using iHS scores instead, where iHS scores are calculated for each population separately for any variant with a minor allele frequency >5% in that population. **b**, Evidence for directional or bidirectional selection on multi-allelic traits. Each trait is associated with at least 10 SNPs in both effect directions in each of the considered populations. We show evidence for a trait increasing over time (left) and evidence for a trait decreasing over time (right) if $p \leq 0.05$. Black boundaries indicate consistency with an additional test that tests for shifts in the DAFs (solid: $p \leq 0.05$, dashed: $p \leq 0.5$).

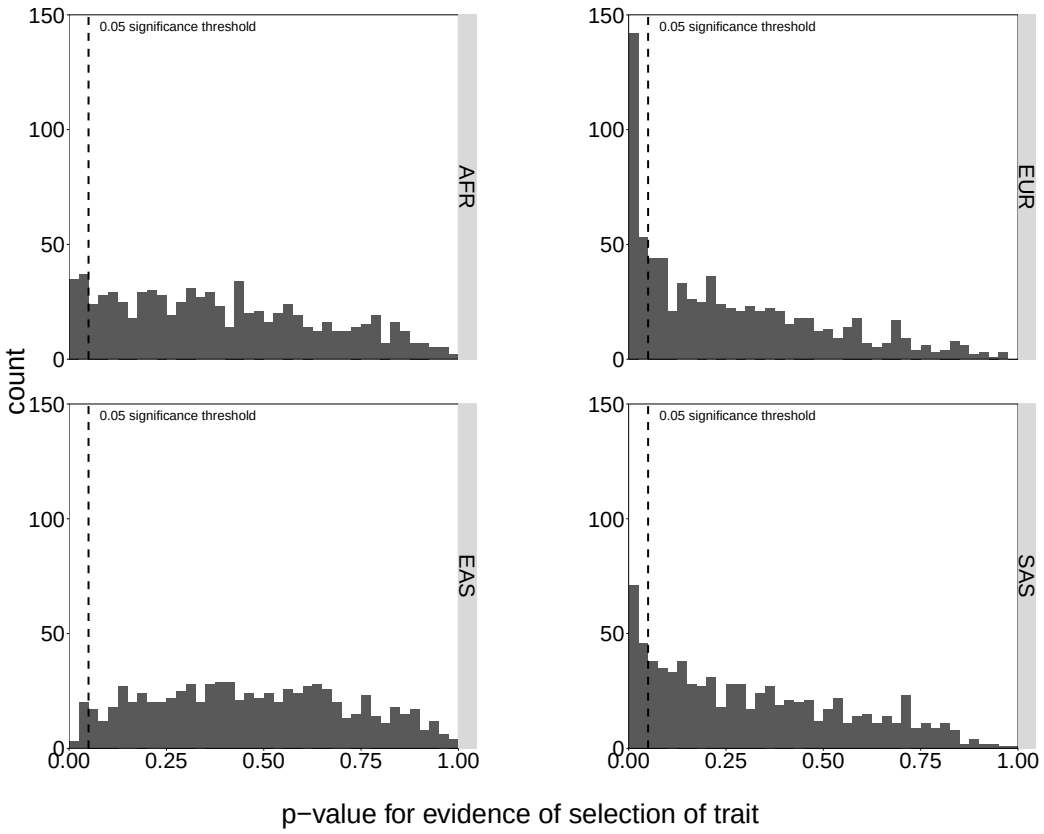


Figure 6.6: Histograms of p-values for evidence of selection on complex traits. We aggregated both effect directions of 84 considered traits, as well as populations in each of the four considered geographic regions (AFR, EAS, EUR, SAS).

6.5.4 Positive control: blond hair colour

As a positive control, we applied our test to GWAS for hair colour conducted for the UK Biobank (Fig. 6.5**a**). As in previous studies [22, 40, 163], we find a signal for SNPs associated with blonder hair colour among European populations, which is absent in South Asian populations [22, 40]. We observe strong selection for a decreased black hair colour, as well as an increase in light brown hair colour among all European populations, and weaker signals in South Asian populations, while East Asian and African populations show no evidence of selection. Testing based on iHS scores identifies only some of these signals, and with significance decreasing around 4 orders of magnitude (Fig. 6.5**a**).

6.5.5 Other traits

Next, we applied the same test to 84 traits: 6 from the UK Biobank, and 78 with at least 10 genome-wide significant GWAS catalogue association signals in each effect direction. We tested all populations except recently admixed groups; 61 of these 84 showed nominal evidence for selection ($p < 0.05$) in at least one population (Fig. 6.5**b**), with strong geographic clustering and the most significant signal ($p = 6 \times 10^{-14}$) for SNPs associated with decreased Body Mass Index (BMI) in CEU. The largest number of selection signals are observed for Europeans, possibly because many GWAS were conducted in these groups. Interestingly, East Asians have the fewest selection signals and no enrichment of low p-values (Fig. 6.6), which may partly be explained by their stronger population bottleneck and the fact that most included GWAS were conducted in Europeans, both of which would theoretically be expected to weaken selection signals in East Asian groups.

Height, Body Mass Index (BMI), and Schizophrenia have been studied previously and show a large number of association signals [18]. While several studies have reported genetic differentiation between populations for these traits [31, 115, 131, 157], evidence for selection remains controversial [11, 12, 35, 40, 131, 142, 157, 169]. It was recently reported that some studies reporting recent selection on increased height in Europeans have been confounded by subtle population stratification [11, 35, 142]. Our test finds an enrichment of selection evidence for both effect directions in the same population for height, across most populations except East Asians, using the large collection of UK Biobank associations. DAFs tend to be larger towards the height-decreasing direction. This complex picture may be a consequence of both negative and positive selection acting on height, as well as pleiotropy; SNPs impacting other traits might also impact height. We also identify strong evidence of selection towards decreased BMI across all populations, with agreement of DAF shifts, indicative of directional selection. For both traits, we detect little evidence of selection in using associations in the smaller GWAS catalogue collection. Schizophrenia has evidence of selection towards decreased risk in Europeans, and

some South and East Asian populations, while African populations show selection evidence towards a risk increase.

Overall, although we find selection evidence for a range of traits, we observe little overlap with traits identified in Ref. [40], which focuses on very recent selection specific to the British population. Among other phenotypes, we see selection evidence for a variety of blood-related phenotypes, with congruent DAF signals. In Europeans and some South Asian groups, we detect a strong signal favouring SNPs increasing diastolic and systolic blood pressures, contrary to previous studies suggesting selection for decreased blood pressure in these groups [22, 168]. The discrepancy to these studies, which have been primarily looking at SNP frequency differences across human groups, needs further investigation and could be a result of temporal changes in selection strength and direction. We moreover find evidence for selection favouring SNPs associated with decreased hemoglobin concentration and other related traits, while platelet-related traits appear to be selected to increase across many populations.

6.5.6 White blood cell traits

Identifying the direction of selection is sometimes challenging, with a number of traits showing selection evidence in both effect directions within the same group (Fig. 6.5). To aid interpretability of these results, we additionally tested in which direction, if any, DAFs of associated SNPs are increased. For many traits with selection evidence in one effect direction only, such as platelet-related traits, the direction of selection inferred using our polygenic selection test and the direction of DAF increase align.

Interestingly, for some traits related to white blood cells (WBC), we observe differences between the frequency conditioned selection signals and shift in DAF (Fig. 6.5**b**). For instance, we detect a signal towards increased granulocyte counts in African populations, but decreasing counts in some European and South Asian populations. While DAFs are strongly different ($p < 0.003$; one-sided Wilcoxon test) and in agreement with the inferred direction of selection in African groups, DAFs remain slightly lower for SNPs associated with decreasing granulocyte counts even in

Europe and South Asia. ($p < 0.09$ for EUR, $p < 0.12$ for SAS; one-sided Wilcoxon test). However, these DAFs for granulocyte-decreasing variants are increased relative to those in African groups ($p < 0.0013$; one-sided paired Wilcoxon test), while DAFs for granulocyte-increasing variants were not significantly different ($p > 0.4$; two-sided paired Wilcoxon test), so a possible resolution is that this trait (or a related trait) was selected to increase in the past, and has more recently been selected to decrease in some non-African groups.

6.5.7 Type-2 diabetes

We additionally investigate selection on Type-2 diabetes (T2D), using the largest trans-ethnic meta analysis of GWAS for T2D to date. This unpublished GWAS meta analysis, lead by Dr Anubha Mahajan (Univ. Oxford) and Prof. Mark McCarthy (Genentech), aggregates 171,262 cases and 1,075,072 controls from diverse populations, doubling the effective sample size from the largest European-only analysis [167]. The GWAS identifies 336 independent loci reaching genome-wide significance at a stringent p-value threshold of 10^{-9}.

We intersect all 336 T2D associations with the 1000 Genomes Project data set and only retain segregating SNPs for each population, which reduces the number of variants to 248 (MSL) - 295 (GBR). We then remove any SNP with a regional GWAS p-value > 0.5, because these have little evidence for an association with T2D in that population and effect size estimates are less accurate (Fig. 6.7). This primarily affects non-European populations and reduces the number of variants to 207 (MSL) – 295 (GBR).

We divide all T2D risk variants into risk increasing or decreasing (after alignment of effect sizes to the derived allele), and find evidence of selection in African populations towards T2D risk increase, with no additional signals elsewhere (Fig. 6.8**a**). This signal aligns with directional differentiation of some of the strongest T2D risk increasing variants towards higher risk in African populations [24], as well as significant correlation of increased T2D risk with African ancestry [25].

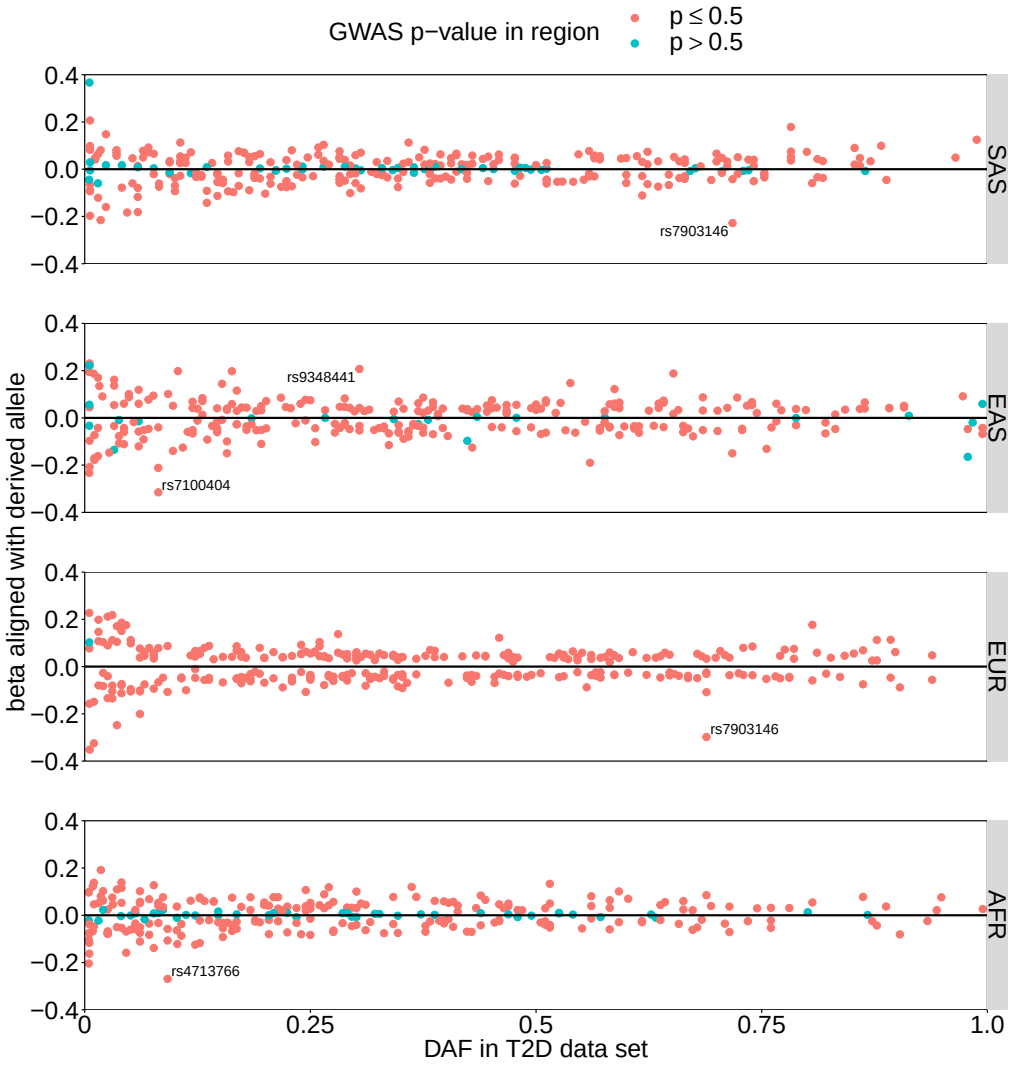


Figure 6.7: Effect sizes of type-2 diabetes associations. SNP effect sizes as a function of derived-allele frequency (DAF) by region coloured by their regional GWAS p-values. Variants with an unusually large effect size for their DAF are annotated.

T2D risk increase is likely not an advantageous trait on its own and positive selection is therefore likely not acting directly on T2D risk increase, but via pleiotropic variants associated with other advantageous traits. We find that variants additionally associated with an increase in arm fat mass are particularly young for their frequency indicative of positive selection and can explain the observed T2D selection signal (Fig. 6.8**a,b**); their exclusion removes the selection signal observed in African populations ($p > 0.05$). For such variants, we observe smaller than expected selection p-values in some European groups despite there being no

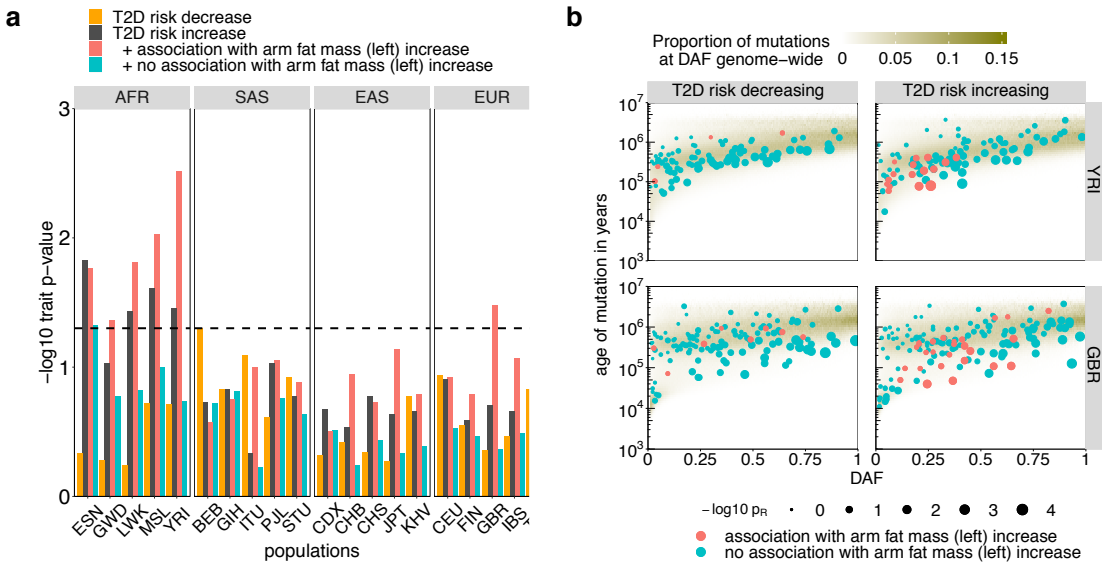


Figure 6.8: Selection on type-2 diabetes risk. **a**, Evidence of positive selection acting on variants associated with T2D risk decrease (orange) and T2D risk increase (grey). Additionally, we partitioned T2D risk increasing variants into those with an association with arm fat mass increase, and those without such an association. **b**, Age of T2D risk variants estimated using Relate by derived-allele frequency (DAF). The heatmap shows the proportion of mutations of a certain age by DAF genome-wide. Sizes indicate $-\log_{10} p_{\mathrm{R}}$, the p-value for selection evidence calculated by Relate. Associations with arm fat mass increase are highlighted in red.

selection evidence for T2D risk increase overall. Using GWAS conducted for the UK Biobank [19], we identify additional traits that may similarly explain the T2D selection signal (Fig. 6.9). These traits are related to Body mass index (BMI), weight, and body fat and share many of the SNPs associated with arm fat mass. Blood related traits and height cannot explain the signal despite possessing similar number of overlapping SNPs with T2D risk variants.

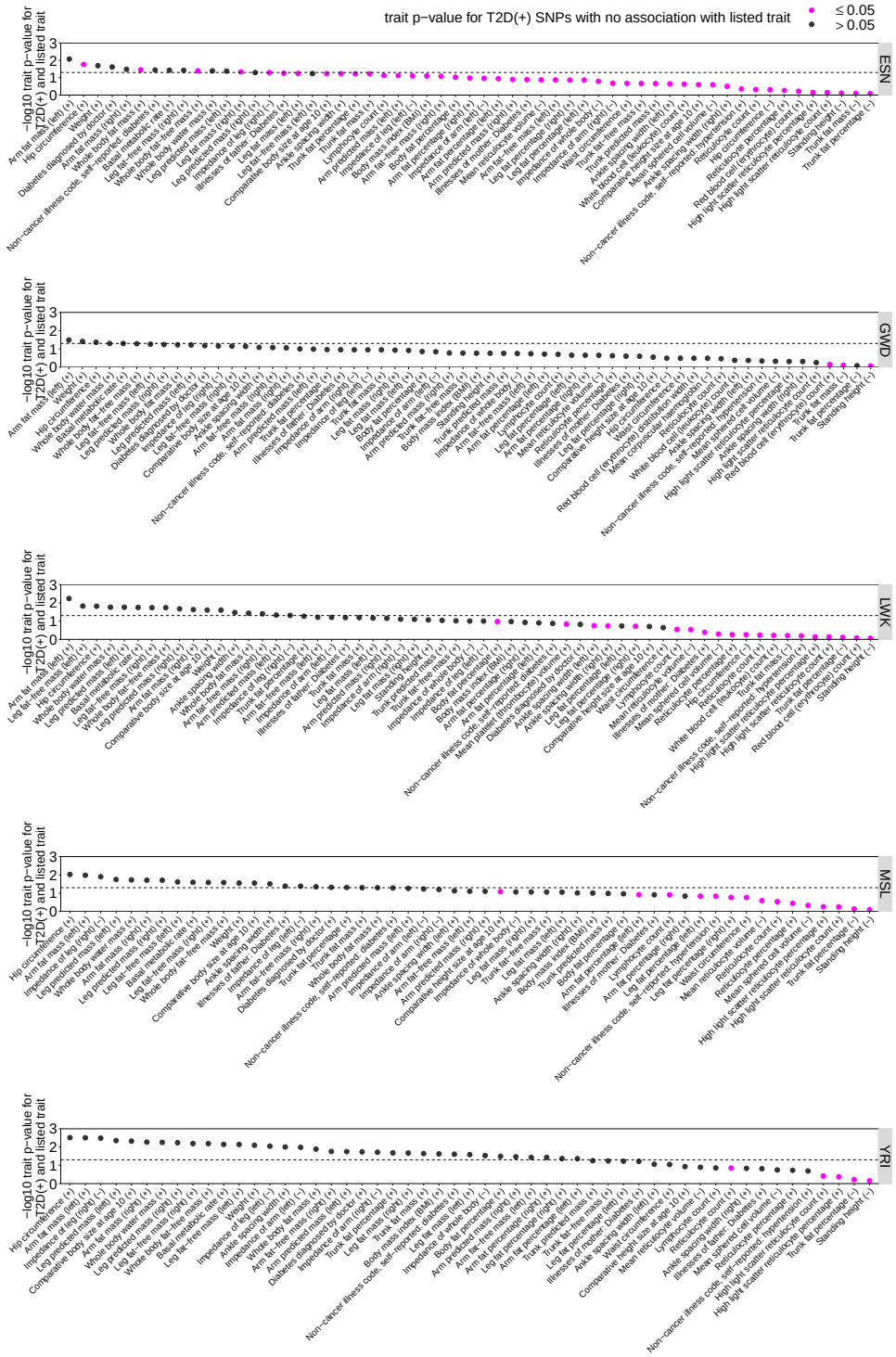


Figure 6.9: Selection on SNPs associated with type-2 diabetes risk and UK Biobank traits. Evidence of selection on subsets of T2D risk increasing variants that have at least 10 associations with the traits shown on the horizontal axis for ESN, GWD, LWK, MSL, and YRI. The dashed line indicates $p = 0.05$. Colours indicate whether the remaining SNPs with no association with the listed trait show evidence of selection.

7

Reconstructing the genealogy of 50 wild mice from France, India, and Taiwan

Contents

So far, we have tested Relate on human data, as well as simulated data with predominantly human-like parameters. To demonstrate that Relate can be successfully applied to non-human data, we apply Relate to 50 wild mice sampled in France, India, and Taiwan. These wild mice are expected to have a much larger population size compared to humans, as well as a shorter generation time and a lower per generation mutation rate [61]. Relate requires computationally phased haplotypes as its input and the accuracy of inferred phases heavily relies on a good, representative reference panel, which does not exist for the mouse samples considered here.

Each of these factors is expected to complicate the estimation of accurate genealogies. Nevertheless, as we will demonstrate in this chapter, the genealogies we obtain are sufficiently informative for a broad range of applications, including

inference of population size histories and split times of the three mice populations.

The analyses presented in this chapter are an extension of an analysis conducted by Prof. Robert Davies (Univ. Oxford) [32]. We apply Relate to estimate a joint genealogy of these 50 wild mice samples, which confirms some of the key insights gained in Ref. [32], including evidence for introgression and positive selection at three target genes: Vkorc1 in French mice, Prl in Indian mice, and Brca2 in Taiwanese mice. Our analysis additionally reveals a signature of GC-biased gene conversion in inferred mutation rates.

7.1 Data set

This data set combines unpublished whole-genome sequencing data for 20 French *M. m. domesticus* and 20 Taiwanese *M. m. castaneus* mice, as well as published data of 10 Indian *M. m. castaneus* mice [62] (Fig. 7.1). The French mice were collected by Dr Amelie Baud and Dr Binnaz Yalcin over an approximately 50km region in the South East of France. Each mouse was sampled at least 1.7 km apart. The Taiwanese mice were collected from the periphery of Taiwan by members of the lab of Prof. Alex Yu. In both populations, a principal component analysis reveals structure within each group (Fig. 7.1). All samples were sequenced to 10X on Illumina HiSeq machines and one French sample was additionally sequenced to 30X for quality control purposes. The Indian mice were from a 130 km transect in northern India with each mouse collected at least 500m apart. Quality control, alignment to a reference genome, as well as genotype calling was conducted by Prof. Robert Davies (see Ref. [32]). This resulted in a total of 48,557,437 biallelic SNPs across all 19 autosomal chromosomes or approximately 2.5 billion base-pairs of the reference genome [29]. Of the total SNP count, 37,739,553 segregated in the Indian mice, 19,903,824 segregated in the Taiwanese mice, and 10,789,452

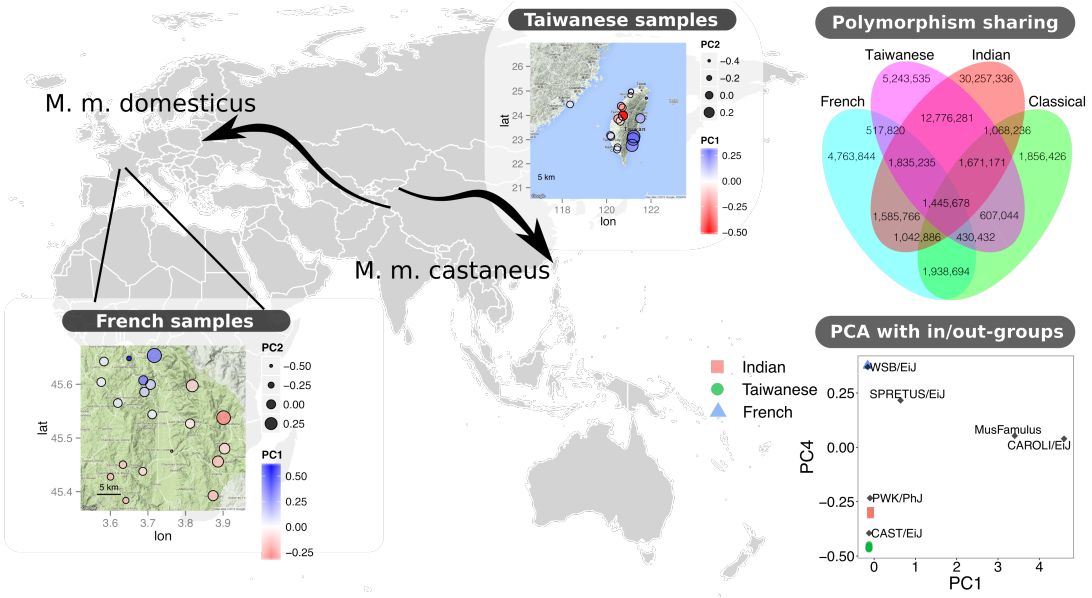


Figure 7.1: Sampling locations of Indian, French, Taiwanese wild mice. For French and Taiwanese mice, we plot sampling locations, with point size and colour indicating the first and second principal components calculated on each population separately. Sampling locations of Indian mice are unavailable. We also conducted a principal component analysis on all three populations, as well as three ourgroup mice (*Mus spretus*, *Mus famulus*, *Mus caroli*) and three labstrains (WSB/EiJ, PWK/PhJ, CAST/EiJ). We show PC1 and PC4 and note that PC1 separates *Mus famulus* and *Mus caroli* from the remaining samples, PC2 separates *Mus famulus* and *Mus caroli*, PC3 separates *Mus spretus* from the remaining samples, and PC4 separates the three *Mus musculus* populations considered. In addition, we show polymorphism sharing across all three groups, as well as 13 classical laboratory strains obtained from Ref. [78].

segregated in the French mice, already indicating a much reduced diversity in both the French and Taiwanese groups.

7.2 Ancestral allele calling and phasing

Application of Relate requires determining ancestral alleles at any biallelic SNP using an outgroup mouse strain. We treat *M. famulus* and *M. caroli* as outgroup mice (Fig. 7.2). We mask out any SNPs that are not biallelic in the 50 wild mice samples and 2 outgroup mice. At any given biallelic SNP, we choose the allele that had a frequency of at least three in the two outgroup mice as the ancestral allele.

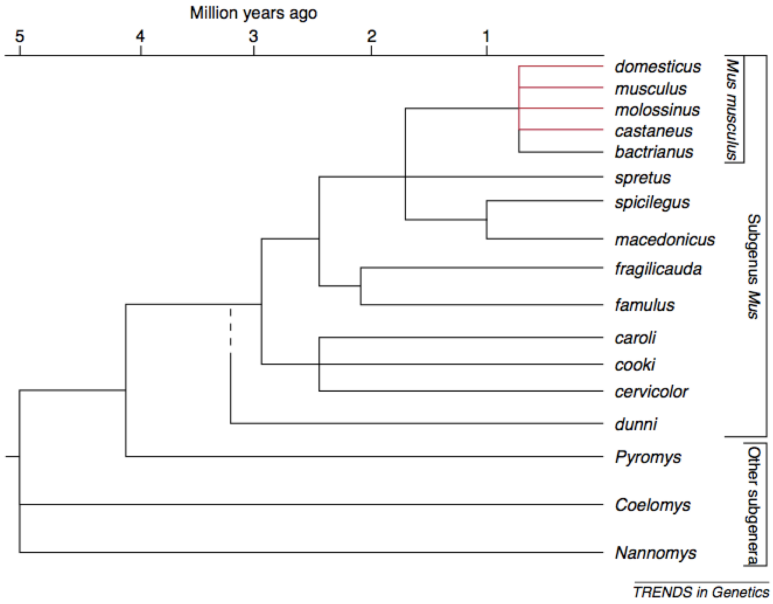


Figure 7.2: Species tree for the genus Mus.
Evolutionary species tree for the genus Mus, taken from Ref. [56]. Samples included in the data set belong to *M. m. castaneus* and *M. m. domesticus*. To determine ancestral alleles, we used *M. famulus* and *M. caroli* strains.

We then apply Shapeit2 to phase the genotypes [33]. We specify a constant diploid population size of 100,000 and use a constant recombination rate of 2×10^{-8} per base per generation because recombination hotspots are highly diverged among the three populations. As no suitable reference panel for these samples exist, we run Shapeit2 without a reference panel, but instead in read-aware mode, such that any reads with multiple SNPs are used towards phasing genotypes. In particular, singletons that have no overlapping reads with other polymorphisms are assigned a random phase.

7.3 Population size and split times

We estimate a joint genealogy of all 50 wild mice samples and estimate historical population sizes and split times (Fig. 7.3**a**). Our date estimates directly scale with the mutation rate, which we set to 3.79×10^{-9} mutations per base per generation.

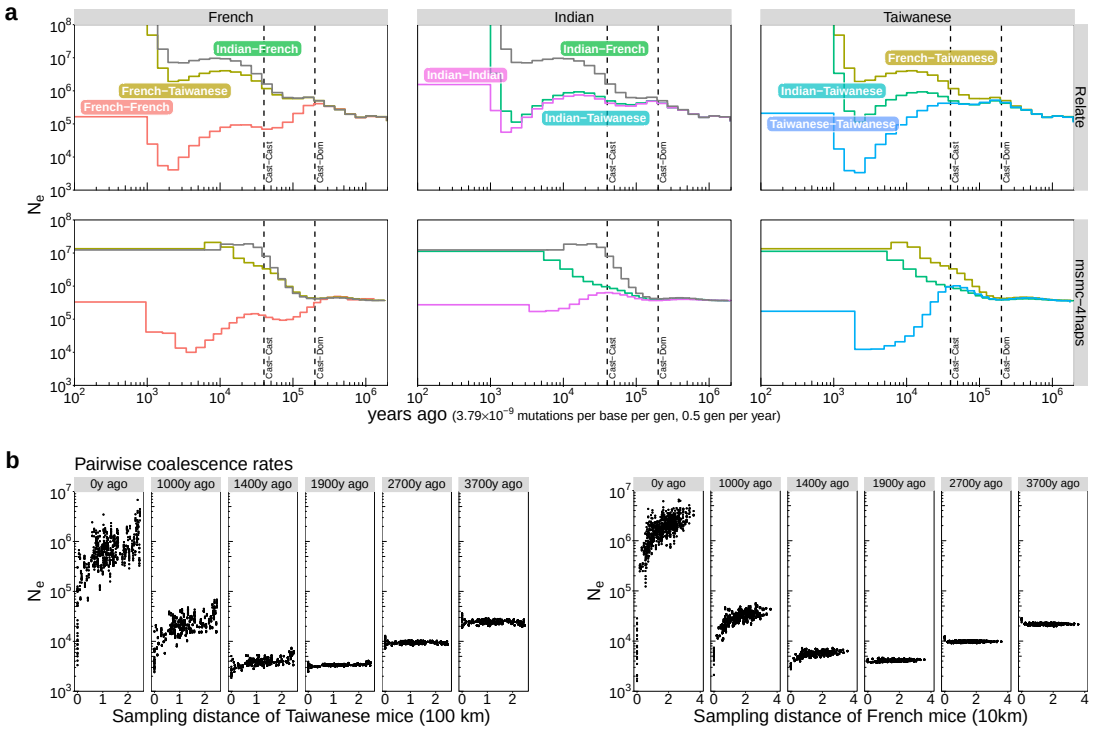


Figure 7.3: Effective population sizes and split times for 50 wild mice. **a**, Diploid effective population sizes and inverse cross-coalescence rates for French, Taiwanese, and Indian mice. (top) Estimates obtained by inferring the joint genealogy of 50 wild mice using Relate, and then jointly fitting coalescence rates and branch lengths. (bottom) Estimates obtained using MSMC2 applied to two samples (4 haplotypes), either both chosen from the same population or each chosen from a different population. Vertical lines indicate possible split times. **b**, Relate-estimated effective population sizes of pairs of haplotypes plotted against sampling distance of Taiwanese (left) and French (right) mice samples. Each panel corresponds to a time epoch.

We also assume an average of two generations per year. Both parameters were estimated in Ref. [61] based on the mouse-rat diverged time. We additionally run MSMC2 on two samples (four haplotypes), choosing both samples from the same population to estimate effective population sizes, as well as choosing samples from different populations to estimate inverse cross-coalescence rates.

Relate and MSMC estimated effective population sizes show a strong bottleneck in French and Taiwanese mice around 1,000 to 10,000 YBP, and a recovery in recent population sizes to approximately 100,000 in both groups. The population size of Indian mice is >100,000 almost throughout their evolutionary history and may be as large as 1,000,000 today. We note that the subtle bottleneck observed in Indian mice around 1,000 to 10,000 YBP is likely an artefact caused by the joint analysis

of the three populations using a panmictic model. This is confirmed in the MSMC estimates of Indian population sizes, which do not show a bottleneck (Fig. 7.3**a**).

The observed population size estimates are consistent with previously reported strong founder effects in the French *M. m. domesticus* and Taiwanese *M. m. castaneus* due to migrations from central Asian populations [56]. The bottlenecks in both French and Taiwanese mice appear to be very recent compared to their respective split times with the Indian mice, peaking at around 2,000 - 3,000 YBP. This may be an artefact shared in both methods, which tend to smooth deep, sudden bottlenecks. However it is also possible that these wild mice only arrived recently; for instance, it is believed that Taiwanese mice only arrived a few thousand years ago, accompanying human migrations [152].

Using inverse cross-coalescence rates, we find that French mice diverged from Taiwanese and Indian mice at around 200,000 YBP, while Taiwanese and Indian mice diverged at around 40,000 – 50,000 YBP (vertical lines in Fig. 7.3**a**). This is consistent with classification of Taiwanese and Indian mice to the same subspecies M. m. castaneus, while French mice are classified to the subspecies *M. m. domesticus*. Previous split time estimates suggest a *M. m. castaneous*-*M. m. domesticus* split at 170,000 - 270,000 YBP, which is consistent with our estimate [152].

Interestingly, we observe that the Relate-estimated Indian–Taiwanese inverse cross-coalescence rate does not diverge and instead remains almost identical to the Indian population size for long periods after the two populations split. In other words, Indian mice coalesce with Taiwanese and Indian lineages at similar rates, while Taiwanese mice have a comparatively higher rate for coalescing with other Taiwanese lineages. This pattern could be explained by weak but persistent directional gene-flow from Indian ancestors into Taiwanese ancestors, which would make some Taiwanese lineages appear similar to Indian lineages. However, it should be noted that MSMC2 estimates qualitatively disagree with Relate on this point and confirmation of this hypothesis will require developing methods that can detect directional migration patterns (see discussion in Chapter 8).

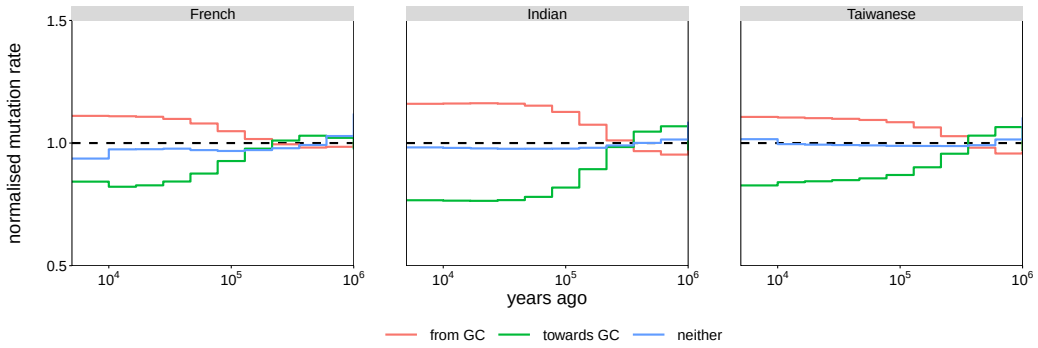


Figure 7.4: Signal of GC-biased gene conversion in mutation rate trends of wild mice.
Evolution of mutation rates for French, Taiwanese, and Indian mice, grouped by whether sites mutate towards G/C, away from G/C, or neither. We eliminate any shared temporal trends by dividing by the average mutation rate in each epoch. Mutation rates are relative to the mean mutation rate through time in the "neither" category.

In Fig. 7.3**b**, we show the Relate-estimated effective population sizes (inverse coalescence rates) of pairs of haplotypes as a function of sampling distance in French and Taiwanese mouse samples. This reveals structure on fine scales, with haplotypes sampled at closer distance tending to have a smaller population size (higher coalescence rate) relative to haplotypes sampled at a larger distance. This structure is strongest in the most recent bin (0 to 1000 years ago), and disappears entirely about 2000 years ago in both populations. This supports localised mating patterns and separation by distance in recent times and disappears roughly coinciding with a strong population size bottleneck in both populations. Further investigation is necessary to conclude whether the disappearance of detectable structure reflects the timing of arrival of these mouse populations in France and Taiwan.

7.4 Mutation rate through time

We find that trends in our estimated historical mutation rates are dominated by an effect that can be explained by GC-biased gene-conversion [34] (Fig. 7.4) and is stronger than the similar effect observed in humans (see Section 5.6). Mutation

rates towards G or C are decreasing, while those from G or C are increasing towards the present and mutations not falling into either categories remain approximately constant. This is consistent with mutations towards G and C diffusing more quickly, such that these are overrepresented further back in time and may also appear older than their actual age, leading to inflation in the mutation rate. In contrast, GC-biased gene conversion predicts that mutations from G and C experience negative selective pressure and may be lost at a higher than expected rate while also appearing younger than their actual age. This would lead to a decreasing mutation rate back in time.

GC-biased gene-conversion is understood to act more effectively in populations with a larger population size, because increased genetic diversity increases the chance of polymorphisms inside a non-crossover event [50, 109]. In addition, it is possible that a greater PRDM9 diversity can lead to less concentrated hotspots such that GC-biased gene conversion impacts a greater proportion of the genome. Consistent with this reasoning, we observe a larger effect in Indian mice compared to French and Taiwanese mice and a larger effect in wild mice compared to humans.

Importantly, the effect of GC-biased gene-conversion on mutation rate estimates may propagate to bias estimates of historical population sizes [121]. For instance, if we estimated branch lengths only using mutations towards G or C, we would tend to estimate a comparatively smaller population size in the past because the Relate algorithm will attempt to elongate old branches to fit a constant mutation rate. It is unclear how inclusion of all present-day observed variation affects estimated population sizes, however these too might be biased because of biased gene-conversion. Future work might include estimating population sizes using different classes of SNPs, such as SNPs in low recombination regions where GC-biased gene conversion is expected to be acting less strongly.

7.5 Evidence of introgression and positive selection at the Vkorc1, Brca2, and Prl genes

Ref. [32] identified three genes with evidence for introgression and positive selection by counting the number of fixed derived variants private to a particular population, as well as calculating Tajima's D statistic and Fay and Wu's H statistic in 1Mb non-overlapping genomic windows (see Section 6.1, Figure 7.5). Here, we extend this analysis using the inferred joint genealogies of the three mouse populations.

In the French mice, a subregion on chromosome 17 has a large excess in singleton mutations in one mouse sample, which is strong evidence for introgression persisting over about a 4Mb region. Calculating Tajima's D and Fay and Wu's H reveals that a region including the Vkorc1 gene is highly enriched with rare variants (Fig. 7.5). Plotting the Relate marginal tree in the Vkorc1 gene reveals that one haplotype is highly diverged from the remaining French haplotypes and coalesces >400,000 YBP with the MRCA of all remaining mice (Figure 7.6). This haplotype persists over a remarkably long region of 4Mb, indicating that the introgressed haplotype may have arrived extremely recently in the French mouse. In Ref. [32], a principal component analysis revealed that this mouse haplotype clustered midway between *M. m. domesticus* and *M. spretus*. On the Relate marginal trees of this region, all singletons of the haplotype of interest have a derived genotype of 2 in *Mus Spretus*, confirming this earlier observation. It is known that there are four non-synonymous mutations close to Vkorc1 that confer resistance to Warfarin [146]. These are believed to originate in Mus Spretus from northern Africa and southern Spain and introgression of this haplotype into *M. m. domesticus* has previously been reported [146]. Remarkably, this French mouse sample carries all four non-synonymous mutations, while these mutations are absent in all other samples.

In Taiwanese mice, it is demonstrated in Ref. [32] that Brca2 gene has seven fixed non-synomymous mutations private to Taiwanese mice, which represents a significant enrichment despite Brca2 being a large gene. We find that one of these non-synonymous mutations is also the youngest fixed, private, non-synonymous

mutation genome-wide. Of any fixed mutations in Taiwanese mice, only 0.00012 are younger than this non-synonymous mutation. However, we note that young fixed mutations are not unexpected given the strong recent bottleneck in Taiwanese mice and therefore further analysis is required to quantify the evidence of positive selection at this locus. The potentially selected phenotype is unknown, however Brca2 is a famous tumour suppressor gene, most notably of breast cancer, and acts as a mediator in the machinery responsible for recombination [132].

In Indian mice, the Prolactin (Prl) gene, which plays an essential role in lactation in mammals [10], has one of the smallest H statistics, as well as a low D statistic (Fig. 7.5). This is the only gene with any fixed, private non-synonymous mutation in Indian mice [32]. It is also the only gene with fixed, private mutations on a branch longer than 1 million years. Plotting the corresponding marginal tree reveals that all Indian mice are highly diverged from French or Taiwanese mice, with their joint TMRCA exceeding 1.5 MYBP (Fig. 7.6). This is therefore strongly suggestive of introgression of into Indian mice. Current analysis suggests that the source group is not *Mus Spretus* and might be as diverged as *Mus Famulus* and *Mus Caroli*, but further investigation is needed to better characterise the origin of this haplotype.

In summary, we have identified evidence of introgression at Vkorc1 and Prl and evidence of a selective sweep at Brca2. Going forward, I plan to extend these analyses to a genome-wide scan of introgression and seletion. This should reveal to what extent the three regions highlighted here are unusual in a genome-wide comparison, which should particularly aid our interpretation of the Brca2 selection signal and may shed light onto whether the introgression signal at Prl is adaptive.

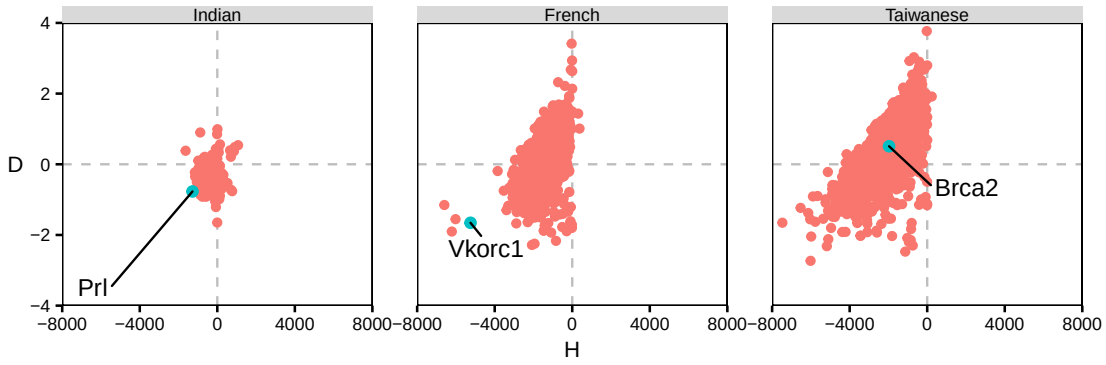


Figure 7.5: Tajima's D and Fay and Wu's H in wild mice.
Tajima's D plotted against Fay and Wu's H calculated for non-overlapping bins of 1Mb in Indian, French, and Taiwanese mice. In each population, we highlighted one gene with evidence for selection identified in Ref. [32] (see Section 7.5).

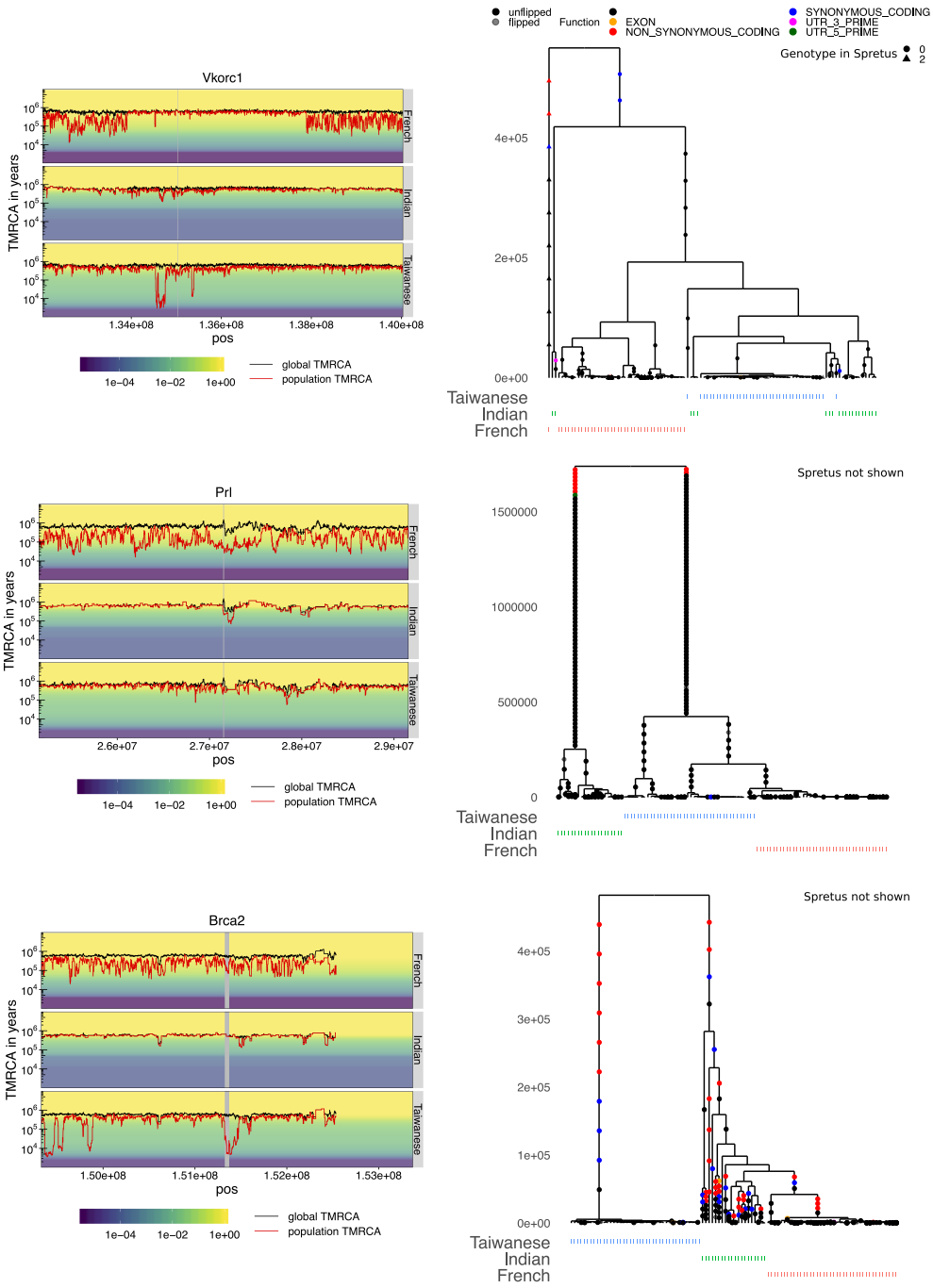


Figure 7.6: TMRCAs and marginal trees at three potential targets of positive selection in mice: Vkorc1, Prl, and Brca2.
Left column shows TMRCAs of all three populations (black) and each individual population (red). Background colour shows the cumulative distribution function of population-specific TMRCAs. The grey bar indicates the location of the gene of interest. The corresponding marginal trees are shown in the right column. At Vkorc1, one French haplotype is highly diverged from the remaining French haplotypes and coalesces >400,000 YBP, possibly due to introgression from Spretus. At Prl, all Indian mice are highly diverged from the remaining mouse groups, again possibly a sign of an introgressed haplotype. At Brca2, all Taiwanese mice appear to coalesce rapidly, which could be signature of a more conventional selective sweep.

8

Discussion and future work

Contents

In this thesis, I developed Relate, a scalable method for estimating genealogies genome-wide and demonstrated its utility on a diverse set of applications. We applied Relate to 2478 modern humans and 50 wild mice. In these applications, results suggest that we have improved in accuracy, resolution, or statistical power on state-of-the-art methods. A strength of a genealogy-based analysis is that all our inferences are derived from the same genealogy, making results across different applications easier to compare. I believe that Relate is the first step towards a broader framework for conducting statistical and population genetic inferences by first estimating genealogy trees and then leveraging these to address almost any statistical or population genetic question of interest.

In the remainder of this chapter, I will outline possible extensions of the Relate framework and potentially fruitful applications of genealogies. The development of methods working with a set of inferred trees is particularly interesting because

genealogy-based inference is highly modular; methods should be applicable regardless of the specific algorithm used for estimating marginal trees.

8.1 Improvements to Relate

With genealogy-based inference proving to be powerful, I am interested in extending Relate into a framework that is easy to use and efficient and integrates well with infrastructure set up by other software, such as tskit (`https://github.com/tskit-dev/tskit`). An essential step towards this goal is to make the Relate output compatible with the succinct tree sequence file format developed for tskit [80], a file format that, in principle, achieves remarkable compression of haplotype sequencing data using genealogies. Compressing data using genealogies has the benefit of simultaneously organising the data in a biologically meaningful way, allowing for fast computation of many population genetic statistics and powerful genealogy-based inferences. Moreover, to facilitate development of novel tree-based techniques by third party individuals, my aim is to provide a minimal working library in C++, as well as higher level programming languages, that enable simple operations on inferred trees (such as input, output, and extracting embedded trees).

There are several natural extensions to Relate itself. A more efficient version of Relate will be required to make use of the increasingly large data sets becoming available. Relate, which is already many orders of magnitude faster than existing methods, scales quadratically in the number of samples for both runtime and memory usage, making it infeasible for data sets comprising more than 10,000 haplotypes. With data sets of more than 100,000 haplotypes already being assembled [49], it is necessary to think more carefully about the computational resources available to us.

It would be desirable to relax some of the assumptions that Relate imposes on its input data. For instance, Relate requires phased haplotypes, as well as knowledge of ancestral alleles. Both requirements are difficult to meet for most species excluding

humans. In principle, genealogical trees contain rich information about likely phases of genotype data, because they indicate which mutations can or cannot co-occur. It should therefore be possible to phase genotypes using genealogical trees and this could potentially be extended to inference of genealogies from unphased sequences.

Another requirement in the current implementation of Relate is that all DNA sequences are sampled in the present. However, in some fast evolving species, such as some bacteria, sample ages can be spread over many generations. For humans, there are now hundreds of ancient DNA (aDNA) samples available (e.g., Ref. [95]). When incorporated correctly, aDNA should drastically improve the resolution of a genealogy in the ancient past. In addition, these ancient samples reveal important insights into the genetic diversity of the past (see e.g., Refs. [45, 141]). However, the inclusion of aDNA poses several challenges. In Relate, it is currently not trivial to fix the age of a sample to a time before present. In addition, aDNA may have substantially higher error rates or more missing data than modern-day individuals, requiring potentially an approach that "threads" sequences through genealogies that are initially built using modern individuals.

8.2 Richer inference framework for detecting natural selection

In Chapter 6 of this thesis, we have developed new tree-based statistics for detecting positive selection, which we have applied to study natural selection either on individual mutations or collections of mutations. We regard the selection statistics introduced here as initial approaches along a path towards a richer inference framework, including e.g. background selection, full selective sweeps, or balancing selection.

A number of statistics for detecting selection evidence using genealogical trees have recently been developed (see Chapter 6 for a discussion). While some of the

strongest selection signals, including the LCT signal, are detected by all methods, reproducibility of signals is otherwise low across different methods, partially owing to the fact that not all methods are applied to the same data set and investigate slightly different modes of selection. I am planning to conduct an extensive simulation study to truly evaluate and compare the performance of these methods which should also reveal limitations - for instance, we suspect that admixture is a likely confounder for our selection statistic.

An alternative and potentially powerful idea for studying selection is the concept of branching rates of lineages, which we could use to infer selection coefficients changing through time and construct statistics robust to population structure and admixture.

8.3 Tracking trait evolution through time

Genealogies reconstruct the history of observed alleles and contain information about historical allele frequencies. Recently, several novel approaches have been developed to infer allele frequency trajectories in non-neutral scenarios, where the inferred genealogical history is biased towards oversampling of beneficial lineages [35, 149].

Using estimated historical allele frequencies of mutations associated with a trait, we can, in theory, calculate polygenic scores through time and predict historical phenotype changes indicative of selection [35]. This approach should aid interpretability of the complex picture of directional adaptation observed in Chapter 6, by disentangling factors such as temporal changes in selection strength and positive/negative selection. Another interesting direction of research would be extrapolating past polygenic scores to the future, particularly for traits with ongoing selection.

8.4 Conditional coalescence rates

In Chapter 3, we developed a maximum-likelihood estimator for coalescence rates of pairs of sequences, adapting previous work by Dr Marie Forest [43]. By averaging over sequences in the same or different groups, we obtained within and across group coalescence rates, which we were able to use to detect population stratification and date split times.

We can extend these ideas by annotating branches and calculating coalescence rates for branches with same or different annotations. For instance, we may be able to detect ancient substructure and/or introgression events by using the lower-end time of a branch as its annotation. In panmictic scenarios, the coalescence rate should not depend on the lower-end time of a branch. However, if the lower-end time is younger than some introgression event, then the coalescence rate is expected to decrease going back in time, indicating that some branches belong to a separate subspecies that is unable to coalesce with the modern human lineage for long time periods, while increasing in the ancient past when this subspecies coalesces with the modern human lineage. Contrary to most existing approaches that heavily rely on sampling of ancient human subspecies to detect introgression, this approach would work without prior knowledge about such extinct lineages. For instance, in Chapter 5, we identified extensive unexplained ancient structure among African populations, likely involving an unknown ancient hominid not closely related to Neanderthals or Denisovans. Using this and similar approaches, we plan to characterise this event by establishing whether it involves a single or multiple unknown human subspecies and dating their departure from the modern human lineage, as well as the time of introgression.

Another application of conditional coalescence rates is the detection of directional migration events, in which gene flow occurred from one group A to another group B over some sustained time span but not conversely. To detect such events, we annotate lineages by whether their lower-ends are the MRCAs of sequences in groups A, B, or a mixture of A and B ((A,A), (B,B), or (A,B)). We can then calculate the coalescence

rate with sequences in groups A and B. In a scenario with directional migration from group A towards group B, we expect an asymmetry in these coalescence rates with rates for ((A,B),A) appearing more similar to ((A,A),A) than ((B,B),A) and rates for ((A,B),B) appearing more similar to ((A,A),B) than ((B,B),B), whereas a perfectly symmetric split of groups A and B would not have such asymmetries.

8.5 Resimulating the genetic history of a sample

Given a genealogy, we can think about resimulating the genetic history of a sample by removing all existing mutations, and subsequently reintroducing new mutations at random. This would generate a new simulated data set that preserves many properties, such as population size changes, migrations, and other historical events. Importantly, this approach assumes independence of genealogical and mutation processes which are not valid in non-neutral settings. This approach could therefore potentially also serve as a test of mutational independence.

Whether such an approach can work, or whether biases and inaccuracies in the genealogy will complicate this idea needs to be investigated. An interesting idea would be to use the resimulated mutation patterns to assess the quality of a genealogy. For instance, one can think of comparing summary statistics of the new simulated variation patterns to those of the original data.

8.6 Recombination rate changes through time

Recombination hotspots are determined by binding sites of PRDM9 which is a zinc finger protein that binds to motifs specified by its zinc finger array [105]. PRDM9 has been shown to evolve remarkably rapidly [107] which has resulted in partially different motifs across modern human groups, such as in African and

European individuals [70]. Inferred genealogies may enable us to date recombination events and then track how recombination hotspots have changed through time, potentially revealing historical motifs of PRDM9.

8.7 Adapting the Relate framework to bacteria

In collaboration with Dr Daniel Falush (Univ. Bath), I am working on adapting the Relate framework to bacterial species. The evolutionary forces acting on bacteria genomes differ substantially from those governing mammalian evolution, as bacteria reproduce clonally and sporadically exchange genetic material across organisms. By developing a genealogy estimation framework for bacteria, we can transfer the advances that have been made with human applications in mind to microbiology, quantifying, for instance, the evolutionary trajectory, including evidence for natural selection, of phenotypes affecting human health.

Appendices

A
Tables

Table A.1: Number of 1000 Genomes Project samples used in our analysis by population label.

ACB	ASW	BEB	CDX	CEU	CHB	CHS	CLM	ESN	FIN	GBR	GIH	GWD
95	60	85	92	98	102	104	93	98	98	90	102	112
IBS	ITU	JPT	KHV	LWK	MSL	MXL	PEL	PJL	PUR	STU	TSI	YRI
106	101	103	98	98	84	63	84	95	103	101	106	107

AMR	(Americas)
CLM	Colombian in Medellin, Colombia
MXL	Mexican Ancestry in Los Angeles, CA, USA
PEL	Peruvian in Lima, Peru
PUR	Puerto Rican in Puerto Rico

SAS	(Southern Asians)
BEB	Bengali in Bangladesh
GIH	Gujarati Indian in Houston, TX, USA
ITU	Indian Telugu in the UK
PJL	Punjabi in Lahore, Pakistan
STU	Sri Lankan Tamil in the UK

EAS	(East Asians)
CDX	Chinese Dai in Xishuangbanna, China
CHB	Han Chinese in Bejing, China
CHS	Southern Han Chinese, China
JPT	Japanese in Tokyo, Japan
KHV	Kinh in Ho Chi Minh City, Vietnam

EUR	(Europeans)
CEU	Utah residents with Northern and Western European ancestry
IBS	Iberian populations in Spain
FIN	Finnish in Finland
GBR	British in England and Scotland
TSI	Toscani in Italy

AFR	(Africans)
ACB	African Caribbean in Barbados
ASW	African Ancestry in Southwest US
ESN	Esan in Nigeria
GWD	Gambian in Western Division, The Gambia
LWK	Luhya in Webuye, Kenya
MSL	Mende in Sierra Leone
YRI	Yoruba in Ibadan, Nigeria

Table A.2: Genome-wide significant hits for positive selection.
Regions containing a SNP with p-value for selection evidence of less than 5×10^{-8} in at least three populations. We list genes, eQTLs, and GWAS at mutations with $r^2 \geq 0.5$, where a bold font corresponds to $r^2 = 1.0$, a plain font corresponds to $r^2 \geq 0.8$, and brackets correspond to $r^2 \geq 0.5$. If a non-synonymous mutation falls within $r^2 \geq 0.5$, we indicate this similarly in the NSM column. BP denotes the base-pair position of the SNP (GRCh37). In the notes column, we indicate whether this region has been highlighted in a previous study. Additionally, we record the statistic that attains a significant value according to the PopHumanScan resource [120], whenever the region reported in PopHumanScan overlaps the region listed in this table and is attributed to a population listed in this table. See Section 6.4.2 for details

GWAS catalogue phenotypes:

a Helix rolling

b Cholesterol, total+;Blood metabolite levels-

c Nonsyndromic cleft lip with cleft palate

d Nonsyndromic cleft lip with cleft palate

e Glaucoma (primary open-angle)

UK BIOBANK phenotypes:

1. Forced vital capacity (FVC)-;Monocyte percentage+;Place of birth in UK - east co-ordinate-;Place of birth in UK - north co-ordinate+

2. Arm fat mass (left)+;Arm fat mass (right)+;Body fat percentage+;Forced vital capacity (FVC)-;Leg fat mass (left)+;Leg fat mass (right)+;Monocyte percentage+;Place of birth in UK - east co-ordinate-;Place of birth in UK - north co-ordinate+;Trunk fat mass+;Trunk fat percentage+;Whole body fat mass+

3. General happiness with own health+;High light scatter reticulocyte count-;High light scatter reticulocyte percentage-;Immature reticulocyte fraction-;Impedance of arm (right)+;Impedance of leg (left)+;Impedance of leg (right)+;Impedance of whole body+;Reticulocyte count-;Reticulocyte percentage-;Standing height+

4. Standing height-

5. White blood cell (leukocyte) count-

6. Impedance of leg (right)-;Impedance of whole body-;Red blood cell (erythrocyte) distribution width+

7. Leg fat-free mass (right)+;Leg predicted mass (right)+;Red blood cell (erythrocyte) distribution width+;Trunk fat-free mass+;Trunk predicted mass+;Whole body fat-free mass+;Whole body water mass+

8. Birth weight-;High light scatter reticulocyte count+;High light scatter reticulocyte percentage+;Nervous feelings+;Reticulocyte count+;Reticulocyte percentage+;Systolic blood pressure, automated reading+

9. Hair colour (natural, before greying): Blonde-;Hair colour (natural, before greying): Dark brown+

10. Hair colour (natural, before greying): Blonde-;Hair colour (natural, before greying): Dark brown+

11. 3mm strong meridian (left)+;3mm weak meridian (left)+;6mm strong meridian (left)+;6mm weak meridian (left)+;Age high blood pressure diagnosed-;Arm fat-free mass (left)-;Arm fat-free mass (right)-;Arm predicted mass (left)-;Arm predicted mass (right)-;Basal metabolic rate-;Basophill count+;Birth weight of first child-;Birth weight-;Blood clot, DVT, bronchitis, emphysema, asthma, rhinitis, eczema, allergy diagnosed by doctor: Hayfever, allergic rhinitis or eczema-;Blood clot, DVT, bronchitis, emphysema, asthma, rhinitis, eczema, allergy diagnosed by doctor: None of the above+;Comparative height size at age 10-;Coronary atherosclerosis+;Diagnoses - main ICD10: I21 Acute myocardial infarction+;Diastolic blood pressure, automated reading+;Diseases of the circulatory system+;Duration to first press of snap-button in each round+;Eosinophill count+;Eosinophill percentage+;Ever smoked+;Haematocrit percentage+;Haemoglobin concentration+;High light scatter reticulocyte count+;High light scatter reticulocyte percentage+;Hip circumference-;Illnesses of father: Heart disease+;Illnesses of siblings: High blood pressure+;Illnesses of siblings: None of the above (group 1)-;Immature reticulocyte fraction+;Impedance of arm (left)+;Impedance of arm (right)+;Impedance of leg (right)+;Impedance of whole body+;Ischaemic heart disease, wide definition+;Leg fat-free mass (left)-;Leg fat-free mass (right)-;Leg predicted mass (left)-;Leg predicted mass (right)-;Long-standing illness, disability or infirmity+;Lymphocyte count+;Lymphocyte percentage+;Major coronary heart disease event excluding revascularizations+;Major coronary heart disease event+;Mean corpuscular haemoglobin+;Mean corpuscular volume+;Mean sphered cell volume-;Mean time to correctly identify matches+;Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Blood pressure medication+;Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: None of the above-;Medication for cholesterol, blood pressure or diabetes: Blood pressure medication+;Medication for cholesterol, blood pressure or diabetes: None of the above-;Monocyte count+;Myocardial infarction+;Myocardial infarction, strict+;Neutrophill count+;Neutrophill percentage-;Non-cancer illness code, self-reported: hypertension+;Non-cancer illness code, self-reported: psoriasis+;Number of self-reported non-cancer illnesses+;Past tobacco smoking-;Platelet count+;Platelet crit+;Platelet distribution width+;Red blood cell (erythrocyte) count+;Reticulocyte count+;Reticulocyte percentage+;Smoking status: Never-;Smoking status: Previous+;Standing height-;Systolic blood pressure, automated reading+;Taking other prescription medications+;Trunk fat-free mass-;Trunk predicted mass-;Weight-;White blood cell (leukocyte) count+;Whole body fat-free mass-;Whole body water mass-

12. Comparative body size at age 10-;Impedance of leg (left)+;Impedance of leg (right)+;Lymphocyte percentage-;Monocyte count+;Monocyte percentage+;Neutrophill count+;White blood cell (leukocyte) count+

13. Arm fat mass (left)-;Arm fat mass (right)-;Arm fat percentage (right)-;Basal metabolic rate-;Body mass index (BMI)-;Comparative body size at age 10-;Duration to first press of snap-button in each round+;Eosinophill percentage+;Haematocrit percentage-;Haemoglobin concentration-;High light scatter reticulocyte count-;High light scatter reticulocyte percentage-;Impedance of leg (left)+;Leg fat-free mass (left)-;Leg fat-free mass (right)-;Leg fat mass (left)-;Leg fat mass (right)-;Leg predicted mass (left)-;Leg predicted mass (right)-;Lymphocyte percentage+;Mean corpuscular haemoglobin+;Mean corpuscular volume+;Mean platelet (thrombocyte) volume+;Mean reticulocyte volume+;Mean sphered cell volume+;Mean time to correctly identify matches+;Nap during day+;Neutrophill count-;Neutrophill percentage-;Platelet count-;Red blood cell (erythrocyte) count-;Red blood cell (erythrocyte) distribution width-;Reticulocyte count-;Waist circumference-;Weight-;Whole body fat mass-

14. Arm fat mass (right)-;Arm fat percentage (right)-;Body mass index (BMI)-;Comparative body size at age 10-;Duration to first press of snap-button in each round+;Eosinophill

percentage+;Haematocrit percentage-;Haemoglobin concentration-;High light scatter reticulocyte count-;High light scatter reticulocyte percentage-;Lymphocyte percentage+;Mean corpuscular haemoglobin+;Mean corpuscular volume+;Mean platelet (thrombocyte) volume+;Mean reticulocyte volume+;Mean sphered cell volume+;Mean time to correctly identify matches+;Nap during day+;Neutrophill count-;Neutrophill percentage-;Platelet count-;Red blood cell (erythrocyte) count-;Red blood cell (erythrocyte) distribution width-;Reticulocyte count-;Reticulocyte percentage-;Sitting height+;Waist circumference-;White blood cell (leukocyte) count-

15. Mean platelet (thrombocyte) volume-;Platelet count+;Platelet distribution width-

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
1:76.1-76.4	BP76111796	KHV (EAS)	6.8×10^{-10}	SLC44A5	(MSH4)			0.03	0.21	0.39	0.63	[160]; Fay & Wu's H
		ITU (SAS)	1.2×10^{-8}									
		STU (SAS)	3.5×10^{-9}									
1:236.5-236.6	rs76420343	FIN (EUR)	1×10^{-9}	**EDARADD**	EDARADD			0.08	0.52	0.38	0.22	The EDARADD protein is known to directly interact with the EDAR protein [134].
		BEB (SAS)	1.4×10^{-8}									
		GIH (SAS)	2.6×10^{-9}									
		ITU (SAS)	4.2×10^{-8}									
		PJL (SAS)	4.8×10^{-8}									
	rs79881482	STU (SAS)	3×10^{-9}	**EDARADD**	EDARADD			0.06	0.52	0.38	0.22	
2:108.9-109.6	rs11123695	CDX (EAS)	7.4×10^{-10}	**GCC2**	(GCC2)	a	EDAR	0	0.01	0.01	0.85	[133, 160]; iHS, XPEHH
		CHB (EAS)	1.2×10^{-12}									
		CHS (EAS)	5.7×10^{-12}									
		KHV (EAS)	8×10^{-12}									
2:135.6-136.9	rs6730157	FIN (EUR)	1.4×10^{-8}	**RAB3GAP1**	**MCM6**	**1**,b	(RAB3GAP1)	0	0.49	0.12	0	[23, 40, 95, 117, 133, 160]
	rs1375131	GBR (EUR)	4.3×10^{-10}	**ZRANB3**	**MCM6**	**1**,b	(RAB3GAP1)	0	0.49	0.12	0	iHS, XPEHH
	rs56369224	CEU (EUR)	3.3×10^{-9}	**R3HDM1**	MCM6	**2**,b	(RAB3GAP1)	0	0.49	0.12	0	Fu & Li's D, α
2:168.4-168.5	rs150960584	CEU (EUR)	1.7×10^{-9}	**U7**				0.02	0.67	0.2	0.04	–
		GBR (EUR)	2.8×10^{-8}									
		IBS (EUR)	7.3×10^{-9}									
		TSI (EUR)	2.5×10^{-9}									
3:48.7-50.5	rs139083518	KHV (EAS)	3.6×10^{-9}	**DAG1**	(NAT6)			0	0.01	0.02	0.6	[23, 160]
	rs201632611	CDX (EAS)	3×10^{-8}	**GNAI2**	NAT6	(3)	(C3orf45)	0	0.01	0.01	0.56	
		CHS (EAS)	1.3×10^{-11}									
4:107.6-107.8	rs1364808	ESN (AFR)	3.1×10^{-8}	**DKK2**				0.86	0.87	0.95	1	[160];
	rs817146	MSL (AFR)	9.7×10^{-9}	**DKK2**				0.93	0.87	0.95	1	Fu & Li's D,
	rs704049	LWK (AFR)	5×10^{-10}	**DKK2**				0.92	0.87	0.95	1	Fu & Li's F
	rs3111706	GWD (AFR)	1×10^{-8}	**DKK2**				0.93	0.87	0.96	1	

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
4:107.8-108	rs6829139	YRI (AFR)	7.8×10^{-10}	**DKK2**				0.9	0.97	0.99	1	in close physical
	rs17037205	LWK (AFR)	1.3×10^{-8}	**DKK2**				0.9	0.97	0.99	1	proximity with
		MSL (AFR)	8.7×10^{-9}									previous region
5:65.2-65.3	rs59755544	GIH (SAS)	1.7×10^{-8}	**ERBB2IP**				0.44	0.85	0.86	0.77	–
		PJL (SAS)	9.3×10^{-10}									
		STU (SAS)	3.5×10^{-8}									
5:178.2-178.3	BP178258803	TSI (EUR)	2.5×10^{-9}	RP11-281O15.3	**AACSP1**			0.13	0.28	0.32	0.25	Fu & Li's D
		BEB (SAS)	1.3×10^{-9}									
		ITU (SAS)	1.7×10^{-9}									
7:19.5-19.6	rs71530658	CEU (EUR)	1.4×10^{-8}	**AC007091.1**		(4)		0.03	0.42	0.15	0.01	F_{ST}
		GBR (EUR)	1.3×10^{-10}									
		IBS (EUR)	2.8×10^{-8}									
		TSI (EUR)	2.8×10^{-9}									
7:98.9-99.1	BP98971118	GWD (AFR)	1.2×10^{-12}	ARPC1A	ARPC1B	(5)		0.5	0.07	0.06	0	–
	rs10229886	ESN (AFR)	9.2×10^{-9}	**ARPC1A**	(ARPC1B)	(5)		0.37	0.03	0.02	0	
	BP98971173	YRI (AFR)	1.1×10^{-10}	ARPC1A	(ARPC1B)	(5)		0.37	0.03	0.02	0	
	BP98971260	LWK (AFR)	2.1×10^{-13}	(ARPC1A)	(GS1-259H13.2)	(5)		0.37	0.03	0.04	0	
8:38-38.3	rs59911155	ESN (AFR)	4.3×10^{-8}	**LSM1**	**DDHD2**	**6**,(c)	(DDHD2)	0.75	0.74	0.91	0.68	–
	rs3739252	GWD (AFR)	3.4×10^{-8}	**DDHD2**	**DDHD2**	**7**,d	(DDHD2)	0.78	0.74	0.91	0.68	
		LWK (AFR)	4.2×10^{-9}									
9:99.4-99.5	BP99411763	YRI (AFR)	4.2×10^{-8}					0.29	0	0	0	–
	BP99411801	GWD (AFR)	5.4×10^{-9}					0.33	0	0	0	
		MSL (AFR)	1.2×10^{-9}									
9:107.3-107.4	rs112315552	ESN (AFR)	3.8×10^{-8}	**OR13C5**	NIPSNAP3A		OR13C2	0.6	0.16	0.38	0.55	–
		GWD (AFR)	2.4×10^{-10}									
		MSL (AFR)	2.1×10^{-9}									
		YRI (AFR)	2.3×10^{-9}									

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
10:0.4-0.6	rs77347335	ESN (AFR)	2.4×10^{-10}	**DIP2C**				0.43	0.01	0.03	0.02	–
		GWD (AFR)	2.5×10^{-8}									
		LWK (AFR)	1.1×10^{-11}									
		MSL (AFR)	4×10^{-8}									
10:104.6-105	rs11191469	LWK (AFR)	8.4×10^{-9}	**CNNM2**	C10orf32	**8**		0.66	0.38	0.25	0.27	[23]
		MSL (AFR)	6.8×10^{-10}									
		YRI (AFR)	5.6×10^{-9}									
10:117.2-117.5	rs150028049	CHB (EAS)	4.3×10^{-8}	**ATRNL1**				0.1	0.23	0.34	0.62	F_{ST}
		CHS (EAS)	1.9×10^{-12}									
		KHV (EAS)	1.2×10^{-8}									
10:122.8-123	rs2246730	IBS (EUR)	3.8×10^{-8}	**RP11-159H3.2**				0.39	0.97	0.96	0.65	Fay & Wu's H
	rs1873446	ITU (SAS)	1.7×10^{-8}	**RP11-159H3.2**				0.37	0.97	0.95	0.64	
	rs10886862	STU (SAS)	4.8×10^{-8}	**RP11-159H3.2**				0.42	0.97	0.95	0.61	
11:65.1-65.2	rs188162087	BEB (SAS)	1.1×10^{-8}	**DPF2**	(AP003068.18)			0	0.29	0.24	0.12	–
		GIH (SAS)	1.7×10^{-11}									
		PJL (SAS)	1.1×10^{-8}									
11:91.8-92	rs11019805	CHS (EAS)	3.7×10^{-8}	**FAT3**	(FAT3)			0.08	0.31	0.54	0.59	[117] for GBR
		BEB (SAS)	6.5×10^{-9}									
		ITU (SAS)	4.1×10^{-8}									
12:79.7-80.2	rs10778678	LWK (AFR)	3.6×10^{-9}	**RP11-359M6.1**	**RP11-530C5.2**			0.72	0.01	0.08	0.21	[133, 160]; iHS
	rs12316084	YRI (AFR)	7.3×10^{-9}	**PAWR**	RP11-530C5.2			0.54	0	0	0.01	
	rs7306681	MSL (AFR)	2.5×10^{-8}	**PAWR**	**RP11-530C5.2**			0.59	0	0	0.01	
12:83-83.1	rs11115333	ESN (AFR)	3.9×10^{-10}	**TMTC2**				0.8	0.77	0.81	0.71	–
		GWD (AFR)	9.2×10^{-10}									
		LWK (AFR)	6.9×10^{-9}									
12:87.3-87.4	rs11104181	GWD (AFR)	4×10^{-8}	**RP11-202H2.1**				0.87	0.43	0.48	0.77	[23] (KITLG for CEU)
	rs11503304	ESN (AFR)	4.8×10^{-9}	**RP11-202H2.1**		(9)		0.82	0.66	0.54	0.78	
	rs2406741	LWK (AFR)	1.8×10^{-10}	**RP11-202H2.1**		(9)		0.85	0.66	0.54	0.78	
	rs7309012	MSL (AFR)	7.4×10^{-9}	**RP11-202H2.1**		**10**		0.84	0.67	0.52	0.61	

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
12:111.7-113	rs7137828	GBR (EUR)	3.3×10^{-11}	**ATXN2**	ALDH2	**11**,**e**	SH2B3	0	0.46	0.07	0	[95]; iHS
		IBS (EUR)	1.2×10^{-9}									
		TSI (EUR)	9.9×10^{-11}									
13:28.6-28.7	rs9554250	CDX (EAS)	7.1×10^{-10}	**FLT3**	**FLT3**	**12**		0.28	0.55	0.45	0.7	–
		CHB (EAS)	1.9×10^{-10}									
		CHS (EAS)	9.2×10^{-11}									
		KHV (EAS)	4.1×10^{-10}									
14:32.9-33	rs7153204	MSL (AFR)	5.6×10^{-9}	**AKAP6**	(AKAP6)			0.88	0.08	0.12	0.29	–
	rs10138310	LWK (AFR)	1.5×10^{-8}	**AKAP6**	(AKAP6)			0.86	0.1	0.21	0.3	
	rs11628486	GWD (AFR)	1.9×10^{-8}	**AKAP6**	(AKAP6)			0.86	0.09	0.2	0.26	
		YRI (AFR)	1.4×10^{-8}									
16:22.9-23.1	rs16974808	YRI (AFR)	4.8×10^{-9}	**HS3ST2**				0.79	0.06	0.11	0.01	[160]
	rs1604799	ESN (AFR)	3.1×10^{-8}	**HS3ST2**				0.79	0.06	0.11	0.01	
	rs8063811	MSL (AFR)	9.5×10^{-9}	**RP11-20G6.2**				0.79	0.06	0.11	0.01	
16:59.5-59.7	rs9929021	ESN (AFR)	3.1×10^{-8}	**U4**				0.76	0.32	0.46	0.29	–
	rs9937266	GWD (AFR)	2.1×10^{-8}	**U4**				0.76	0.35	0.45	0.32	
		YRI (AFR)	4.2×10^{-8}									
16:81.4-81.5	rs310010	ESN (AFR)	2.3×10^{-8}	**CMIP**				0.5	0.06	0.13	0.1	–
		GWD (AFR)	3.3×10^{-8}									
		MSL (AFR)	5.1×10^{-10}									
		YRI (AFR)	6.3×10^{-10}									
17:44-44.9	BP44262496	STU (SAS)	2.6×10^{-8}	KANSL1	**RP11-798G7.5**	**13**	(KANSL1)	0.03	0.31	0.6	0.02	[117] (MYL4 in GBR);
	BP44363740	ITU (SAS)	1.9×10^{-9}	(KANSL1)	**RP11-798G7.5**	**14**		0.03	0.31	0.63	0.04	Fay & Wu's H
		PJL (SAS)	2.4×10^{-10}									
18:37.7-37.8	rs7236000	GWD (AFR)	1.1×10^{-9}	**RP11-653G8.2**				0.62	0.15	0.26	0.62	–
		MSL (AFR)	5.7×10^{-10}									
	rs1943603	ESN (AFR)	4.5×10^{-9}	**RP11-653G8.2**				0.63	0.15	0.27	0.59	

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
19:31.7-31.8	BP31733665	LWK (AFR)	3.3×10^{-11}	TSHZ3				0.44	0.13	0.2	0.18	–
	rs62101246	ESN (AFR)	7.4×10^{-10}	**TSHZ3**				0.47	0.14	0.24	0.19	
		MSL (AFR)	3.2×10^{-9}									
19:45.8-45.9	rs12609631	MSL (AFR)	2.9×10^{-8}	**KLC3**		**15**		0.71	0.28	0.21	0.35	[133]
	rs10853773	ESN (AFR)	3.5×10^{-8}	**KLC3**		**15**		0.72	0.27	0.21	0.34	
		YRI (AFR)	3.5×10^{-8}									
21:17.5-17.7	rs2823681	GWD (AFR)	7×10^{-9}	**LINC00478**				0.47	0	0	0	–
		LWK (AFR)	3.4×10^{-10}									
		MSL (AFR)	2.8×10^{-8}									
		YRI (AFR)	4×10^{-8}									
22:23-23.2	rs2003444	CEU (EUR)	6.2×10^{-9}	**IGLV3-16**	**IGLV3-12**			0.49	0.62	0.65	0.64	Fu & Li's D
		IBS (EUR)	1.8×10^{-8}									
		TSI (EUR)	2×10^{-9}									
		BEB (SAS)	8.3×10^{-10}									
		GIH (SAS)	3.8×10^{-10}									
		ITU (SAS)	6.6×10^{-11}									
		PJL (SAS)	9.4×10^{-9}									
		STU (SAS)	1.4×10^{-9}									

Bibliography

[1] 1000 Genomes Project Consortium, A global reference for human genetic variation, *Nature*, 526:68–74, 2015.

[2] 1000 Genomes Project dataset, phased; accessed 13 Jan 2017, `https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html`.

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

[4] L. Anderson-Trocmé et al., Legacy Data Confounds Genomics Studies, *bioRxiv:624908,* 2019.

[5] A. Auton and G. A. T. McVean, Recombination rate estimation in the presence of hotspots, *Genome research*, 17:1219–1227, 2007.

[6] C. J. Bae, K. Douka, and M. D. Petraglia, On the origin of modern humans: Asian perspectives, *Science*, 358:eaai9067, 2017.

[7] M Bahlo and RC Griffiths, Inference from gene trees in a subdivided population, *Theoretical Population Biology*, 57:79–95, 2000.

[8] F. Baudat et al., PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice, *Science*, 327:836–840, 2010.

[9] A. C. Beichman, T. N. Phung, and K. E. Lohmueller, Comparison of single genome and allele frequency data reveals discordant demographic histories, *G3: Genes, Genomes, Genetics*, 7:3605–3620, 2017.

[10] N. Ben-Jonathan, C. R. LaPensee, and E. W. LaPensee, What can we learn from rodents about prolactin in humans?, *Endocrine Reviews*, 29:1–41, 2007.

[11] J. J. Berg et al., Reduced signal for polygenic adaptation of height in UK Biobank, *ELife*, 8:e39725, 2019.

[12] J. J. Berg and G. Coop, A population genetic signal of polygenic adaptation, *PLOS Genetics*, 10:e1004412, 2014.

[13] T. Bersaglieri et al., Genetic signatures of strong recent positive selection at the lactase gene, *The American Journal of Human Genetics*, 74:1111–1120, 2004.

[14] A. Bird, DNA methylation patterns and epigenetic memory, *Genes & development*, 16:6–21, 2002.

[15] M. Bordewich and C. Semple, On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance, *Annals of Combinatorics*, 8:409–423, 2005.

[16] S. R. Browning and B. L. Browning, Haplotype phasing: existing methods and new developments, *Nature Reviews Genetics*, 12:703–714, 2011.

[17] J. Bryk et al., Positive selection in East Asians for an EDAR allele that enhances NF-κB activation, *PLOS One*, 3:e2209, 2008.

[18] B. K. Bulik-Sullivan et al., LD Score regression distinguishes confounding from polygenicity in genome-wide association studies, *Nature Genetics*, 47:291–295, 2015.

[19] C. Bycroft et al., The UK Biobank resource with deep phenotyping and genomic data, *Nature*, 562:203–209, 2018.

[20] R. L. Cann, M. Stoneking, and A. C. Wilson, Mitochondrial DNA and human evolution, *Nature*, 325:31–36, 1987.

[21] J. Carlson et al., Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans, *Nature Communications*, 9:3753, 2018.

[22] A. M. Casto and M. W. Feldman, Genome-Wide Association Study SNPs in the Human Genome Diversity Project populations: Does Selection Affect Unlinked SNPs with Shared Trait Associations?, *PLOS Genetics*, 7:e1001266, 2011.

[23] H. Chen, J. Hey, and M. Slatkin, A hidden Markov model for investigating recent positive selection through haplotype structure, *Theoretical Population Biology*, 99:18–30, 2015.

[24] R. Chen et al., Type 2 diabetes risk alleles demonstrate extreme directional differentiation among human populations, compared to other diseases, *PLOS Genetics*, 8:e1002621, 2012.

[25] C. Cheng et al., African ancestry and its correlation to type 2 diabetes in African Americans: a genetic admixture analysis in three US population cohorts, *PLOS One*, 7:e32840, 2012.

[26] E. Y. Cheng et al., Meiotic Recombination in Human Oocytes, *PLOS Genetics*, 5:1–15, 2009.

[27] H. Chheda et al., Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom, *European Journal of Human Genetics*, 25:477–484, 2017.

[28] 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, *Nature*, 467:1061–1073, 2010.

[29] Mouse Genome Sequencing Consortium et al., Initial sequencing and comparative analysis of the mouse genome, *Nature*, 420:520–562, 2002.

[30] G. Coop and R. C. Griffiths, Ancestral inference on gene trees under selection, *Theoretical Population Biology*, 66:219–232, 2004.

[31] B. Crespi, K. Summers, and S. Dorus, Adaptive evolution of genes underlying schizophrenia, *Proceedings of the Royal Society of London B: Biological Sciences*, 274:2801–2810, 2007.

[32] R. W. Davies, Factors influencing genetic variation in wild mice, PhD thesis, University of Oxford, 2015.

[33] O. Delaneau, J. Marchini, and J.-F. Zagury, A linear complexity phasing method for thousands of genomes, *Nature Methods*, 9:179–181, 2012.

[34] L. Duret and N. Galtier, Biased gene conversion and the evolution of mammalian genomic landscapes, *Annual Review of Genomics and Human Genetics*, 10:285–311, 2009.

[35] M. D. Edge and G. Coop, Reconstructing the history of polygenic scores using coalescent trees, *Genetics*, 211:235–262, 2019.

[36] N. S. Enattah et al., Identification of a variant associated with adult-type hypolactasia, *Nature Genetics*, 30:233–237, 2002.

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

[38] J. C. Fay and C. Wu, Hitchhiking Under Positive Darwinian Selection, *Genetics*, 155:1405–1413, 2000.

[39] P. Fearnhead and P. Donnelly, Estimating recombination rates from population genetic data, *Genetics*, 159:1299–1318, 2001.

[40] Y. Field et al., Detection of human adaptation during the past 2000 years, *Science*, 354:760–764, 2016.

[41] R. A. Fisher, XXI. – On the Dominance Ratio, *Proceedings of the Royal Society of Edinburgh*, 42:321–341, 1923.

[42] A. Fledel-Alon et al., Broad-Scale Recombination Patterns Underlying Proper Disjunction in Humans, *PLOS Genetics*, 5:1–7, 2009.

[43] M. Forest, Simultaneous estimation of population size changes and splits times using importance sampling, PhD thesis, University of Oxford, 2014.

[44] Q. Fu et al., An early modern human from Romania with a recent Neanderthal ancestor, *Nature*, 524:216–219, 2015.

[45] Q. Fu et al., The genetic history of Ice Age Europe, *Nature*, 534:200–205, 2016.

[46] J. Galway-Witham and C. Stringer, How did Homo sapiens evolve?, *Science*, 360:1296–1298, 2018.

[47] Z. Gao et al., Overlooked roles of DNA damage and maternal age in generating human germline mutations, *Proceedings of the National Academy of Sciences of the USA*, 116:9491–9500, 2019.

[48] Genomic accessibility masks; accessed 20 July 2017, `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/`.

[49] Genomics England, The 100,000 Genomes Project Protocol v3, 2017.

[50] S. Glémin et al., Quantification of GC-biased gene conversion in the human genome, *Genome Research*, 25:1215–1228, 2015.

[51] R. E. Green et al., A Draft Sequence of the Neandertal Genome, *Science*, 328:710–722, 2010.

[52] R. C. Griffiths and P. Marjoram, Ancestral inference from samples of DNA sequences with recombination, *Journal of Computational Biology*, 3:479–502, 1996.

[53] R. C. Griffiths and P. Marjoram, IMA Volume on Mathematical Population Genetics, in: ed. by P. Donnelly and S. Tavarè, Berlin/Heidelberg/New York: Springer, 1996, chap. An ancestral recombination graph, pp. 257–270.

[54] R. C. Griffiths and S. Tavaré, The age of a mutation in a general coalescent tree, *Stochastic Models*, 14:273–295, 1998.

[55] GTEx eQTL; accessed 13 Jan 2019, `https://storage.googleapis.com/gtex_analysis_v7/single_tissue_eqtl_data/GTEx_Analysis_v7_eQTL.tar.gz`.

[56] J. L. Guénet and F. Bonhomme, Wild mice: an ever-increasing contribution to a popular mammalian model, *Trends in Genetics*, 19:24–31, 2003.

[57] D. Gusfield, Efficient algorithms for inferring evolutionary trees, *Networks*, 21:19–28, 1991.

[58] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data, *PLOS Genetics*, 5:e1000695, 2009.

[59] GWAS Catalog; accessed 9 Nov 2017, `https://www.ebi.ac.uk/gwas/api/search/downloads/full`.

[60] J. G. Hacia et al., Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays, *Nature Genetics*, 22:164–167, 1999.

[61] D. L. Halligan et al., Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents, *PLOS Genetics*, 9:1–14, Dec. 2013.

[62] D. L. Halligan et al., Evidence for Pervasive Adaptive Protein Evolution in Wild Mice, *PLOS Genetics*, 6:1–9, 2010.

[63] M. F. Hammer et al., Genetic evidence for archaic admixture in Africa, *Proceedings of the National Academy of Sciences of the USA*, 108:15123–15128, 2011.

[64] HaploReg; accessed 21 Oct 2017, `http://archive.broadinstitute.org/mammals/haploreg/data/haploreg_v4.0_20151021.vcf.gz`.

[65] K. Harris, Evidence for recent, population-specific evolution of the human mutation rate, *Proceedings of the National Academy of Sciences of the USA*, 112:3439–3444, 2015.

[66] K. Harris and J. K. Pritchard, Rapid evolution of the human mutation spectrum, *eLife*, 6:e24284, 2017.

[67] J. Hein, Reconstructing evolution of sequences subject to recombination using parsimony, *Mathematical Biosciences*, 98:185–200, 1990.

[68] G. Hellenthal et al., A genetic atlas of human admixture history, *Science*, 343:747–751, 2014.

[69] D. Henderson, S. (Joe) Zhu, and G. Lunter, Demographic inference using particle filters for continuous Markov jump processes, *bioRxiv:382218,* 2018.

[70] A. G. Hinch et al., The landscape of recombination in African Americans, *Nature*, 476:170–175, 2011.

[71] K. E. Holsinger and B. S. Weir, Genetics in geographically structured populations: defining, estimating and interpreting F_{st}, *Nature Reviews Genetics*, 10:639–650, 2009.

[72] R. E. Howes et al., The global distribution of the Duffy blood group, *Nature Communications*, 2:266, 2011.

[73] J.-J. Hublin, The origin of Neandertals, *Proceedings of the National Academy of Sciences of the USA*, 106:16022–16027, 2009.

[74] R. R. Hudson, Properties of a neutral allele model with intragenic recombination, *Theoretical Population Biology*, 23:183–201, 1983.

[75] R. R. Hudson, Two-locus sampling distributions and their application, *Genetics*, 159:1805–1817, 2001.

[76] R. R. Hudson and N. L. Kaplan, Statistical properties of the number of recombination events in the history of a sample of DNA sequences, *Genetics*, 111:147–164, 1985.

[77] Human ancestor genome; accessed 20 July 2017, `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/`.

[78] T. M. Keane et al., Mouse genomic variation and its effect on phenotypes and gene regulation, *Nature*, 477:289–294, 2011.

[79] J. Kececioglu and D. Gusfield, Reconstructing a history of recombinations from a set of sequences, *Discrete Applied Mathematics*, 88:239–260, 1998.

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

[81] J. Kelleher et al., Inferring whole-genome histories in large population datasets, *Nature Genetics*, 51:1330–1338, 2019.

[82] M. Kendall and C. Colijn, Mapping phylogenetic trees to reveal distinct patterns of evolution, *Molecular Biology and Evolution*, 33:2735–2743, 2016.

[83] A. Kiezun et al., Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency, *PLOS Genetics*, 9:e1003301, 2013.

[84] M. Kimura, The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, *Genetics*, 61:893–903, 1969.

[85] J. F. C. Kingman, On the genealogy of large populations, *Journal of Applied Probability*, 19:27–43, 1982.

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

[87] S. Leslie et al., The fine-scale genetic structure of the British population, *Nature*, 519:309–314, 2015.

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

[89] M. Li et al., Integrative functional genomic analysis of human brain development and neuropsychiatric risks, *Science*, 362:eaat7615, 2018.

[90] N. Li and M. Stephens, Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, *Genetics*, 165:2213–2233, 2003.

[91] X. Liu and Y.-X. Fu, Exploring population size changes using SNP frequency spectra, *Nature Genetics*, 47:555–559, 2015.

[92] J. MacArthur et al., The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog), *Nucleic Acids Research*, 45:D896–D901, 2016.

[93] P. Marjoram, L. Markovtsova, and S. Tavarè, *Estimation in ancestral recombination graphs using Markov chain Monte Carlo*, tech. rep., University of Southern California, 2001.

[94] T. Maruyama, The age of an allele in a finite population, *Genetics Research*, 23:137–143, 1974.

[95] I. Mathieson et al., Genome-wide patterns of selection in 230 ancient Eurasians, *Nature*, 528:499–503, 2015.

[96] I. Mathieson and G. A. T. McVean, Estimating selection coefficients in spatially structured populations from time series data of allele frequencies, *Genetics*, 193:973–984, 2013.

[97] S. Mathieson and I. Mathieson, FADS1 and the timing of human adaptation to agriculture, *Molecular Biology and Evolution*, 35:2957–2970, 2018.

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

[99] G. A. T. McVean et al., The fine-scale structure of recombination rate variation in the human genome, *Science*, 304:581–584, 2004.

[100] P. Menozzi, A. Piazza, and L. Cavalli-Sforza, Synthetic maps of human gene frequencies in Europeans, *Science*, 201:786–792, 1978.

[101] M. Meyer et al., A high-coverage genome sequence from an archaic Denisovan individual, *Science*, 338:222–226, 2012.

[102] M. M. Miretti et al., A High-Resolution Linkage-Disequilibrium Map of the Human Major Histocompatibility Complex and First Generation of Tag Single-Nucleotide Polymorphisms, *The American Journal of Human Genetics*, 76:634–646, 2005.

[103] S. Mirzaei and Y. Wu, RENT+: An improved method for inferring local genealogical trees from haplotypes with recombination, *Bioinformatics*, 33:1021–1030, 2017.

[104] S. R. Myers and R. C. Griffiths, Bounds on the minimum number of recombination events in a sample history, *Genetics*, 163:375–394, 2003.

[105] S. R. Myers et al., A common sequence motif associated with recombination hot spots and genome instability in humans, *Nature Genetics*, 40:1124–1129, 2008.

[106] S. R. Myers et al., A fine-scale map of recombination rates and hotspots across the human genome, *Science*, 310:321–324, 2005.

[107] S. R. Myers et al., Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination, *Science*, 327:876–879, 2010.

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

[109] T. Nagylaki, Evolution of a finite population under gene conversion, *Proceedings of the National Academy of Sciences of the USA*, 80:6278–6281, 1983.

[110] V. M. Narasimhan et al., Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes, *Nature Communications*, 8:303, 2017.

[111] M. Nei, T. Maruyama, and R. Chakraborty, The bottleneck effect and genetic variability in populations, *Evolution*, 29:1–10, 1975.

[112] H. L. Norton et al., Genetic evidence for the convergent evolution of light skin in Europeans and East Asians, *Molecular Biology and Evolution*, 24:710–722, 2006.

[113] J. Novembre et al., Genes mirror geography within Europe, *Nature*, 456:98–101, 2008.

[114] J. Novembre and N. H. Barton, Tread lightly interpreting polygenic tests of selection. *Genetics*, 208:1351–1355, 2018.

[115] D. Novick et al., Sex differences in the course of schizophrenia across diverse regions of the world, *Neuropsychiatric Disease and Treatment*, 12:2927–2939, 2016.

[116] J. A. Palacios, J. Wakeley, and S. Ramachandran, Bayesian nonparametric inference of population size changes from sequential genealogies, *Genetics*, 201:281–304, 2015.

[117] P. F. Palamara, J. Terhorst, Y. A. Song, and A. L. Price, High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability, *Nature Genetics*, 50:1311–1317, 2018.

[118] B. Peng and M. Kimmel, simuPOP: A forward-time population genetics simulation environment, *Bioinformatics*, 21:3686–3687, 2005.

[119] V. Plagnol and J. D. Wall, Possible Ancestral Structure in Human Populations, *PLOS Genetics*, 2:e105, 2006.

[120] PopHumanScan resource; accessed 13 Jan 2019, `https://pophumanscan.uab.cat`.

[121] F. Pouyet, S. Aeschbacher, A. Thiéry, and L. Excoffier, Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences, *Elife*, 7:e36317, 2018.

[122] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, New approaches to population stratification in genome-wide association studies, *Nature Reviews Genetics*, 11:459–463, 2010.

[123] J. K. Pritchard, M. Stephens, and P. Donnelly, Inference of population structure using multilocus genotype data, *Genetics*, 155:945–959, 2000.

[124] K. Prüfer et al., The complete genome sequence of a Neanderthal from the Altai Mountains, *Nature*, 505:43–49, 2014.

[125] A. P. Ragsdale and S. Gravel, Models of archaic admixture and recent history from two-locus statistics, *PLOS Genetics*, 15:1–19, June 2019.

[126] M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel, Genome-wide inference of ancestral recombination graphs, *PLOS Genetics*, 10:e1004342, 2014.

[127] D. E. Reich et al., Genetic history of an archaic hominin group from Denisova Cave in Siberia, *Nature*, 468:1053–1060, 2010.

[128] D. E. Reich et al., Linkage disequilibrium in the human genome, *Nature*, 411:199–204, 2001.

[129] D. Reich et al., Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania, *The American Journal of Human Genetics*, 89:516–528, 2011.

[130] D. F. Robinson and L. R. Foulds, Comparison of phylogenetic trees, *Mathematical Biosciences*, 53:131–147, 1981.

[131] M. R. Robinson et al., Population genetic differentiation of height and body mass index across Europe, *Nature Genetics*, 47:1357–1362, 2015.

[132] R. Roy, J. Chun, and S. N. Powell, BRCA1 and BRCA2: different roles in a common pathway of genome protection, *Nature Reviews Cancer*, 12:68–78, 2012.

[133] P. C. Sabeti et al., Genome-wide detection and characterization of positive selection in human populations, *Nature*, 449:913–918, 2007.

[134] A. Sadier, L. Viriot, S. Pantalacci, and V. Laudet, The ectodysplasin pathway: from diseases to adaptations, *Trends in Genetics*, 30:24–31, 2014.

[135] S. Sankararaman et al., The date of interbreeding between Neandertals and modern humans, *PLOS Genetics*, 8:e1002947, 2012.

[136] S. Schiffels and R. Durbin, Inferring human population size and separation history from multiple genome sequences, *Nature Genetics*, 46:919–925, 2014.

[137] Schizophrenia GWAS study conducted by the Psychiatric Genomics Consortium; accessed 23 Nov 2018, `https://www.med.unc.edu/pgc/results-and-downloads`.

[138] G. Sella and N. H. Barton, Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies, *Annual Review of Genomics and Human Genetics*, 20:461–493, 2019.

[139] J. Shendure and J. M. Akey, The origins, determinants, and consequences of human mutations, *Science*, 349:1478–1483, 2015.

[140] Y. B. Simons, K. Bullaughey, R. R. Hudson, and G. Sella, A population genetic interpretation of GWAS findings for human quantitative traits, *PLOS Biology*, 16:e2002985, 2018.

[141] P. Skoglund et al., Reconstructing prehistoric African population structure, *Cell*, 171:59–71, 2017.

[142] M. Sohail et al., Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies, *eLife*, 8:e39702, 2019.

[143] R. Sokal and C. Michener, A statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin*, 38:409–1438, 1958.

[144] Y. S. Song and J. Hein, Constructing minimal ancestral recombination graphs, *Journal of Computational Biology*, 12:147–169, 2005.

[145] Y. S. Song and J. Hein, Parsimonious reconstruction of sequence evolution and haplotype blocks: finding the minimum number of recombination events, in: ed. by G. Benson and R. Page, vol. Proceedings of the Third International Workshop on Alogarithms in Bioinfomatics (WABI 2003), LNCS 2812, Budapest, Hungary: Springer-Verlag, New York, 2003.

[146] Y. Song et al., Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice, *Current Biology*, 21:1296–1301, 2011.

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

[148] M. Stephens, N. J. Smith, and P. Donnelly, A new statistical method for haplotype reconstruction from population data, *The American Journal of Human Genetics*, 68:978–989, 2001.

[149] A. J. Stern, P. R. Wilton, and R. Nielsen, An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data, *bioRxiv:592675,* 2019.

[150] P. H. Sudmant et al., An integrated map of structural variation in 2,504 human genomes, *Nature*, 526:75–81, 2015.

[151] F. Sun et al., Variation in MLH1 distribution in recombination maps for individual chromosomes from human males, *Human molecular genetics*, 15:2376–2391, 2006.

[152] H. Suzuki et al., Evolutionary and dispersal history of Eurasian house mice Mus musculus clarified by more extensive geographic sampling of mitochondrial DNA, *Heredity*, 111:375–390, 2013.

[153] J. Sved and A. Bird, The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences of the USA*, 87:4692–4696, 1990.

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

[155] J. Terhorst, J. A. Kamm, and Y. S. Song, Robust and scalable inference of population history from hundreds of unphased whole-genomes, *Nature Genetics*, 49:303–309, 2017.

[156] K. M. Teshima and H. Innan, mbs: Modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection, *BMC Bioinformatics*, 10:166, 2009.

[157] M. C. Turchin et al., Evidence of widespread selection on standing variation in Europe at height-associated SNPs, *Nature Genetics*, 44:1015–1019, 2012.

[158] UK Biobank GWAS summary statistics; accessed 4 Oct 2018, `http://www.nealelab.is/uk-biobank`.

[159] J. J. Vitti, S. R. Grossman, and P. C. Sabeti, Detecting natural selection in genomic data, *Annual Review of Genetics*, 47:97–120, 2013.

[160] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard, A map of recent positive selection in the human genome, *PLOS Biology*, 4:e72, 2006.

[161] L. Wang, K. Zhang, and L. Zhang, Perfect phylogenetic networks with recombination, *Journal of Computational Biology*, 8:69–78, 2001.

[162] L. D. Ward and M. Kellis, HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants, *Nucleic Acids Research*, 40:D930–D934, 2011.

[163] S. Wilde et al., Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y, *Proceedings of the National Academy of Sciences of the USA*, 111:4832–4837, 2014.

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

[165] S. Wright, Evolution in Mendelian populations, *Genetics*, 16:97–159, 1931.

[166] Y. Wu, New methods for inference of local tree topologies with recombinant SNP sequences in populations, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8:182–193, 2011.

[167] A. Xue et al., Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes, *Nature Communications*, 9:2941, 2018.

[168] J. H. Young et al., Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion, *PLOS Genetics*, 1:e82, 2005.

[169] G. Zhang et al., Signatures of natural selection on genetic variants affecting complex human traits, *Applied & Translational Genomics*, 2:78–94, 2013.