People assume that time is a strict progression of cause to effect,
but, actually, from a non-linear, non-subjective point of view,
it's more like a big ball of... wibbily-wobbly... timey-wimey... stuff.
— Doctor Who (David Tennant)

1 Estimation of rare allele age

1.1 Introduction

The inference of the genealogical history of a sample is of interest to a myriad of applications in genetic research, both in population and medical genetics. The "age" of an allele, which simply refers to the time since the allele was created by a mutation event, is of particular interest; for example, to observe demographic processes and events, or to better understand the effects of disease-related variants by their time of emergence in a population.

In this chapter, I propose a novel method to estimate the age of an allele, which is based on a collection of statistical models that derive from coalescent theory. Composite likelihood methods recently have gained in popularity for various applications in genetic research. In particular, such methods can be useful in situations where the full likelihood cannot be known analytically or its calculation is computationally prohibitive. Applications that use the composite likelihood based on the coalescent have been pioneered by Hudson (2001) and have been used successfully, for example, for the fine-scale estimation of recombination rates (McVean *et al.*, 2004; Myers *et al.*, 2005). The age estimation method developed here operates in a Bayesian setting to obtain the posterior probability of the time of coalescent events between pairs of haplotypes. These are then consolidated using an approach similar to methods that use the composite likelihood.

In contrast to existing methods for allele age estimation (*e.g.*, see review by Slatkin and Rannala, 2000), the method I present in this chapter does not require prior knowledge about past demographic processes or events. Although an assumption of certain population parameters is required, such as effective population size (N_e), as

well as mutation and recombination rates, these are expected to affect the scaling of time, such that differences between age estimates for different alleles are expected to be proportionally constant.

The age estimation framework presented in this chapter is based on allele sharing at a particular variant site observed in the sample, where the underlying IBD structure is inferred locally around the chromosomal position of the variant under consideration. The methodology for targeted IBD detection presented in Chapters 3 and 4 is therefore essential for this approach; *i.e.* the `tidy` algorithm which includes the four-gamete test (FGT), discordant genotype test (DGT), and the probabilistic IBD model for inference using a Hidden Markov Model (HMM). Additionally, I present a novel haplotype-based HMM method for shared haplotype inference, which can be seen as the logical conclusion of the previously developed genotype-based HMM.

I implemented the age estimation method as a computational tool written in `C++`, referred to as the **`rvage`** algorithm (for rare variant age estimation) which incorporates the full functionality of the previously presented `tidy` algorithm for IBD detection, as well as the novel haplotype-based HMM that is presented in this chapter.*

I begin this chapter by introducing the concept of the method, which is followed by a detailed description of the statistical framework. The method is evaluated in extensive simulation studies, which also consider data error as a source of estimation bias. Although the method can be applied to single-nucleotide polymorphisms (SNP) occurring at any frequency, here, I focus on rare alleles in particular. Finally, I apply this method to data from the 1000 Genomes Project (1000G) Phase III.

1.2 Approach

The mutation that gave rise to a particular allele of interest can be seen as distinguishing event in the history of a population. Immediately after the mutation event, there was only one chromosome in the population that carried the mutant allele. Given a sample of haplotypes, where more than one chromosome carries the focal allele, it is assumed that all copies of the allele were co-inherited from that one chromosome in which the mutation occurred at some point in the past.

According to coalescent theory, any two haplotypes that share the allele are expected to have coalesced more recently than the time of the focal mutation event. Conversely, the coalescent event between one haplotype carrying the allele and one haplotype not

* Rare variant age estimation (`rvage`): `https://github.com/pkalbers/rvage`

carrying the allele is expected to date back to a point in time before the mutation event occurred. This insight is of particular interest as it suggests that the actual time of the mutation event lies somewhere in between two such points in time.

Here, allele age is estimated on the basis of a (composite) Bayesian analysis, where the posterior probability distribution of the time to the most recent common ancestor (T_{MRCA}) is obtained in a pairwise analysis of the haplotypes in a sample. With respect to a given focal site whose age is attempted to be estimated, the following presents the theory in which it is assumed that the shared haplotype structure around that site is known for any pair of haplotypes. In particular, it is assumed that the *breakpoints* of the recombination events that delimit the shared haplotype region are known, and that no recombination has occurred in the interval between breakpoints for the two haplotypes considered. Later sections in this chapter extend the theory to the case where breakpoints are inferred.

There are two main sources of information available from sample data which relate to the T_{MRCA}. First, mutation events occur independently in each lineage and mutations accumulate along the sequence as the haplotype is passed on over generations. Second, recombination events break down the length of the haplotype in each generation independently in each lineage. Thus, the T_{MRCA} between a given pair of chromosomes can be estimated from the number of mutations which segregate in two haplotypes, as well as the genetic length of the haplotype region that is shared between two chromosomes in the sample. In the following (Section 1.2.1), I derive the formulations for three T_{MRCA} estimators. These are referred to as follows.

- Mutation clock, denoted by $\mathcal{T}_{\mathcal{M}}$
- Recombination clock, denoted by $\mathcal{T}_{\mathcal{R}}$
- Combined clock, denoted by $\mathcal{T}_{\mathcal{MR}}$

I then explain the age estimation method in detail in Section 1.2.2 (page 7).

1.2.1 Coalescent time estimators

The posterior probability is proportional to the prior probability of the time to coalescence multiplied by the likelihood of the time. The derivation of the prior distribution on the coalescent time follows from the results given in **??** (page ??), but is briefly described below.

Let t be the number of discrete generations that separate two haplotypes in relation to the most recent common ancestor (MRCA). As shown by Tajima (1983), the probability that two haplotypes find a common ancestor at exactly t generations in the past is

$$f(t) \approx \frac{1}{2N_e} e^{-\frac{t}{2N_e}} \quad \text{CORRECTED} \tag{1.1}$$

where N_e is the effective population size. The expression above relates to the probability distribution of the branch length in the underlying genealogical tree. It is convenient to use a continuous time approximation and measure time in units of $2N_e$ generations such that $\tau = t/2N_e$. Hence, the prior distribution of the coalescent time is $\tau \sim \text{Exp}(1)$, written as

$$\pi(\tau) \propto e^{-\tau}. \quad \text{CORRECTED} \tag{1.2}$$

1.2.1.1 Mutation clock model ($\mathcal{T}_M$)

CORRECTION Section partially rewritten with revised notation

Let the physical length of a shared haplotype region be denoted by h, measured in basepairs. The number of mutational differences along the sequence between a pair of haplotypes is denoted by the discrete random variable S, which is the number of segregating sites in a sample of $n = 2$ haplotypes, for which the infinite sites model is assumed without recombination; *e.g.* see Watterson (1975) and Tavaré *et al.* (1997). Mutations are assumed to occur only once at each site in the history of the sample (Kimura, 1969), such that S reflects the total number of mutation events that have occurred along both lineages since the split from the MRCA.

Mutation events are Poisson distributed, as each mutation represents an independent Bernoulli trial over a large number of sites, where each site has a small probability of mutation. The mutation rate per site per generation is given by μ. In the coalescent, the mutation rate is scaled by population size, which is expressed by the composite mutation parameter $\theta = 4N_e\mu$. It follows that θh is equal to the expected number of pairwise differences per coalescent time unit over the length of the segment.

The number of pairwise differences therefore is modelled as $S \sim \text{Pois}(\theta h \tau)$, for which the probability mass function (PMF) is given as

$$f_S(s) = P(S = s \mid \theta, h, \tau) = \frac{(\theta h \tau)^s}{s!} e^{-\theta h \tau}. \tag{1.3}$$

The likelihood function for the time parameter τ is proportional to Equation (1.3), but requires only those terms that involve τ and where constant terms can be dropped, such that

$$\mathcal{L}(\tau \mid \theta, h, s) \propto \tau^s e^{-\theta h \tau}. \tag{1.4}$$

The posterior probability of the time to coalescence can now be obtained as

$$\begin{aligned} p(\tau \mid \theta, h, s) &\propto \mathcal{L}(\tau \mid \theta, h, s) \times \pi(\tau) \\ &\propto \tau^s \, \mathrm{e}^{-\tau(\theta h+1)} \end{aligned} \tag{1.5}$$

where $\pi(\tau)$ is the coalescent prior, reflecting the general assumption that the expected time to a coalescent event is exponentially distributed.

In the above, the density of the posterior probability is specified up to a missing normalising constant. Note that Equation (1.5) is proportional to (has the form of) the Gamma probability density function (PDF), namely

$$g(\tau \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \, \mathrm{e}^{-\beta\tau}$$

where α is the shape and β the rate parameter. The coalescent prior $\pi(\tau)$ follows the Exponential distribution, which is a special case of the Gamma distribution and therefore is conjugate with the Poisson likelihood. Thus, by using $\alpha = s + 1$ and $\beta = \theta h + 1$, the posterior density can be computed as

$$p(\tau \mid \theta, h, s) = g(\tau \mid s + 1, \theta h + 1). \tag{1.6}$$

1.2.1.2 Recombination clock model ($\mathcal{T}_\mathcal{R}$)

CORRECTION Section partially rewritten with revised notation

The length of a shared haplotype region is delimited by two recombination events that occurred on either side. For either the left or right-hand side, independently, the distance to the first occurrence of a recombination breakpoint follows a Geometric distribution, but where the *genetic* distance can be approximated by the Exponential distribution if time is continuously measured and provided that N_e is large; *e.g.* see Hein *et al.* (2004). The recombination rate per site per generation is given by ρ; again, the rate is scaled by population size and the composite recombination parameter $\psi = 4N_e\rho$ is used.* Suppose, for now, that recombination is uniform.

The distance is modelled such that $D \sim \mathrm{Exp}(\psi\tau)$, where D is a random variable used to denote the physical distance between a given focal position and a recombination breakpoint. Hence, the PDF of the distance until a recombination breakpoint is

$$P(D = d \mid \psi, \tau) = \psi\tau \, \mathrm{e}^{-\psi\tau d}. \tag{1.7}$$

* Note that the literature often specifies ρ as the population-scaled recombination rate and r as the rate per site per generation.

However, in boundary cases where the shared haplotype segment is delimited by the chromosomal end, it follows from the Exponential distribution that

$$P(D > d \mid \psi, \tau) \;=\; \mathrm{e}^{-\psi\tau d}. \tag{1.8}$$

Equations (1.7) and (1.8) above can be simplified to

$$f_D(d) \;=\; (\psi\tau)^b \; \mathrm{e}^{-\psi\tau d} \tag{1.9}$$

where b is the result of an indicator function of the breakpoint defined as

$$b \;:=\; \mathbf{1}_d \;=\; \begin{cases} 0 & \text{if } D > d \text{ (\textit{i.e.} boundary case)} \\ 1 & \text{otherwise.} \end{cases}$$

Considering Equation (1.9), the likelihood function for τ can now be written as

$$\mathcal{L}(\tau \mid \psi, d, b) \;\propto\; \tau^b \; \mathrm{e}^{-\psi d\tau} \tag{1.10}$$

but which can be extended to consider the distances observed on the left and right-hand side relative to a given focal position. The observed physical length of the shared haplotype segment is now expressed as the sum of both left and right distances; *i.e.* $h = d_L + d_R$. Hence, the likelihood function in support of τ is

$$\mathcal{L}(\tau \mid \psi, h, b_L, b_R) \;\propto\; \tau^{b_L + b_R} \; \mathrm{e}^{-\psi h\tau} \tag{1.11}$$

where b_L, b_R indicate the breakpoint on the left and right-hand side, respectively.

Importantly, the term ψh refers to the genetic length of the shared haplotype region, but where ψ is rarely constant along a chromosome. It is straightforward to compute the value of ψh by using a chromosome-specific recombination map from which the genetic distance between breakpoint positions can be derived.

The posterior probability is obtained as

$$\begin{aligned} p(\tau \mid \psi, h, b_L, b_R) \;&\propto\; \mathcal{L}(\tau \mid \psi, h, b_L, b_R) \;\times\; \pi(\tau) \\ &\propto\; \tau^{b_L + b_R} \; \mathrm{e}^{-\tau(\psi h + 1)}. \end{aligned} \tag{1.12}$$

As in the previous section, the form of the posterior probability obtained above suggests a Gamma PDF with $\alpha = b_L + b_R + 1$ and $\beta = \psi h + 1$. Thus, the posterior density can be computed as

$$p(\tau \mid \psi, h, b_L, b_R) \;=\; g(\tau \mid b_L + b_R + 1, \psi h + 1). \tag{1.13}$$

1.2.1.3 Combined clock model ($\mathcal{T}_{\mathcal{MR}}$)

CORRECTION Section partially rewritten with revised notation

The parameters defined in the mutation clock and recombination clock models given above are combined in the following way. The likelihood function in support of τ considers Equations (1.3) and (1.9) on page 4 and page 6 and is given as

$$\mathcal{L}(\tau \mid \theta, \psi, h, s, b_L, b_R) \propto \tau^{s+b_L+b_R} \, \mathrm{e}^{-\tau h(\theta+\psi)}.$$

However, it is convenient to replace the term $h(\theta + \psi)$ above with $h_p + h_g$, where $h_p = \theta h$ and $h_g = \psi h$, so as to consider the physical and genetic lengths separately; *e.g.* when recombination is not uniform and ψh is determined from the distances given in a genetic map. Therefore,

$$\mathcal{L}(\tau \mid h_p, h_g, s, b_L, b_R) \propto \tau^{s+b_L+b_R} \, \mathrm{e}^{-\tau(h_p+h_g)} \tag{1.14}$$

from which the posterior probability is obtained as

$$\begin{aligned} p(\tau \mid h_p, h_g, s, b_L, b_R) &\propto \mathcal{L}(\tau \mid h_p, h_g, s, b_L, b_R) \times \pi(\tau) \\ &\propto \tau^{s+b_L+b_R} \, \mathrm{e}^{-\tau(h_p+h_g+1)}. \end{aligned} \tag{1.15}$$

As was done in both the mutation and recombination clock models, here, the Gamma PDF is used with $\alpha = s + b_L + b_R + 1$ and $\beta = h(\theta + \psi) + 1 = h_p + h_g + 1$ to compute the posterior density, *i.e.*

$$p(\tau \mid h_p, h_g, s, b_L, b_R) = g(\tau \mid s + b_L + b_R + 1, h_p + h_g + 1). \tag{1.16}$$

Note that a similar derivation has been used by Schroff (2016).

1.2.2 Inference of allele age from coalescent time posteriors

Consider a focal allele of interest that is shared by some of the haplotypes in a sample. The time at which this allele was created by a mutation event is bound by the times of the two coalescent events that delimit the length of the branch on which the mutation occurred in the underlying coalescent tree; see the example provided in Figure 1.1 (next page). The haplotypes which inherited the allele (*carriers*) are distinguished from the haplotypes that do not carry the allele (*non-carriers*). Thus, the sample is divided into two disjoint subsamples; let X_c denote the set of chromosomes which share a given allele, and X_d the set of chromosomes which do not carry that allele.

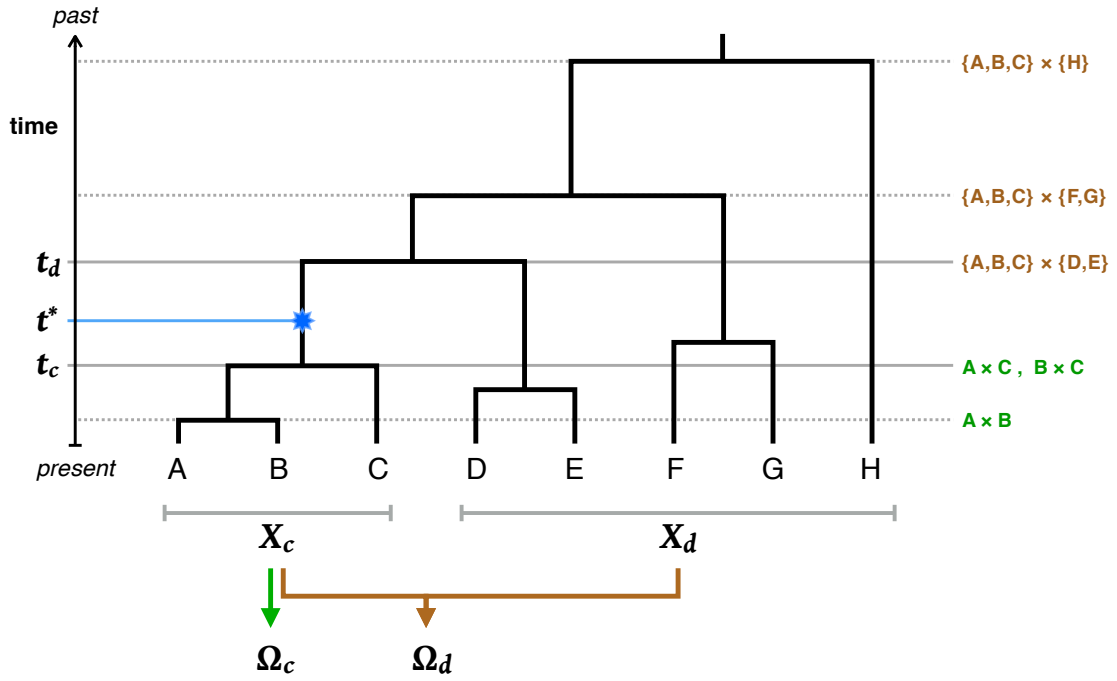


Figure 1.1: Allele age in relation to concordant and discordant pairs. The genealogy of a sample of eight haplotypes is shown of which A, B, and C share a focal allele that derived from a mutation event as indicated in the tree (*star*). These chromosomes constitute the set of *carriers*, denoted by X_c, which are distinguished from the set of *non-carriers*, denoted by X_d. Horizontal lines indicate the time of each coalescent event in the history of the sample within the local genealogy. The time of the focal mutation event is denoted by t^*; the two coalescent events at time t_c and t_d define the length of the branch on which the focal mutation event occurred. In particular, t_c and t_d correspond to the time until all haplotypes in X_c have coalesced and the time at which the derived lineage joins the ancestral lineage of the most closely related haplotype in X_d, respectively.

It follows that all lineages in X_c coalesce before any of them can coalesce with a lineage in X_d. Any coalescent event between two lineages in X_c must have occurred *earlier* than the focal mutation event (back in time). On the other hand, any coalescent event between one lineage in X_c and one lineage in X_d must have occurred *later* than the focal mutation event (back in time). Pairs of haplotypes in X_c are referred to as *concordant* pairs, whereas pairs formed by strictly taking one haplotype from X_c and another from X_d are *discordant* pairs. The sets Ω_c and Ω_d are defined to contain all concordant and discordant pairs, respectively.

In the following, I describe how the posterior density of the T_{MRCA} is obtained for concordant and discordant pairs to eventually arrive at an estimate of allele age. To distinguish the population-scaled time τ, as defined for the T_{MRCA}, from the time of the mutation event, let the latter be denoted by the likewise population-scaled time t. Informally, the actual time of a mutation event is found at the "sweet spot" in between the earlier coalescent event at time t_c and the later coalescent event at time t_d; see Figure 1.1.

1.2.2.1 Cumulative coalescent function (CCF)

CORRECTION Section partially rewritten with revised notation

At a given focal site at which the possible concordant and discordant pairs in the sample have been sorted into the sets Ω_c and Ω_d, respectively, each pair is analysed in turn to obtain a posterior on their T_{MRCA}. Importantly, to find the time of the focal mutation event, it is of interest to obtain the probability distribution of the T_{MRCA} relative to t. Here, this task is accomplished by introducing the cumulative coalescent function (CCF) which is defined as the posterior cumulative distribution function (CDF) with respect to a given pair of haplotypes, denoted by i, j. In simple terms, the CCF is expressed as

$$\Phi_{ij}(t) \;=\; \begin{cases} P(\tau \leq t) & \text{if } \{i,j\} \subseteq \Omega_c \quad (\textit{i.e.}\text{ concordant pairs}) \\ P(\tau > t) = 1 - P(\tau \leq t) & \text{if } \{i,j\} \subseteq \Omega_d \quad (\textit{i.e.}\text{ discordant pairs}). \end{cases} \tag{1.17}$$

Specifically, the term $P(\tau \leq t)$ implies that concordant pairs have coalesced *earlier* than or at the time of the focal mutation event (back in time), and $P(\tau > t)$ implies that discordant pairs have coalesced *later* than the mutation event (back in time).

Since each clock model defines the posterior using the Gamma distribution, it is straightforward to obtain the CCF from the Gamma CDF; formally given as

$$G(t) \;=\; P(\tau \leq t \mid \alpha, \beta) \;=\; \int_0^t g(u \mid \alpha, \beta)\, du \tag{1.18}$$

where α, β are defined according to the clock model used, with parameter values obtained from the analysis of a given haplotype pair at a focal site in the genome. Notably, because α is a positive integer in each of the clock models considered, the Gamma distribution simplifies to the Erlang distribution, such that the above becomes equal to (Papoulis and Pillai, 2002)

$$F(t) \;=\; P(\tau \leq t \mid \alpha, \beta) \;=\; 1 - \mathrm{e}^{-\beta t} \sum_{i=0}^{\alpha-1} \frac{(\beta t)^i}{i!}. \tag{1.19}$$

Further, to obtain point estimates from the posterior of the T_{MRCA}, it follows from the Gamma (Erlang) distribution that the mean is $\frac{\alpha}{\beta}$ and the mode is $\frac{\alpha-1}{\beta}$. Note that no simple closed form exists for the median, but which in practise is straightforward to approximate by scanning the CCF to find, for example, the times of the 1st, 2nd (*i.e.* median), and 3rd quartiles.

1.2.2.2 Allele age estimation from the composite posterior distribution

CORRECTION Section partially rewritten with revised notation

A at given focal site, the CCF is obtained for concordant and discordant pairs. Because the T_{MRCA} of concordant pairs would extend to a point below the focal mutation event and the T_{MRCA} of discordant pairs above that point in time, ideally, it would be expected that the age of an allele can be derived from the structure of posteriors. Here, the CCF posteriors are combined in the following way.

$$\Lambda_k(t) \propto \prod_{i,j \in A_k} \Phi_{ij}(t \mid \alpha, \beta) \tag{1.20}$$

for focal site k at which haplotype pairs have been sorted into the collection $A_k = \{\Omega_c, \Omega_d\}$, according to allele sharing at that site. Again, α, β are defined by the clock model used and obtained from parameter values observed for pair i, j. In the following, the term *composite posterior* is used to refer to the result obtained using Equation (1.20).

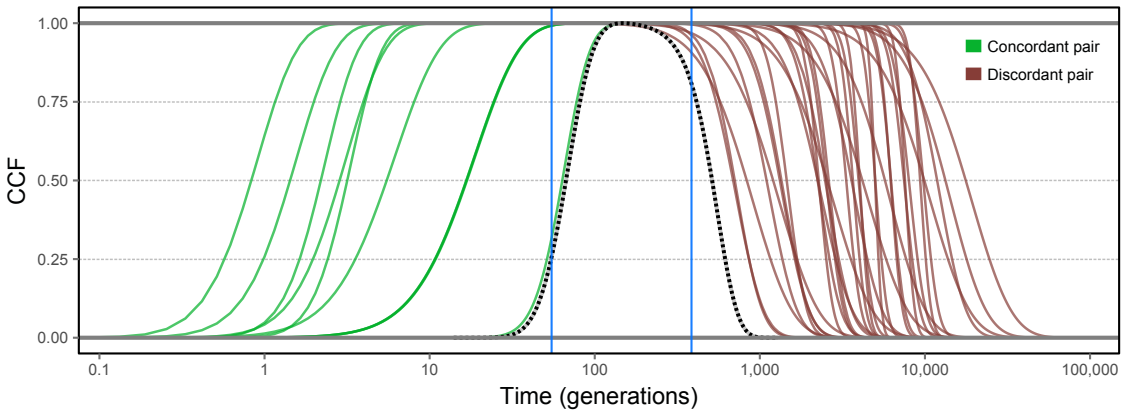


Figure 1.2: Example of concordant and discordant posterior distributions and the resulting composite posterior. A target variant was randomly selected from simulated data. The CCF was obtained for the set of possible concordant pairs and a random subset of discordant pairs. The thicker *dotted* line shows the distribution of the maximised composite posterior. The *blue* lines mark the actual time of coalescent events below and above the focal mutation event; *i.e.* t_c (*left*) and t_d (*right*), determined from simulation records. Their distance corresponds to the length of the branch on which the focal mutation event occurred.

The composite posterior distribution can now be obtained over $t \in (0, \infty)$. However, in practise, it is unlikely that the relationship of i, j can be traced back further than a small multiple of N_e (*e.g.* ~ 10). An example is given in Figure 1.2 (this page), showing the CCF for concordant and discordant pairs, as well as the maximised composite posterior

distribution. Notably, the "width" of the distribution is expected to be determined by the underlying branch length. In the following, the mode of the composite posterior distribution is taken as a point estimate for the age of an allele, denoted by $\hat{t}$.

1.2.2.3 Note on composite likelihood methods

ADDITION Section included

There is extensive literature on the topic of composite likelihood methods and their application to problems where the full likelihood function cannot be known or is intractable. In its general form, the composite likelihood is defined as the weighted product of the likelihoods associated with a set of events $\{X_1, \ldots, X_z\}$; *i.e.* (Lindsay, 1988)

$$\mathcal{CL}(\vartheta \mid y) = \prod_{z \in Z} \mathcal{L}_z(\vartheta \mid y)^{w_z} \tag{1.21}$$

where $\mathcal{L}_z(\vartheta \mid y)$ is the likelihood function proportional to density $f(y \in X_z \mid \vartheta)$ with parameter (vector) ϑ, and w_z are non-negative weights.

The use of the composite likelihood in a Bayesian setting has been discussed, for example, by Pauli *et al.* (2011) who argued that, formally, a posterior distribution can be obtained with the composite likelihood; *i.e.*

$$p_{\mathcal{CL}}(\vartheta \mid y) \propto \pi(\vartheta) \times \mathcal{CL}(\vartheta \mid y) \tag{1.22}$$

where $\pi(\vartheta)$ is a suitable prior on the parameter. The properties of the above were described by Pauli *et al.* (2011) who, as a result, proposed an adjustment to the composite likelihood by choosing appropriate weights (w_z) to improve approximation of the full posterior distribution.

The "composite posterior" given in Equation (1.20) on page 10 follows a similar approach in context of the above, but is defined proportional to the product of posteriors. While $\Lambda_k(t)$ itself cannot be regarded as a composite likelihood, it can be argued that the proposed method is an (*ad hoc*) approach equivalent to using the composite likelihood in a Bayesian setting, yet without specifying an appropriate weighting function.

1.2.2.4 Note on the computational burden

A major caveat is the computationally demanding analysis of each haplotype pair in Ω_c and Ω_d per target site. The number of concordant and discordant pairs, denoted by n_c and n_d, respectively, varies dependent on the observed frequency of the focal allele and

sample size. For a given f_k variant, the number of possible concordant pairs is

$$\max[n_c] = \binom{k}{2} = \frac{k(k-1)}{2} \tag{1.23}$$

where k is the number of allele copies observed in the sample. The number of possible discordant pairs is given by

$$\max[n_d] = k(2N-k) \tag{1.24}$$

where N refers to the diploid sample size. The total number of pairwise analyses conducted per target site is the sum of n_c and n_d.

The estimation process for a single focal allele quickly becomes intractable if the allele is observed at higher frequencies or if sample size is large. This can be particularly problematic if many target sites are considered. For example, for $N = 1{,}000$, each f_2 variant has $n_c = 2$ and $n_d = 3{,}996$, whereas each f_{20} variant already has $n_c = 190$ and $n_d = 19{,}600$. Therefore, in practise, the computational burden is reduced by employing a sampling regime where, for example, pairs in Ω_c and Ω_d are picked at random.

REMOVED Section “Anticipated limitations”

1.3 Evaluation

The following sections describe the data used in this chapter, as well as the metrics used to evaluate age estimation results.

1.3.1 Data generation

The following simulated datasets were available. First, sample data were simulated under a simple demographic model of constant population size ($N_e = 10{,}000$) with mutation rate $\mu = 1 \times 10^{-8}$ per site per generation and constant recombination rate $\rho = 1 \times 10^{-8}$ per site per generation, using `msprime` (Kelleher *et al.*, 2016). Note that by setting the mutation and recombination rates to constant and equal values, the physical and genetic lengths are identical when measured in Megabase (Mb) and centiMorgan (cM), respectively. The size of the simulated dataset was 2,000 haplotypes, which were randomly paired to form a sample of $N = 1{,}000$ diploid individuals. The length of the simulated region was 100 Mb (100 cM), resulting in 326,335 variant sites. This dataset is denoted by $\mathcal{D}_A$.

Second, the dataset simulated in Chapter 3 was included here to evaluate the age estimation method in presence of data error. Briefly, the simulation was performed under a demographic model that recapitulates the human expansion out of Africa; following Gutenkunst *et al.* (2009). A sample of 5,000 haplotypes was simulated with $N_e = 7{,}300$, a mutation rate of $\mu = 2.35 \times 10^{-8}$ per site per generation, and variable recombination rates taken from human chromosome 20; Build 37 of the International HapMap Project (HapMap) Phase II (International HapMap Consortium *et al.*, 2007; International HapMap 3 Consortium *et al.*, 2010), yielding 0.673 million segregating sites over a chromosomal length of 62.949 Mb (108.267 cM). The simulated haplotypes were randomly paired to form a sample of $N = 2{,}500$ diploid individuals. Haplotype data were converted into genotypes and subsequently phased using `SHAPEIT2` (Delaneau *et al.*, 2008, 2013). This facilitated the assessment of the impact of phasing error on the age estimation process.

Third, the dataset described above was retrofitted in Chapter 4 to include realistic proportions of empirically estimated error, which was equally distributed in the derived genotype and haplotype datasets (both "true" and phased haplotype data). Here, data *before* and *after* the inclusion of error are distinguished by referring to dataset $\mathcal{D}_B$ and dataset $\mathcal{D}_B^*$, respectively. Note that in the following the term *genotype error* is used, even in analyses that operate on haplotype data, as error proportions were estimated from misclassified genotypes in 1000G data ("1000G.A" in Chapter 4, **??**, page ??).

In each dataset, simulation records were queried to determine the underlying IBD structure of each pair of individuals analysed in this work. Note that the simulated genealogy underlying $\mathcal{D}_B$ was identical to $\mathcal{D}_B^*$, such that direct comparisons were possible between results obtained before and after error. True IBD intervals were found in simulated genealogies by scanning the sequence until the MRCA of a given pair of haplotypes changed, on both sides of a given target position. Interval breakpoints were identified on basis of the observed variant sites in the sample, such that the resulting true IBD segment defined the smallest interval detectable from available data.

1.3.2 Accuracy analysis

Coalescent simulators may not define the exact time point at which a mutation event occurred, because mutations are independent of the genealogical process (if simulated under neutrality) and can therefore be placed randomly along the branches of the simulated tree. Mutation times are not specified in `msprime`, but the times of coalescent events are recorded.

In simulations, the probability of placing a mutation on a particular branch is directly proportional to its length, which itself is delimited by the time of the coalescent event below (joining the lineages that derive from that branch) and the time of the coalescent event above (joining that branch with the tree back in time). Here, the times of coalescence below and above a particular mutation event are denoted by t_c and t_d, respectively, against which the accuracy of the estimated allele age $\hat{t}$ is measured.

Although the true time of a mutation event was not known from the simulations performed, an indicative value for the age of an allele was derived from the logarithmic "midpoint" (or *log-average*) between coalescent events, which is denoted by t_m and calculated as the geometric mean of t_c and t_d, namely

$$t_m = \sqrt{t_c\, t_d}\,. \quad \text{CORRECTED} \tag{1.25}$$

However, note that the arithmetic mean, $\frac{1}{2}(t_c + t_d)$, would be appropriate given that mutation events can be placed uniformly between t_c and t_d. The geometric mean is nonetheless useful and was chosen for practical reasons (*e.g.*representation on log-scale).

Accuracy was measured using Spearman's rank correlation coefficient, r_S, which is a robust measure for the strength of the monotonic relationship between two variables; *i.e.* the inferred allele age ($\hat{t}$) and true time proxies (t_c, t_m, or t_d). Note that the squared Pearson correlation coefficient, r^2, was used in previous chapters but was regarded as being less suitable here, as both the inferred and true age are expected to vary on log-scale, and the Pearson coefficient measures the linear relationship between variables. However, for example, r^2 of log-transformed age could have been used, but which was not additionally considered here, in order to keep the analysis brief. Also, to indicate bias, the root mean squared logarithmic error (RMSLE) was calculated as a descriptive score for the magnitude of error (here defined on $\log_{10}$).

To better illustrate the distribution of age estimates obtained in an analysis, the *relative age* was computed, $\hat{t}_{rel}$, for each allele by normalising the time scale conditional on the time interval between the coalescent events at t_c and t_d, such that age estimates were "mapped" on the same scale relative to the branch length spanned between t_c and t_d; this was calculated as below.

$$\hat{t}_{rel} = \frac{\log\left[\frac{\hat{t}}{t_c}\right]}{\log\left[\frac{t_d}{t_c}\right]} \tag{1.26}$$

As a result, the times of coalescent events at t_c and t_d are mapped to 0 and 1, respectively. It follows that that $\hat{t}_{rel} < 0$ indicates underestimation and $\hat{t}_{rel} > 1$ overestimation in relation to the true interval in which the mutation event could have occurred.

Further, an age estimate was counted as being "correct" if $t_c \leq \hat{t} \leq t_d$, which is equal to the condition $0 \leq \hat{t}_{rel} \leq 1$. The proportion of age estimates that fall within this interval is reported.

1.4 Validation of the method

The allele age estimation method relies on an (ideally) correct inference of the haplotype region shared by descent between two chromosomes relative to a target site. Several approaches for targeted, pairwise inference of the shared haplotype have been developed in the previous chapters, which are applied further below. To first establish *proof of concept* of the age estimation method, its performance was evaluated given complete knowledge of the underlying shared haplotype structure. That is, the "true" shared haplotype segments were determined from simulation records and analysed using haplotype data in dataset $\mathcal{D}_A$.

Further, because an exhaustive analysis of all possible haplotype pairs becomes computationally intractable, it is convenient to reduce the number of pairwise analyses that are conducted per target allele. In particular, because the current analysis focused on rare alleles, the number of discordant pairs, n_d, was reduced such that Ω_d consisted of a substantially smaller set of randomly retained pairs. Here, the impact on estimation accuracy was assessed under different nominal thresholds applied to n_d (listed below).

n_d	Pairwise analyses
10	0.462 million
50	0.862 million
100	1.362 million
500	5.362 million
1,000	10.366 million

A number of 10,000 rare variants were randomly selected as target sites at allele frequency $\leq 1\,\%$ ($f_{[2,20]}$). Each clock model was considered separately and the same set of sites was analysed under each threshold. However, note that because discordant pairs were chosen at random, these differed in each analysis. Model parameters in `rvage` were specified according to parameters used for the simulation of $\mathcal{D}_A$ ($N_e = 10{,}000$; $\mu = 1 \times 10^{-8}$; $\rho = 1 \times 10^{-8}$).

1.4.1 Results

An overview is illustrated in Figure 1.3 (next page), which shows the density of true and estimated age under each clock model; results are shown for $n_d = 10$, $n_d = 100$, and $n_d = 1{,}000$, to better distinguish differences visually. Note that, here, true age was set to t_m (the geometric mean of t_c and t_d).

Despite the substantial difference in the number of pairwise analyses, overall accuracy was high for each threshold and under each clock model. A higher n_d threshold was generally found to improve overall accuracy. At lower thresholds, each model showed a tendency to overestimate allele age, which most likely resulted from the smaller set of discordant pairs, as the individuals that are more closely related to the focal haplotypes may or may not be captured.

The proportion of target alleles for which age was correctly estimated ($t_c \leq \hat{t} \leq t_d$) increased with higher n_d thresholds under each clock model. This was lowest in $\mathcal{T}_{\mathcal{M}}$, where 36.610 %, 51.110 %, and 66.280 % were correctly inferred for n_d at 10, 100, and 1,000, respectively, and relatively high in $\mathcal{T}_{\mathcal{R}}$, where 55.790 %, 70.600 %, and 70.510 % were correct, respectively. The highest proportion of correct alleles was 79.930 % in $\mathcal{T}_{\mathcal{MR}}$ and $n_d = 1{,}000$. The proportion of overestimated alleles ($\hat{t} > t_d$) decreased in all clock models at higher n_d thresholds, showing a modest decrease in $\mathcal{T}_{\mathcal{M}}$ (63.380 % to 32.660 % for n_d at 10 and 1,000, respectively), a substantial decrease in $\mathcal{T}_{\mathcal{R}}$ (43.450 % to 6.450 %, respectively), and a notable decrease in $\mathcal{T}_{\mathcal{MR}}$ (46.780 % to 15.640 %, respectively). Since $\mathcal{T}_{\mathcal{M}}$ showed a tendency to overestimate allele age, the proportion of underestimated alleles was low (1.060 % for $n_d = 1{,}000$), which was similarly low in $\mathcal{T}_{\mathcal{MR}}$ (4.430 %), and highest in $\mathcal{T}_{\mathcal{R}}$ (23.040 %).

A complete summary of results is given in Table 1.1 (page 18). Throughout, rank correlation (r_S) was highest for $n_d = 1{,}000$; see Table 1.1. However, for all thresholds, correlations with t_c were higher than correlations with t_m, which in turn were higher than correlations with t_d. Such a pattern may be expected as the number of concordant pairs, n_c, was not reduced, such that the t_c was inferred with higher accuracy. Highest accuracy was seen for the mutation clock model, $\mathcal{T}_{\mathcal{M}}$, where r_S for $n_d = 1{,}000$ was 0.923, 0.904, and 0.723 for t_c, t_m, and t_d, respectively. By comparison, the recombination clock, $\mathcal{T}_{\mathcal{R}}$, yielded the lowest levels of overall accuracy at each threshold, but did not differ markedly from $\mathcal{T}_{\mathcal{M}}$; *e.g.* r_S for $n_d = 1{,}000$ was 0.889, 0.895, and 0.739 for t_c, t_m, and t_d, respectively. The combined clock, $\mathcal{T}_{\mathcal{MR}}$, was found to be more accurate for t_m and t_d at higher thresholds.

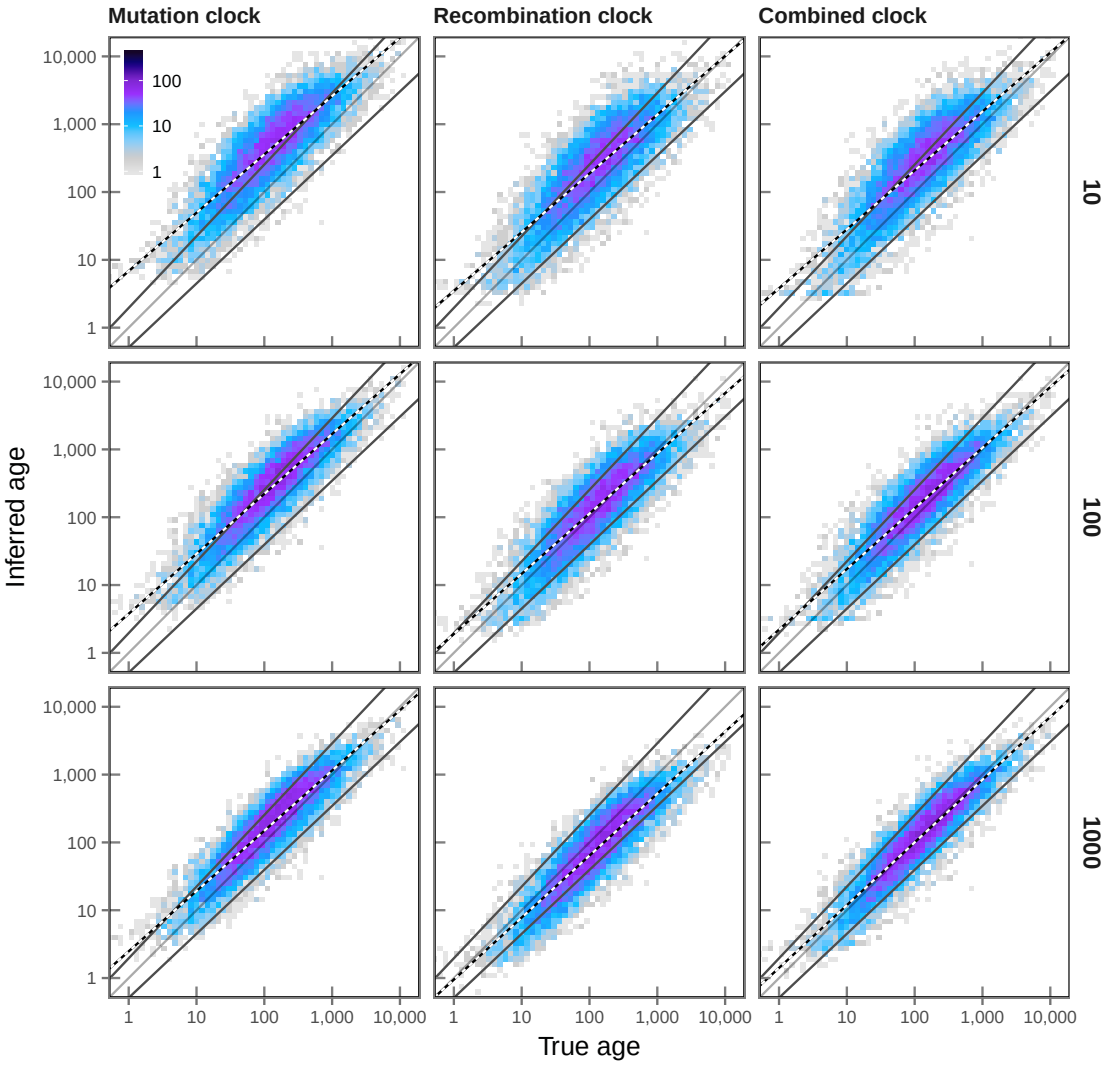


Figure 1.3: True and inferred age under varying numbers of discordant pairs. A set of 10,000 target sites was randomly drawn in $f_{[2,20]}$ (shared allele frequency ≤ 1%) in a simulated sample of 2,000 haplotypes. Different numbers of sampled discordant pairs were analysed on the same set of target variants, which is shown for $n_d = 10$, $n_d = 100$, and $n_d = 1{,}000$ (indicated at the *right* of each row). True IBD was used to estimate allele age. IBD breakpoints were determined from simulation records and defined as the first variant sites observed in the data following the two recombination events on each side of a given focal position. Age was estimated under each of the three clock models; *i.e.* mutation clock, $\mathcal{T}_{\mathcal{M}}$, recombination clock, $\mathcal{T}_{\mathcal{R}}$, and combined clock, $\mathcal{T}_{\mathcal{MR}}$ (indicated at the *top* of each column). Each panel shows the density distribution of true and inferred age (numbers indicated by the colour-gradient). Note that the "true age" of a focal allele was set to t_m, which is the geometric mean of t_c and t_d, *i.e.* the true time of the coalescent event from which the focal allele derived (t_c) and the true time of the coalescent event immediately preceding that event (t_d) in the history of the sample, respectively; these are indicated by their regression trend lines *below* and *above* the dividing line at t_m, respectively. The *black-white* line indicates the line of best fit resulting from linear regression of age estimates, using the posterior mode of the composite likelihood distribution as the inferred age value. True and inferred age are both shown on log-scale.

Table 1.1: Estimation accuracy under varying numbers of discordant pairs. Different thresholds for the number of randomly formed discordant pairs, n_d, were analysed to evaluate the impact on the accuracy of allele age estimation. Note that all possible concordant pairs were included in each analysis; *i.e.* n_c was not reduced. True IBD segments were used to focus on the differences induced by varying n_d thresholds. Each analysis was conducted on the same set of 10,000 randomly selected rare variants at allele frequency ≤ 1%. Accuracy was measured using the rank correlation coefficient, r_S, and the magnitude of error, RMSLE, between the estimated age, $\hat{t}$ and the the times of coalescent events; *i.e.* the time until all haplotypes in X_c have coalesced, t_c, and the time of the immediately preceding coalescent event, t_d, which joined the lineages in X_c and X_d back in time, as well as the geometric mean of both, t_m.

Clock	n_d	Rank correlation (r_S)			RMSLE		
		t_c	t_m	t_d	t_c	t_m	t_d
$\mathcal{T}_\mathcal{M}$	10	**0.907**	0.842	0.632	0.963	0.624	**0.574**
	50	**0.918**	0.872	0.674	0.823	**0.487**	0.528
	100	**0.920**	0.884	0.692	0.763	**0.431**	0.521
	500	**0.920**	0.907	0.731	0.626	**0.308**	0.533
	1,000	**0.923**	0.904	0.723	0.606	**0.299**	0.547
$\mathcal{T}_\mathcal{R}$	10	**0.881**	0.816	0.612	0.714	**0.443**	0.609
	50	**0.889**	0.844	0.651	0.578	**0.349**	0.633
	100	**0.892**	0.857	0.671	0.519	**0.319**	0.653
	500	**0.892**	0.886	0.720	0.390	**0.304**	0.728
	1,000	0.889	**0.895**	0.739	0.345	**0.329**	0.772
$\mathcal{T}_{\mathcal{MR}}$	10	**0.891**	0.829	0.624	0.745	**0.455**	0.589
	50	**0.901**	0.865	0.675	0.624	**0.348**	0.586
	100	**0.905**	0.881	0.699	0.574	**0.309**	0.593
	500	0.909	**0.914**	0.753	0.469	**0.243**	0.626
	1,000	0.911	**0.914**	0.751	0.464	**0.243**	0.629

The magnitude of error, measured by RMSLE scores, was lowest for t_m, indicating that the majority of alleles were correctly dated between t_c and t_d; except in $\mathcal{T}_\mathcal{M}$ for $n_d = 10$, in which allele age was overestimated and therefore closer to t_d.

The difference between $n_d = 500$ and $n_d = 1{,}000$ was small overall (see Table 1.1), suggesting that further improvements in accuracy may not be attained by increasing the threshold.

A comparison of the inferred age distributions at distinct f_k ranges is presented in Figure 1.4 (next page), again shown for $n_d = 10$, $n_d = 100$, and $n_d = 1{,}000$. Notably, the accuracy of target alleles at lower frequencies was overall higher compared to alleles observed at higher frequencies. This difference was consistent across n_d thresholds under the mutation clock model, $\mathcal{T}_\mathcal{M}$. For example, at $n_d = 10$, the proportion of correctly dated alleles was higher in the $f_{[2,5]}$ range (48.356 %) compared to alleles at $f_{(5,10]}$ (29.445 %). At $n_d = 1{,}000$, overall accuracy was increased but the difference for alleles at lower and higher frequencies remained; *i.e.* 77.819 % and 60.834 % at $f_{[2,5]}$ and $f_{(5,10]}$, respectively. Under the recombination clock model, $\mathcal{T}_\mathcal{R}$, these differences were reduced at higher n_d

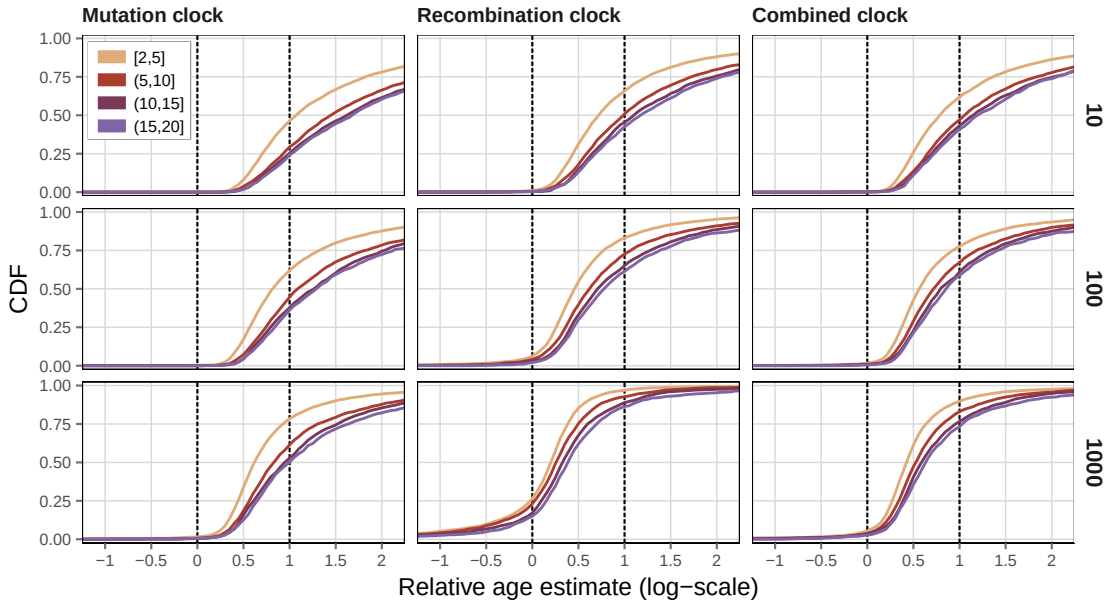


Figure 1.4: Relative age under varying numbers of discordant pairs. A randomly drawn set of 10,000 target sites at allele frequency ≤ 1%, *i.e.* $f_{[2,20]}$, was analysed under each of the three clock models (indicated at the *top* of each column) and with different numbers of sampled discordant pairs; $n_d = 10$, $n_d = 100$, and $n_d = 1{,}000$ (indicated at the *right* of each row). The analysis was conducted using the true IBD breakpoints as derived from simulation records, defined as the first variant sites observed in the data that immediately follow the two recombination events on each side distal to a given focal site. The relative age, $\hat{t}_{rel}$, was calculated as given in Equation (1.26), such that the true times of concordant and discordant coalescent events, t_c and t_d, sit at 0 and 1, respectively (*dashed* lines). Note that $\hat{t}_{rel}$ is defined on log-scale. The CDF of relative age estimates is shown per f_k group, where target variants were pooled by their allele count in the data, in ranges of $f_{[2,5]}$, $f_{(5,10]}$, $f_{(10,15]}$, and $f_{(15,20]}$.

thresholds. At $n_d = 10$, 66.608 % and 50.344 % of alleles were correctly dated at $f_{[2,5]}$ and $f_{(5,10]}$, respectively, whereas at $n_d = 1{,}000$ these proportions were 72.258 % and 69.826 % at the same frequency ranges, respectively.

1.4.2 Discussion

In summary, the method as well as the clock models proposed were able to estimate allele age from IBD information alone, without prior knowledge of the demographic history of the sample. However, because data were simulated under a simple demographic model (dataset $\mathcal{D}_A$), further evaluation is appropriate (*e.g.* using datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$; see next section). The analysis considered true IBD segments and therefore evaded the effects that would result from inexact IBD detection. Since true IBD was determined conditional on the observed variation in the data, the analysis reflected the practical feasibility of age estimation given available data.

The implemented sampling regime for discordant pairs sought to find a compromise between computational tractability and the chance of randomly selecting haplotypes that are informative for the estimation. However, ideally, to minimise the computational burden while simultaneously improving estimation accuracy, it would be desirable to consider the nearest neighbours to the focal shared haplotypes in the local genealogy. If the nearest neighbours are found among the haplotypes in X_d and paired with the focal haplotypes in X_c they are likely to coalesce more closely to t_d and would therefore be more informative for the estimation of focal allele age.

For instance, a simple approach would be to compute the Hamming distance between haplotypes in X_c and X_d within a short region around the position of a given target site, such that a subset of presumed nearest neighbours can be selected based on a distance ranking. In practice, however, there are two caveats to such an approach. First, it would be computationally expensive to conduct an additional pairwise analysis for the (whole) sample at each target site, which may not outweigh the improvement gained through the reduction of n_d. Second, a dilemma arises in presence of data error, as the identification of nearest neighbours is likely to give preference to haplotypes in which the focal allele has been missed. Such *false negatives* distort the estimation of allele age as the CCF computed for false discordant pairs could bias the resulting composite posterior distribution.

It is important to note that the problem of finding false negatives in the data cannot be avoided if discordant pairs are formed by a random sampling process, but the chance of including false negatives is reduced if n_d is small in comparison to the (haploid) sample size. Hence, the n_d threshold defines a balance between accuracy and expected bias.

1.5 Age estimation using inferred shared haplotype segments

The `tidy` algorithm for targeted IBD detection (see Chapters 3 and 4) was fully integrated in `rvage`, such that several methods for IBD detection were available to inform allele age estimation; namely the FGT, DGT, and the genotype-based HMM.

In this section, two main analyses were conducted. First (Section 1.5.2), dataset $\mathcal{D}_A$ was analysed to provide a comparison to Section 1.4 (page 15), where true shared haplotype information was used to validate the age estimation method. Second (Section 1.5.3), an extensive analysis was performed on datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$ to assess the impact of data error on age estimation. Note that the impact of phasing error was also evaluated in the latter, but which affected only the FGT.

The IBD detection methods used here were originally designed to infer shared haplotype segments in individuals sharing a focal allele. While this condition is fulfilled when considering concordant pairs, the IBD detection in discordant pairs is problematic. In the section below, I describe the modifications made to infer shared haplotypes in discordant pairs.

1.5.1 Modifications of IBD detection methods

Four-gamete test (FGT). The FGT is applied to the four haplotypes observed in two diploid individuals. A recombination event is inferred to have occurred between two variant sites if all four possible allelic configurations are observed. Let the focal site be denoted by b_i and another, distal site by b_j. In the four haplotypes, the alleles observed at (b_i, b_j) confirm a breakpoint if, for example, $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$ are observed, where 0 denotes the ancestral allelic state and 1 the derived state. Since breakpoints are inferred on both sides of a given focal variant, the genotypes at the focal site are both heterozygous in concordant pairs. But because the two individuals considered in a discordant pair do not share the focal allele, the required configuration cannot be observed.

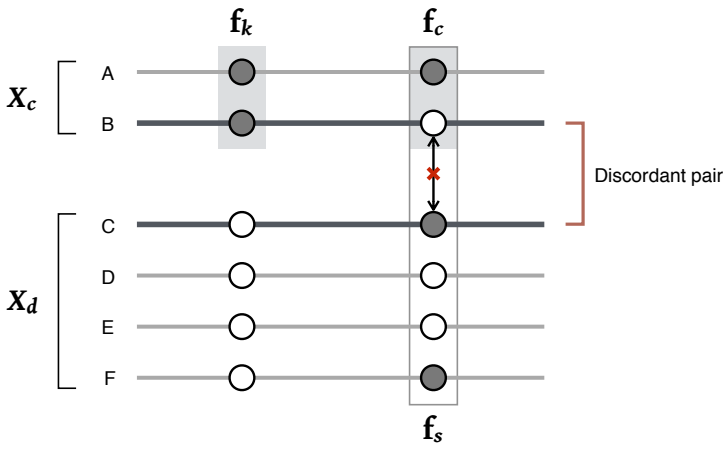


Figure 1.5: Breakpoint detection in discordant pairs. A discordant pair is formed by one haplotype from X_c (which share the focal allele) and one haplotype from X_d (which do not share the focal allele). The lines indicate the chromosomal sequence where the alleles at two sites are indicated; allelic states are distinguished as the ancestral (*hollow* circle) and derived state (*solid*). The conditions that lead to the detection of a recombination breakpoint is indicated between the focal site (*left*) and another, distal site (*right*), where f_k denotes the number of allele copies at the focal site within the subsample X_c, f_c denotes the number of allele copies observed at the distal site within the subsample X_c, and f_s denotes the number of allele copies at the distal site within the whole sample. The FGT is passed if all four allelic configurations are observed at four haplotypes in the sample.

However, breakpoints in discordant pairs can be detected as based on the allele frequencies observed in the sample. Let f_k denote the number of allele copies at focal site b_i. At a distal site, b_j, let f_c denote the number of allele copies observed only within the subsample X_c (who carry the focal allele at the focal position). Also, let f_s be the number of allele copies at b_j in the whole sample. A recombination breakpoint is indicated at b_j if the two haplotypes carry different alleles and if $f_c < f_k$ and $f_c < f_s$; additionally $f_s > 1$ to exclude singletons and $(f_s - f_c) > (2N - f_k)$ to exclude sites that are monomorphic within X_d, where $2N$ refers to the number of haplotypes in the sample. The condition implies the existence of the four allelic configurations at any of the haplotypes in the sample but is not bound by haplotype occurrence in two diploid individuals. The FGT thereby still holds but is practically inverted. An example is illustrated in Figure 1.5 (page 21).

Discordant genotype test (DGT). Recall that the DGT is a special case of the FGT which detects breakpoints at genotypic configurations that would also pass the FGT if haplotypes were available. Given the two heterozygous genotypes at the focal variant, a breakpoint is found at a distal site if opposite homozygous genotypes are observed. Again, in discordant pairs, such a configuration cannot be observed. The observation of opposite homozygous genotypes nonetheless implies that the two individuals do not share a haplotype at this site and is therefore also applied for breakpoint detection in discordant pairs.

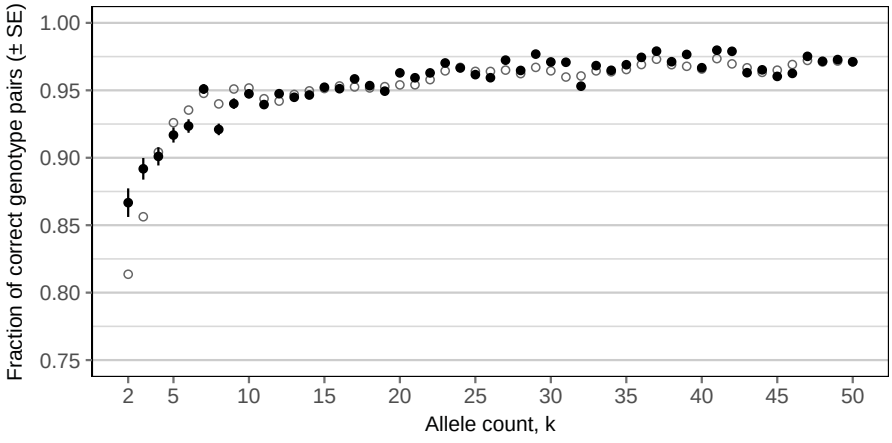


Figure 1.6: Initial state probability of discordant pairs in the Hidden Markov Model (HMM). The proportion of discordant pairs that were correctly identified by their genotypes was empirically determined from data before and after the inclusion of realistic genotype error rates. The mean per f_k was used as the initial state probability of the HMM-based approach for IBD detection around target sites. For comparison, the initial state probability of concordant pairs is shown (*hollow* circles).

Genotype-based Hidden Markov Model. The HMM includes a probabilistic model for observing each possible genotype pair in pairs of diploid individuals in *ibd* and *non*, which are the hidden states defined in the underlying IBD model; see Chapter 4. Both the emission and initial probabilities were determined empirically, from data before and after the inclusion of realistic genotype error rates.

The initial state probability corresponds to the probability of correctly observing a concordant pair through allele sharing, *i.e.* the true positive rate of observing heterozygous genotypes at a given target site where both individuals share the focal allele, which was determined per focal allele frequency (f_k). The empirical model was extended such that there was an additional initial state probability available and applied to discordant pairs. By comparing the data before and after error ($\mathcal{D}_B$ and $\mathcal{D}_B^*$), the initial state probability was determined as follows. For each f_k category, I randomly selected 1,000 target sites in the dataset "before error" for which I randomly selected 1,000 discordant pairs per target site. I then compared these genotypes to the corresponding genotypes in the dataset "after error" to determine the true positive rate. The mean per f_k was taken as the empirical initial state probability. The resulting distribution is shown in Figure 1.6 (page 22); the initial state probabilities used for discordant pairs are given for comparison. However, the initial state probability for the discordant case is similar to the concordant one. A possible explanation is that this is particularly driven by the heterozygous status being false.

The same empirical emission model was applied to discordant pairs as it follows from the coalescent (under the assumption of the infinite sites model) that the relationship at any site in the genome can be traced back to a common ancestor if looking back far enough. However, it must be noted that the current model was constructed to consider recent IBD. It can be expected that inference at discordant pairs is therefore less accurate.

Note that both the DGT and the HMM-based approach may operate on genotype data alone. Importantly, if haplotype information is not available, the sets X_c and X_d are formed by assigning all individuals that are heterozygous to X_c while all others are assigned to X_d, but excluding individuals that are homozygous for the focal allele. Since haplotype data are required to determine pairwise differences along haplotype sequences, $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$ cannot be used with genotype data. Here, analyses using the DGT and the HMM-based approach were performed on haplotypes although genotype data alone would suffice.

1.5.2 Comparison of IBD detection methods

Dataset $\mathcal{D}_A$ was used to compare the different IBD detection methods. Age was estimated for each of the three clock models, using a threshold of $n_d = 1{,}000$. The results presented in this section were obtained on the previously selected 10,000 rare allele target sites; see Section 1.4 (page 15). Again, the parameters of the age estimation method were specified according to simulation parameters ($N_e = 10{,}000$; $\mu = 1 \times 10^{-8}$ per site per generation; $\rho = 1 \times 10^{-8}$ per site per generation).

The density of true and inferred allele age is given in Figure 1.7 (next page). In all three methods, a tendency to overestimate allele age was seen, in particular under the mutation clock, $\mathcal{T}_\mathcal{M}$. This overestimation was elevated when the DGT was used, and less prominent for the FGT or HMM. The latter methods showed similar distributions in $\mathcal{T}_\mathcal{M}$ and under the combined clock model, $\mathcal{T}_{\mathcal{MR}}$, in which age appeared to be less overestimated. Under the recombination clock, $\mathcal{T}_\mathcal{R}$, alleles were underestimated in each method, but more severely in both the DGT and HMM.

Specifically, the method with the highest proportion of correctly estimated alleles was the FGT in all three clock models, where accuracy was highest under $\mathcal{T}_\mathcal{R}$ (72.6 %), followed by $\mathcal{T}_{\mathcal{MR}}$ (55.4 %) and $\mathcal{T}_\mathcal{M}$ (34.5 %). The HMM achieved similar proportions, but which was low in $\mathcal{T}_\mathcal{R}$ (10.950 %) compared to $\mathcal{T}_{\mathcal{MR}}$ (51.876 %) and $\mathcal{T}_\mathcal{M}$ (32.415 %). Throughout, the lowest proportions were found for the DGT (14.554 %, 8.226 %, and 29.659 % for $\mathcal{T}_\mathcal{M}$, $\mathcal{T}_\mathcal{R}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively).

Table 1.2: Estimation accuracy per IBD detection method. The accuracy was measured in analyses based on IBD detected by different methods; namely the FGT, DGT, and the HMM-based approach. See Table 1.1 (page 18) for comparison to results obtained using true IBD segments (for $n_d = 1{,}000$).

Clock	Method	Rank correlation (r_S)			RMSLE		
		t_c	t_m	t_d	t_c	t_m	t_d
$\mathcal{T}_\mathcal{M}$	FGT	**0.841**	**0.839**	**0.686**	**1.011**	**0.653**	**0.554**
	DGT	0.830	0.813	0.650	1.460	1.086	0.832
	HMM	0.806	0.806	0.662	1.078	0.725	0.607
$\mathcal{T}_\mathcal{R}$	FGT	**0.899**	**0.887**	**0.718**	**0.339**	**0.330**	**0.775**
	DGT	0.820	0.749	0.554	0.577	0.941	1.396
	HMM	0.821	0.751	0.556	0.533	0.892	1.348
$\mathcal{T}_{\mathcal{MR}}$	FGT	**0.863**	**0.873**	**0.723**	**0.755**	**0.422**	**0.524**
	DGT	0.840	0.829	0.669	1.083	0.727	0.600
	HMM	0.826	0.834	0.692	0.806	0.485	0.554

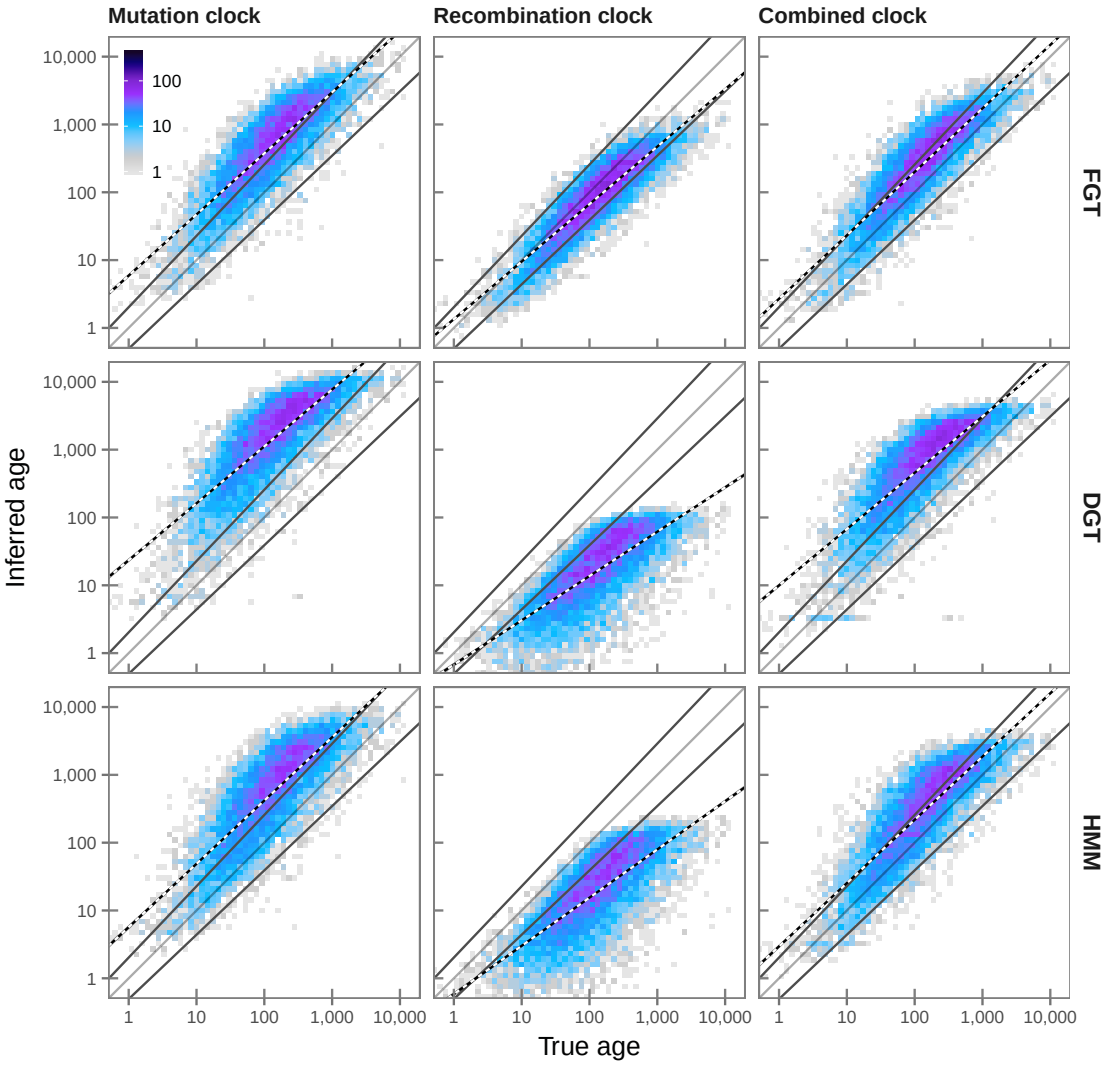


Figure 1.7: Distribution of true and inferred age using different IBD detection methods. The three IBD detection methods FGT, DGT, and HMM were compared under each clock model, on the same set of target sites that were drawn from $f_{[2,20]}$ variants (allele frequency $\leq 1\%$) in $\mathcal{D}_A$. Each panel shows the density of true age (t_m) and inferred age. Lines *below* and *above* the dividing line are regression trend lines of the corresponding true coalescent times around each mutation event, t_c and t_d, respectively. The regression of inferred age ($\hat{t}$) is given by the *black-white* line.

Summary metrics for each analysis are given in Table 1.2 (page 24). Throughout, the FGT showed a higher accuracy compared to the other IBD detection methods under each clock model. Relative age estimates are shown for distinct f_k ranges in Figure 1.8 (next page), where the relative age for corresponding results obtained by using true IBD information is given for comparison per clock model; see Figure 1.4 (page 19). Analyses under $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$ showed a substantial difference between alleles at lower and higher

frequencies; *e.g.* overall accuracy of $f_{[2,5]}$ variants was increased compared to f_k variants at higher frequencies in each method. This difference was reduced under $\mathcal{T}_{\mathcal{R}}$, but the DGT showed an accuracy decrease for $f_{[2,5]}$ variants.

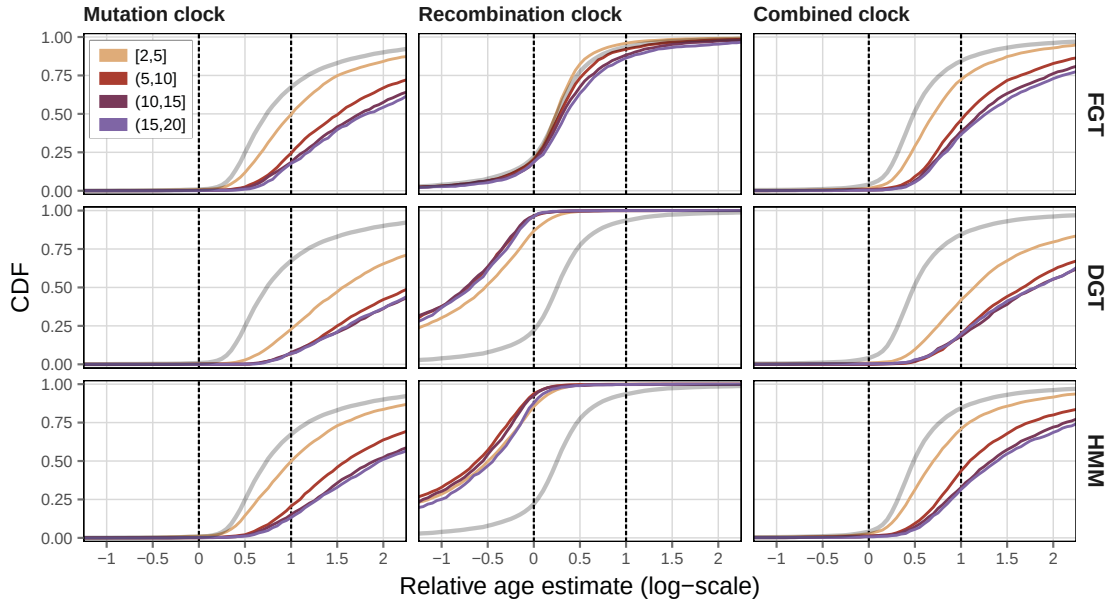


Figure 1.8: Relative age using different IBD detection methods. The three IBD detection methods implemented in `rvage` were compared, *i.e.* FGT, DGT, and HMM (indicated at the *right* of each row), under each clock model (indicated at the *top* of each column). The relative age, $\hat{t}_{rel}$, was calculated as given in Equation (1.26), such that t_c and t_d sit at 0 and 1 (*dashed* lines). The CDF of relative age estimates is shown for different frequency ranges; namely $f_{[2,5]}$, $f_{(5,10]}$, $f_{(10,15]}$, and $f_{(15,20]}$. The *grey* line provides a comparison to age estimated using true IBD information as shown in Figure 1.4 (page 19), but for $f_{[2,20]}$.

These results suggested that the accuracy of estimated allele age is crucially dependent on correct inference of the underlying IBD structure. The clock models behave differently when the length of an IBD segment is over or underestimated. It can be expected that $\mathcal{T}_{\mathcal{M}}$ may indicate an older allele age if IBD length is overestimated, due to potentially including a larger number of mutational differences which suggests an older T_{MRCA}. Conversely, $\mathcal{T}_{\mathcal{R}}$ may indicate a younger age, because a more recent T_{MRCA} is suggested when IBD length is relatively long.

I found that the FGT was the best performing method for the targeted detection of IBD segments, as the accuracy of estimated age was similar to the expectations defined by true IBD information in Section 1.4. However, the estimation was more accurate for target sites at lower allele frequencies. The DGT was least accurate in terms of estimated allele age in this comparison.

Recall that the probabilistic model of the HMM was developed to overcome the effects of genotype error encountered in real data (see Chapter 4). Thus, the results in this section reflect theoretical limitations of age estimation given IBD detected in flawless data, but may change drastically in presence of genotype error. This was explored in the section below.

1.5.3 Impact of genotype error on allele age estimation

Allele age was estimated using datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$, to compare the accuracy of the estimation method before and after error. Shared haplotype inference was performed using the FGT, DGT, and the genotype-based HMM. In addition, age was estimated using true IBD information as determined from simulation records.

In total, 5,000 target sites were randomly selected at allele frequency $\leq 0.5\%$ ($f_{[2,25]}$). Note that these were sampled from the subset of variants unaffected by error in $\mathcal{D}_B^*$, to ensure that alleles correctly identified haplotype sharing. A threshold of $n_d = 2{,}500$ was applied to randomly select concordant pairs at a given target site. Note that statistically phased data was available for both $\mathcal{D}_B$ and $\mathcal{D}_B^*$, which were included here to assess the impact of phasing error, but which can only affect the FGT.

1.5.3.1 Age estimation using true shared haplotype information

First, estimation based on the true IBD structure of the sample is compared before and after error. Results are shown in Figure 1.9a (next page). The most striking discovery is the extent of overestimation after error under the mutation clock model, $\mathcal{T}_{\mathcal{M}}$, which was similarly high in the combined clock, $\mathcal{T}_{\mathcal{MR}}$. It is suggested that age was overestimated due to misclassified alleles which may have substantially increased the number of observed mutational differences observed within the shared haplotype interval.

Rank correlation decreased in $\mathcal{T}_{\mathcal{M}}$ from $r_S = 0.870$ to $r_S = 0.518$ with regard to t_c, before and after error. This was similar in $\mathcal{T}_{\mathcal{MR}}$, where r_S at t_c decreased from 0.884 to 0.593, respectively. The proportion of correctly estimated alleles ($t_c < \hat{t} < t_d$) in $\mathcal{T}_{\mathcal{M}}$ was 75.4 % before and 24.1 % after error, which was similar in $\mathcal{T}_{\mathcal{MR}}$, where 80.5 % of alleles were correct before but only 39.4 % after error.

The estimation under the recombination clock model, $\mathcal{T}_{\mathcal{R}}$, was not affected by genotype error, due to using true IBD information to derive segment lengths. Note that analyses were performed on the same sets of concordant and discordant pairs, which is why the results in $\mathcal{T}_{\mathcal{R}}$ are identical before and after error. Allele age showed a tendency to be underestimated in $\mathcal{T}_{\mathcal{R}}$. Overall, 42.891 % of alleles were correctly inferred, and rank correlation was relatively high (r_S: 0.818, 0.843, and 0.666 at t_c, t_m, and t_d, respectively).

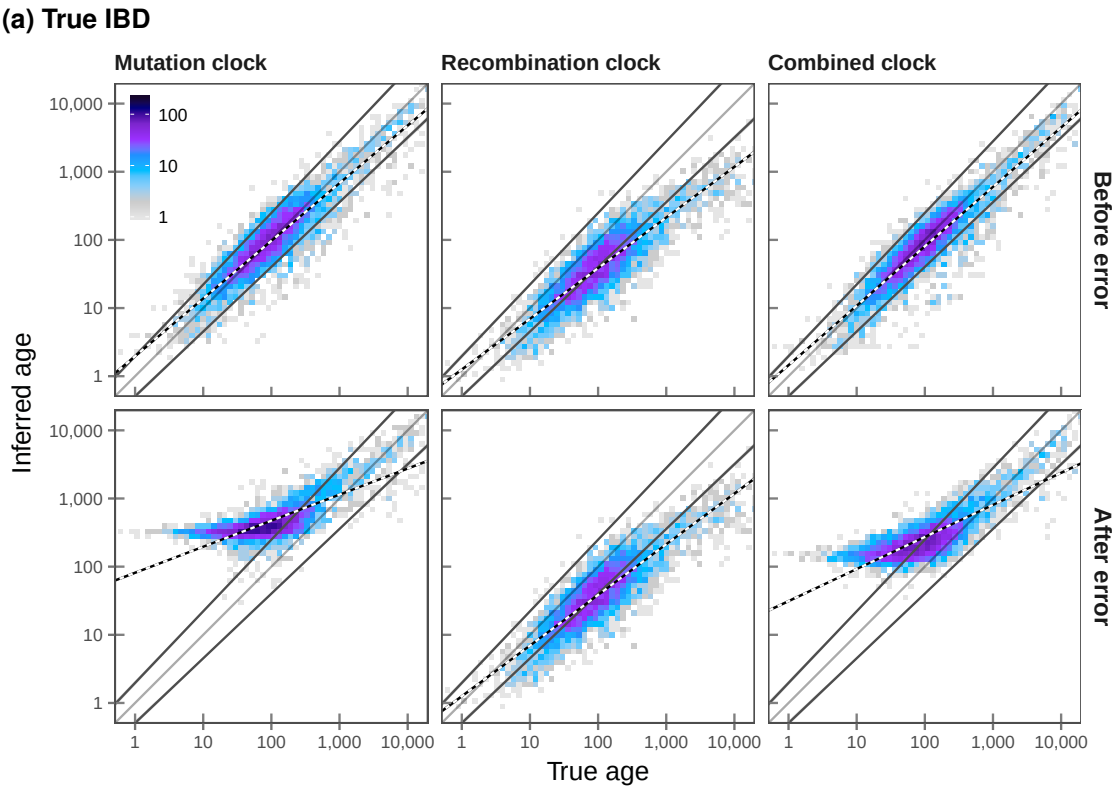
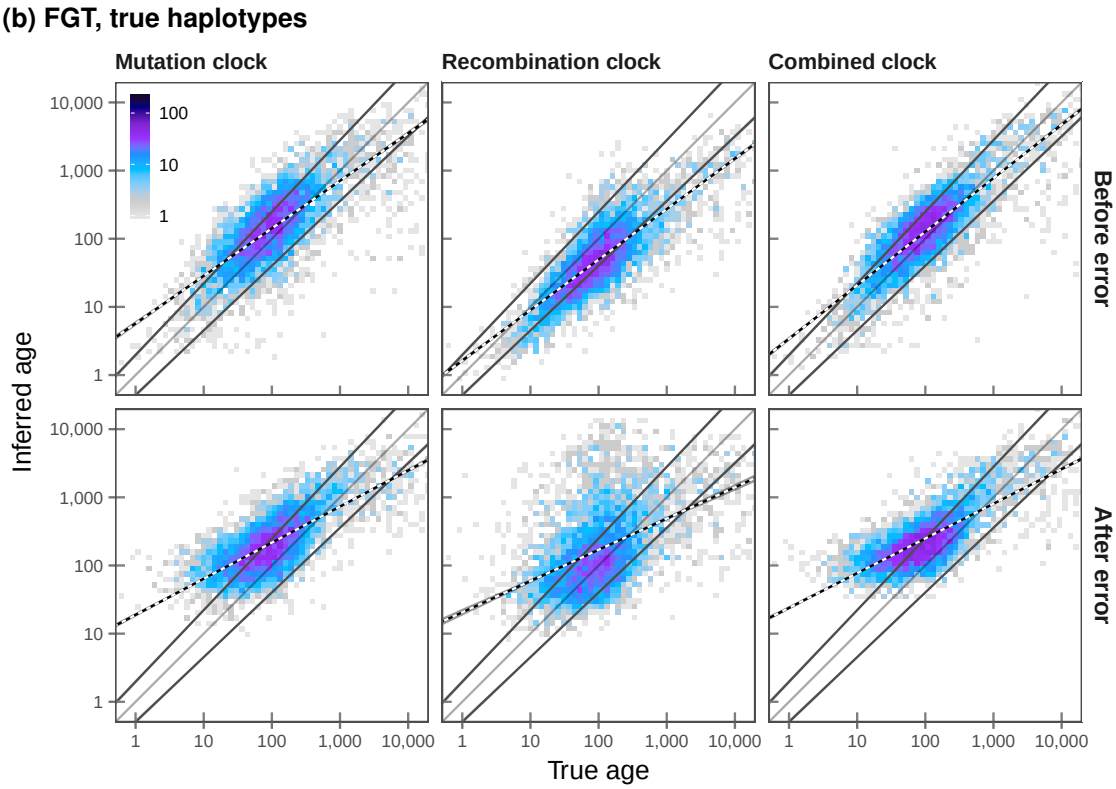


Figure 1.9: Density of allele age before and after error in simulated data. The effects on the estimation process *before* and *after* error are compared. Note that the "true age" was set to t_m, which is the geometric mean of t_c and t_d. Lines *below* and *above* the dividing line correspond to the regression lines over t_c and t_d; *i.e.* of the times of coalescent events delimiting the branch on which a focal mutation occurred. The *black-white* line gives the regression for the inferred age ($\hat{t}$). This panel (**a**) compares the distributions of true and inferred ages, which were estimated on basis of the true IBD structure of the sample as determined from simulation records. The other panels show estimation results based on the different IBD detection methods; FGT on both true and phased haplotypes (**b**, **c**; next page), DGT (**d**; page 30), and the genotype-based HMM (**e**; page 31). Each analysis was conducted on the same set of 5,000 randomly selected target variants at $f_{[2,25]}$.

1.5.3.2 Age estimation based on inferred shared haplotypes

The different IBD detection methods are compared below. Results based on the FGT are shown in Figure 1.9b and 1.9c (next page), which show age estimates based on IBD detected in true (simulated) and phased haplotype data, respectively, both before and after error.

Before error, 53.021 %, 50.847 %, and 60.040 % of alleles were correctly inferred from true haplotype data in $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively. When phased data were used, this changed only slightly; 50.828 %, 51.366 %, and 59.182 % in $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively. Notably, the proportion of correctly inferred alleles increased in $\mathcal{T}_{\mathcal{R}}$ due to phasing error. It is suggested that the tendency for underestimation that was generally seen in $\mathcal{T}_{\mathcal{R}}$ may have been mitigated by further reduction of IBD segment lengths resulting from



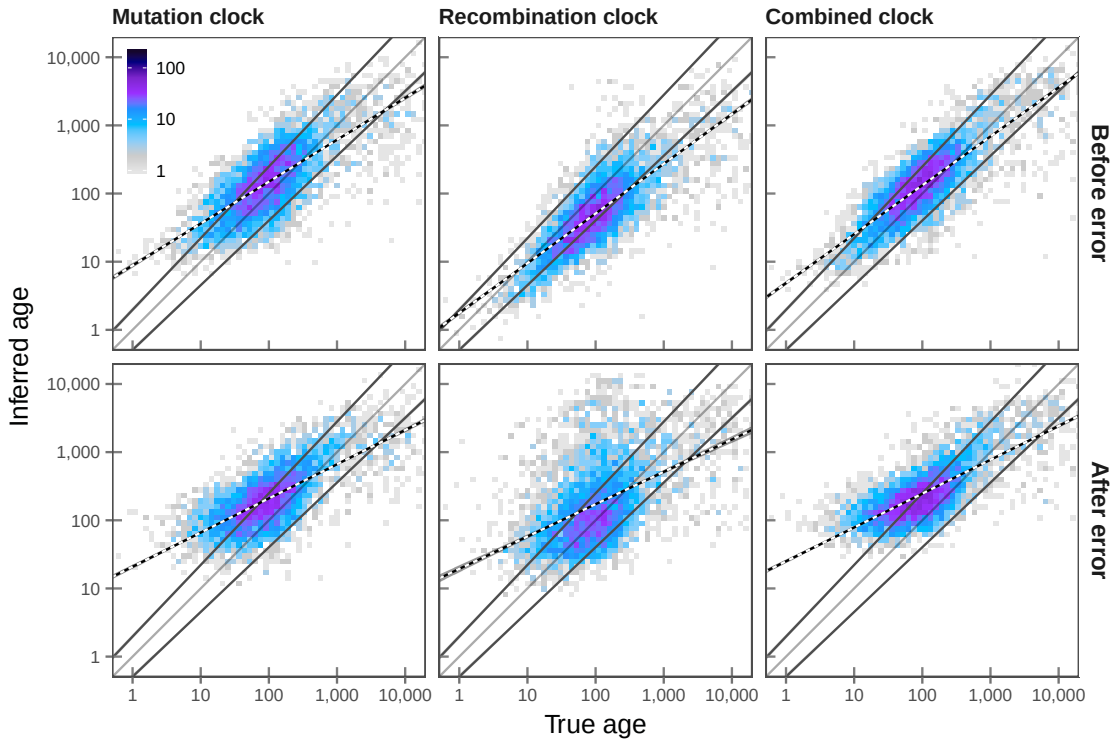


Figure 1.9: Continued.

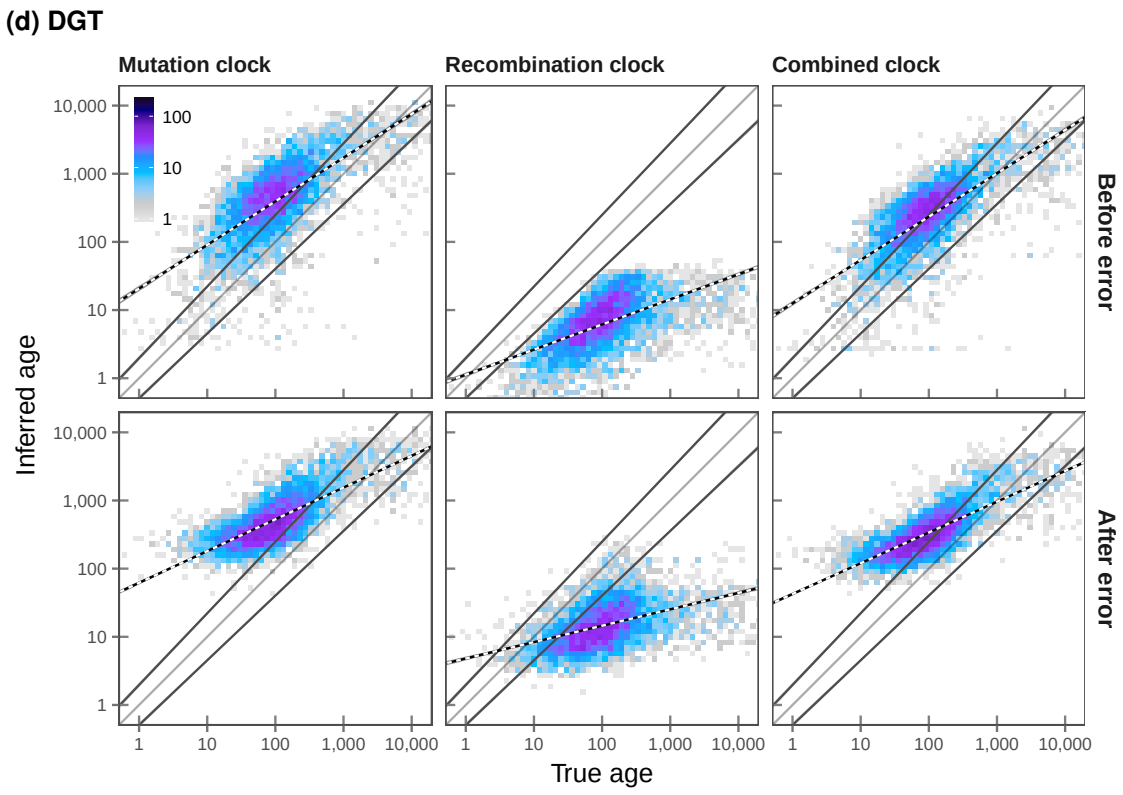


Figure 1.9: Continued.

phasing error. The small difference between true and phased data was further reflected in r_S, which changed from 0.680 to 0.660 in $\mathcal{T}_{\mathcal{M}}$, 0.780 to 0.764 in $\mathcal{T}_{\mathcal{R}}$, and 0.742 to 0.731 in $\mathcal{T}_{\mathcal{MR}}$, with regards to t_d.

After error, the overall proportion of correct alleles was reduced, but again the differences between true and phased data were small. On true haplotypes, the proportion of correct alleles was 44.267 %, 45.025 %, and 42.034 % in $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively, whereas 43.549 %, 46.002 %, and 41.635 % of alleles were correct using phased haplotypes. Likewise, r_S and RMSLE scores did not suggest notable differences between estimation results from true and phased haplotypes; see Table 1.3 (page 32).

In the previous analysis using true IBD, it was suggested that genotype error may induce an overestimation of allele age in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$. However, this was reduced here, possibly because phasing error may result in truncated shared haplotype intervals, such that shorter intervals may mitigate the effects of data error on observed pairwise differences in a pair haplotypes.

Estimation results based on the DGT are shown in Figure 1.9d (this page). Before error, the proportions of correctly inferred alleles were the lowest in the present comparison in each clock model. For $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$, the estimation resulted in 26.341 % and 36.949 %

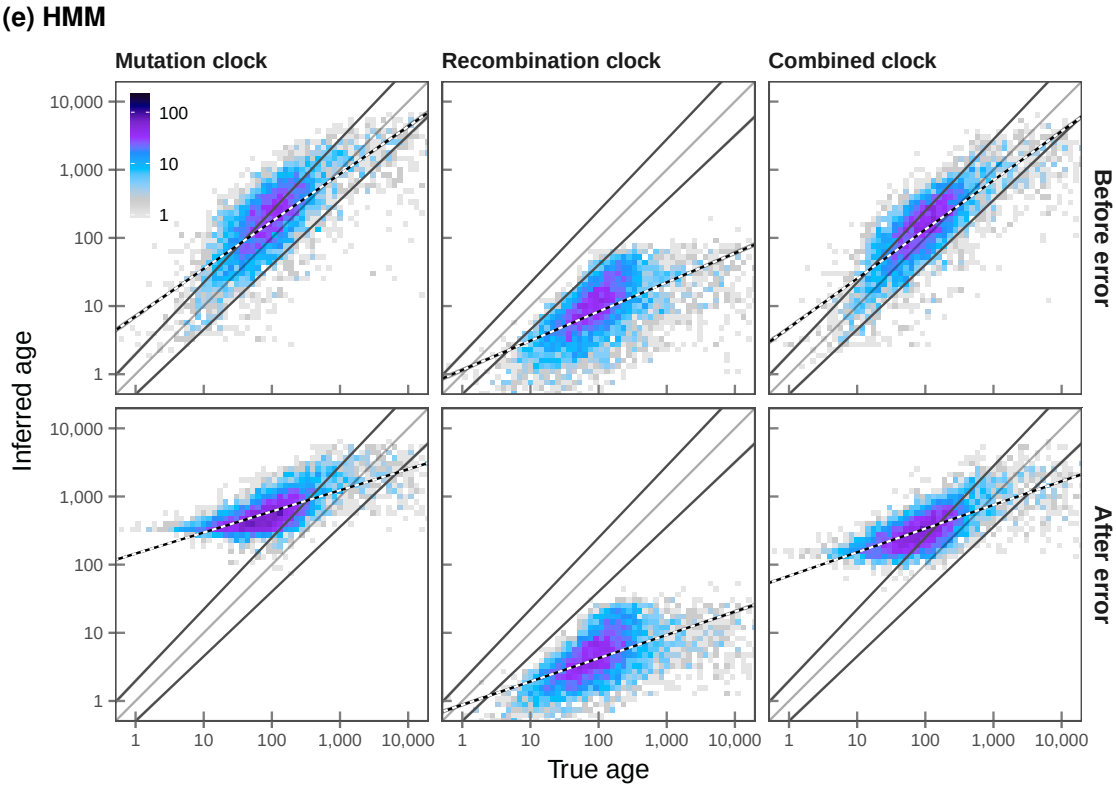


Figure 1.9: Continued.

of correct alleles, respectively, whereas only 2.413 % were correct for $\mathcal{T}_{\mathcal{R}}$. The tendency to overestimate allele age was increased after error. This was also seen for $\mathcal{T}_{\mathcal{R}}$, where the proportion of correctly inferred alleles increased to 15.693 %, but at a loss of accuracy. Rank correlation, r_S, was 0.746, 0.628, and 0.406 at t_c, t_m, and t_d before error, and 0.588, 0.504, and 0.328 after error; see Table 1.3 (next page).

The accuracy of age estimation using the genotype-based HMM was relatively high before error; that is, more accurate in comparison to the FGT in $\mathcal{T}_{\mathcal{M}}$, similar in accuracy to the DGT in $\mathcal{T}_{\mathcal{R}}$, and similar to the FGT in $\mathcal{T}_{\mathcal{MR}}$. The density of inferred allele age based on the HMM is given in Figure 1.9e (this page).

Before error, the proportion of correct alleles was 47.537 % in $\mathcal{T}_{\mathcal{M}}$, 3.629 % in $\mathcal{T}_{\mathcal{R}}$, and 57.827 % in $\mathcal{T}_{\mathcal{MR}}$. Allele age was generally underestimated in $\mathcal{T}_{\mathcal{R}}$ (96.351 %). This was increased after error, resulting in an underestimated proportion of 98.305 % in $\mathcal{T}_{\mathcal{R}}$, as the proportion of correct alleles was overall reduced; 16.650 % and 27.657 % in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$, respectively. Also, RMSLE scores were lowest for the HMM under each clock model after error; see Table 1.3 (next page). Rank correlation before and after error, for r_S at t_c, decreased from 0.702 to 0.535 in $\mathcal{T}_{\mathcal{M}}$, and from 0.733 to 0.569 in $\mathcal{T}_{\mathcal{MR}}$. Although rank correlation for the HMM was high in $\mathcal{T}_{\mathcal{R}}$, *e.g.* r_S at t_c was 0.751 before and 0.737 after error, estimation bias was substantial both before and after error.

Table 1.3: Effect of genotype error on age estimation accuracy. Allele age was estimated based on IBD inferred using the FGT, DGT, and HMM on the same set of 5,000 rare allele target sites randomly selected at allele frequency ≤ 0.5% ($f_{[2,25]}$) in simulated data before and after error (datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$). The number of discordant pairs was limited to n_d = 2,500 in each analysis. True IBD refers to age estimation conducted using knowledge of the actual shared haplotype structure of the sample, as determined from simulation records. CORRECTED

Clock	Method	Before error			After error		
		t_c	t_m	t_d	t_c	t_m	t_d
Rank correlation coefficient (r_S)							
$\mathcal{T}_{\mathcal{M}}$	FGT*	0.680	0.736	0.597	0.556	0.696	0.615
	FGT**	0.660	0.711	0.576	0.543	0.673	0.591
	DGT	0.618	0.685	0.563	**0.577**	**0.724**	**0.649**
	HMM	**0.702**	**0.738**	**0.599**	0.535	0.686	0.621
	True IBD	0.870	0.871	0.673	0.518	0.694	0.646
$\mathcal{T}_{\mathcal{R}}$	FGT*	**0.780**	**0.782**	0.601	0.405	0.481	0.407
	FGT**	0.764	0.780	**0.603**	0.406	0.485	**0.414**
	DGT	0.746	0.628	0.406	0.588	0.504	0.328
	HMM	0.751	0.632	0.411	**0.737**	**0.621**	0.398
	True IBD	0.818	0.843	0.666	0.818	0.843	0.666
$\mathcal{T}_{\mathcal{MR}}$	FGT*	**0.742**	**0.792**	**0.644**	0.528	0.689	0.629
	FGT**	0.731	0.787	0.643	0.520	0.679	0.619
	DGT	0.666	0.727	0.597	**0.596**	**0.757**	**0.689**
	HMM	0.733	0.781	0.641	0.569	0.693	0.606
	True IBD	0.884	0.885	0.696	0.593	0.735	0.655
Root mean squared logarithmic error (RMSLE)							
$\mathcal{T}_{\mathcal{M}}$	FGT*	**0.696**	**0.436**	0.639	0.864	**0.516**	**0.524**
	FGT**	0.715	0.444	**0.623**	**0.859**	0.524	0.547
	DGT	1.083	0.743	0.657	1.190	0.809	0.606
	HMM	0.754	0.478	0.633	1.250	0.882	0.681
	True IBD	0.454	0.255	0.666	1.146	0.770	0.587
$\mathcal{T}_{\mathcal{R}}$	FGT*	**0.380**	**0.471**	0.909	0.881	**0.638**	0.728
	FGT**	0.413	0.480	**0.903**	0.890	0.641	**0.722**
	DGT	0.905	1.252	1.690	**0.703**	0.991	1.413
	HMM	0.796	1.141	1.585	1.031	1.380	1.814
	True IBD	0.337	0.504	0.960	0.337	0.504	0.960
$\mathcal{T}_{\mathcal{MR}}$	FGT*	**0.624**	**0.364**	0.626	**0.915**	**0.548**	**0.496**
	FGT**	0.641	0.367	**0.608**	0.916	0.551	0.503
	DGT	0.869	0.557	0.611	1.019	0.645	0.523
	HMM	0.644	0.398	0.647	1.021	0.672	0.585
	True IBD	0.381	0.260	0.716	0.919	0.555	0.506

* FGT applied to true haplotypes
** FGT applied to phased haplotypes

1.5.4 Discussion

I demonstrated the validity of the age estimation framework using simulated data where I showed that age can be estimated with very high accuracy. However, certain problems may arise when working with real data. Notably, ihe impact of phasing error is small in comparison to genotypic (or allelic) misclassification, which is likely to bias the estimation process.

Generally, imperfect data may affect the estimation of allele age in two ways. First, the method was shown to be highly susceptible to inaccurate IBD inference, where each clock model behaves differently to the over or underestimation of IBD length. However, second, even for a method that can infer shared haplotype segments with high accuracy, the alleles observed at a focal site in the sample may wrongly identify haplotype sharing. To account for the possibility that some concordant pairs may actually be discordant pairs (or *vice versa*), for example, a separate filtering method would be needed to exclude pairs based on patterns of the inferred haplotype structure. But because such a method would effectively predict falsely called or typed alleles in the data, a solution to this problem may not be straightforward.

In conclusion, a substantial amount of estimation bias was seen for any of the evaluated methods for shared haplotype inference. Due to the dependency on finding genuine haplotype intervals, the allele age estimation method may therefore not be regarded as being reliable in applications to real data. However, a solution is attempted in the following section, where I present a novel haplotype-based HMM as an advancement over of the current genotype-based method.

REMOVED Section “Generation of error correction models”

REMOVED Section “Age of alleles with predicted effects in 1000 Genomes data”

ADDITION All sections below were added to include new results as presented in the viva

1.6 A haplotype-based HMM for shared haplotype inference

As shown in the previous sections, the allele age estimation method was able to infer the time of mutation events with high accuracy, but where it was suggested that this is largely dependent on (ideally) correct inference of the underlying shared haplotype

structure around a given target site. The inference methods presented so far were either rule-based (FGT and DGT) and thus highly susceptible to data error, or probabilistic but genotype-based (HMM) where accuracy may easily be influenced by the shared haplotype structure at the "unshared" haplotypes in the individuals considered. Although the latter was implicitly impervious to data error due to phasing, it was shown that such errors were less problematic on average, but where error due to misclassification of alleles may be regarded as the main source of estimation bias.

In this section, I present a novel haplotype-based HMM for the targeted and pairwise detection of shared haplotypes. This method can be seen as the conclusion of the previous genotype-based approach and is therefore algorithmically defined in a similar way. I describe the model in the section below. Note that I implemented additional modifications in the age estimation method, which I describe in Section 1.6.2 (page 39).

As done previously, I evaluated the haplotype-based HMM in terms of the resulting age estimation accuracy in data before and after error (Section 1.6.3, page 40). The method was further compared to the Pairwise Sequentially Markovian Coalescent (PSMC) in terms of the inferred T_{MRCA} and subsequent estimation of allele age (Section 1.6.4, page 44). Lastly, I briefly present age estimation results from an analysis using 1000 Genomes Project (1000G) Phase III data (Section 1.6.5, page 52).

1.6.1 Description of the model

The general algorithm and model specifications of the HMM presented here follow the previously described genotype-based HMM; see **??** (page ??). Briefly, at a given target position in the genome, the allelic sequence of two selected haplotypes is independently traversed to the left and right-hand side from that position, until the end of the chromosome is reached. As before, two hidden states are assumed; *ibd* and *non*. The observation states are defined as the possible allelic pairs $(0, 0)$, $(0, 1)$, and $(1, 1)$, where the order of alleles is not relevant; *i.e.* observing $(0, 1)$ in haplotypes A and B is equivalent to observing $(1, 0)$. The model is illustrated in Figure 1.10 (next page), where φ denotes a transition parameter and the probabilities of emission from *ibd* are denoted by $\delta_{h_A h_B}$ and from *non* by $\eta_{h_A h_B}$. Model transitions, emissions, and initial state probabilities are described below.

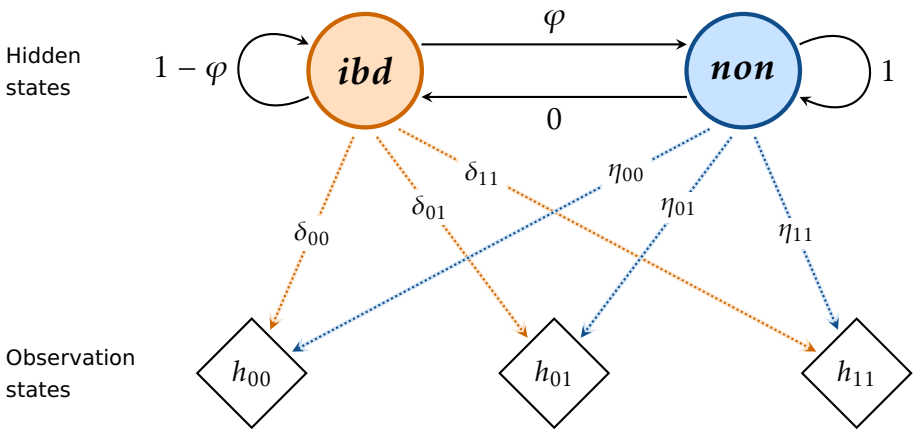


Figure 1.10: Illustration of the haplotype-based HMM for shared haplotype inference. Two hidden states are assumed to generate the observations in a Markov process; *ibd* and *non*. Transitions from each state into any state are indicated by *solid* lines. The probability of transition from *ibd* to *non* is denoted by φ, and from *non* to *ibd* is set to zero; hence, once the chain proceeds into the *non* state it cannot go back into *ibd*. This is because the process is modelled such that only the innermost segment is inferred, relative to the focal position which sits at the start of the sequence. The input sequence consists of sequence data from a pair of haplotypes, resulting in three possible observation states; denoted by $h_{h_A h_B}$, where $h_A, h_B \in \{0, 1, \}$. The probabilities of emitting each possible haplotype pair given each hidden state are denoted by $\delta_{h_A h_B}$ and $\delta_{h_A h_B}$ for *ibd* and *non*, respectively; indicated by the *dotted* lines. The direction of arrows indicates conditional dependence; *i.e.* the transition from one hidden state into another state, or emission of a genotype pair while being in *ibd* or *non*.

Transition probabilities. The genetic distance to a recombination event follows the Exponential distribution when measured in population-scaled time units. The probability of transition from *ibd* to *non* is modelled as

$$\varphi = 1 - \mathrm{e}^{-2N_e r_j \frac{\xi_k}{2}} \tag{1.27}$$

where j indicates the position at the current site in the sequence and k indicates the the focal position. The value of r_j is the genetic distance between the current and the Immediately previous site observed in the sequence. The value of ξ_k is the expectation of the age of the focal allele after Kimura and Ota (1973), calculated using **??** on page ??. The probability of remaining in *ibd* is $1 - \varphi$. The model employs a "left-to-right" architecture, where transition from *non* to *ibd* have zero probability; *i.e.* once *ibd* has been left the sequence stays in *non* with probability equal to one. The intuition and shortcomings of such a transition model were discussed in **??** (page ??).

Emission probabilities. An empirical emission model was generated from simulated haplotype data, both before and after the integration of empirically determined error rates. In the previous genotype-based model, the empirical rate of observing possible

genotype pairs was determined from haplotype segments shared between individuals who carried any allele observed at low frequency ($f_{[2,25]}$). The same was done here, but such that the empirical probabilities were measured based on the T_{MRCA} between haplotype pairs. A full representation of observation rates found by T_{MRCA} and allele frequency is given in Figure 1.11 (this page). Note that, again, datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$ were used to obtain emission models before and after error.

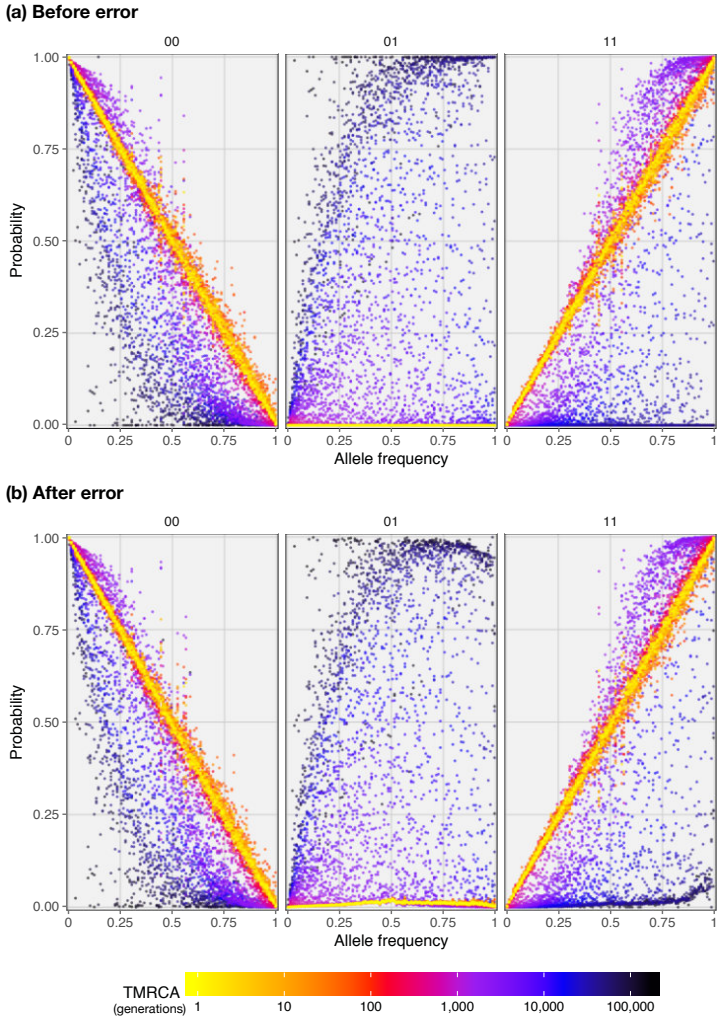


Figure 1.11: Empirical probability to observe allelic pairs dependent on T_{MRCA}. Using data before and after error (datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$), the rate at which the possible allelic pairs $(0, 0)$, $(0, 1)$, and $(1, 1)$ were observed is shown by allele frequency, distinguished by the T_{MRCA} between two haplotypes at a given site. Empirical observation rates were measured at 100,000 randomly selected haplotype pairs. The resulting rates were averaged over 100 equally sized allele frequency bins and 100 time intervals of equal size on log-scale. Panels **(a)** and **(b)** show the empirical distribution before and after error, respectively.

However, the generation of the emission model was not straightforward as any pair of haplotypes is "identical by descent" at any position in the genome; according to the coalescent and the assumptions of the infinite sites model. The definition of the hidden

states as being "in IBD" and "not in IBD" may therefore be seen as arbitrary. Nonetheless, to generate empirical probabilities, I defined a nominal "cutoff" for the T_{MRCA} at 100 generations to calculate the mean empirical observation rates for each state. The average rate seen ≤ 100 generations was taken for *ibd*, and > 100 generations for *non*, at each allele frequency bin (rates were averaged over 100 equally sized bins). Linear interpolation was used to obtain rates at sites with frequencies not captured by the model. The resulting model is illustrated in Figure 1.12 (this page).

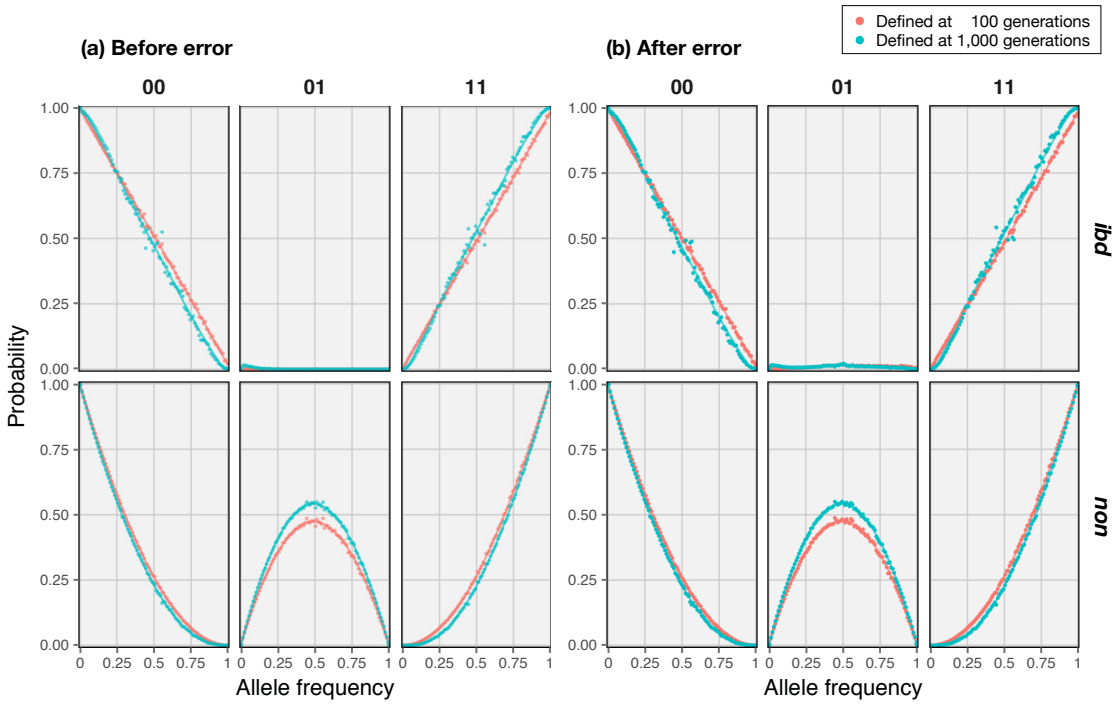


Figure 1.12: Empirical emission model used in the haplotype-based HMM. The emission probabilities used in the haplotype-based HMM were determined empirically for each possible allelic pair, using datasets $\mathcal{D}_B$ (before error) and $\mathcal{D}_B^*$ (after error). To distinguish *ibd* from *non*, a nominal cutoff was applied to the result shown in Figure 1.11 (page 36). For comparison, the resulting distributions for a T_{MRCA} cutoff at 100 generations (*red*) are compared to a cutoff at 1,000 generations (*blue*). Note that the empirical model employed here used a cutoff at 100 generations.

One caveat of applying such a cutoff is that the model is implicitly conditioned on observations deriving from relatively recent coalescent events. But, notably, when comparing cutoffs at 100 and 1,000 generations, differences between averaged distributions were relatively small; as shown in Figure 1.12. In the following, the emission models generated from dataset $\mathcal{D}_B$ were used in all analyses of error-free data; otherwise, the model generated from $\mathcal{D}_B^*$ was used.

Initial state probabilities. The estimated true positive rate of observing allelic pairs in data before and after error was taken as an empirical model of the initial state probabilities. Note that for error-free data implies the initial probability of being in *ibd* is equal to one, as it is certain that a focal haplotype pair shares an allele by descent at a given target site; given the assumptions of the infinite sites model. Again, datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$ were used.

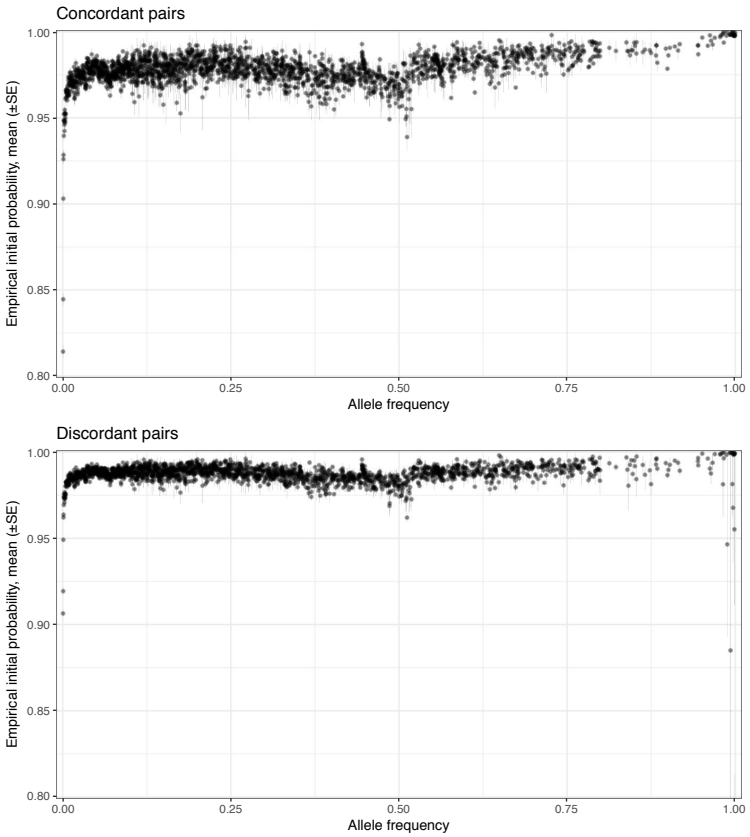


Figure 1.13: Empirical initial state probabilities used in the haplotype-based HMM. The rate of correctly observing allelic combinations in pairs of haplotypes was measured by comparing data points before and after error, using simulated datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$. At a given focal site, haplotypes were sorted into concordant and discordant pairs as determined from allele sharing in the sample before error. The allelic state at each pair was then compared to the same location and pair in data after error, to measure the true positive rate of observing the $(1, 1)$ allelic combination in concordant pairs and $(0, 1)$ in discordant pairs. Rates are shown as the mean for observations at a given focal allele count (±SE). The number of focal sites observed at a specific frequency in the sample differed along the site frequency spectrum, but was capped at a maximum of 1,000 randomly sampled sites found at a given allele count.

Empirical rates were estimated as follows. At a given site, the sample was distinguished into carrier and non-carrier haplotypes for the allele at that site to form all possible concordant and discordant pairs. Each pair was analysed in turn where allele combinations were compared before and after error to establish the rate at which alleles were observed

correctly at that site and that pair. That is, allelic combinations $(1, 1)$ in concordant pairs and $(0, 1)$ or $(1, 0)$ in discordant pairs. This was done at each value of the allele count in the simulated dataset (N = 5,000 haplotypes), but where a maximum of 1,000 randomly selected sites was analysed from the sites found at the same allele count. Measured rates were then averaged per bin; the result is shown in Figure 1.13 (page 38). When applied to data of different size N, rates were linearly interpolated based on allele frequency.

1.6.2 Modifications of the age estimation method

I attempted to improve the allele age estimation method as a result of the lessons learned from the analyses conducted in previous sections. Note that the implemented modifications entail (minor) changes to the algorithm, but where model specifications remained untouched.

1.6.2.1 Nearest neighbour selection of discordant pairs

Discordant pairs are formed by taking one haplotype from set X_c (*carriers*) and one from set X_d (*non-carriers*). The intuition is that discordant pairs can be used to to indicate the time of coalescent events above the the point in time at which the focal mutation occurred (back in time), such that the "upper limit" of the branch can be found. It would be beneficial to select those haplotypes from X_d that are the nearest neighbours to the sub-tree spanned by the lineages leading to X_c. Otherwise, if selection of pairs is random, the number of discordant pairs required to include nearer neighbours may be relatively high on average, dependent on the composition of the sample.

I implemented a simple approach to calculate the Hamming distance in each discordant pair in the sample, along a fixed region around the focal site. Discordant pairs with the lowest distance are prioritised. Here, the first 5,000 sites to the left and right-hand side of the focal position were queried.

Importantly, in presence of data error, it is likely that false negatives (missed focal alleles) would be selected preferentially. To reduce the chance of selecting false negatives, I used a "relaxed" nearest neighbour approach. Discordant pairs are first sorted by their distance (lowest to highest) and each pair is scanned in turn. I then identify X_d haplotypes that are repeated more than once along the queue of pairs and place these at the end of the queue. The first n_d pairs are then retained, where n_d can be specified.

The implementation of this algorithm substantially improved computation time due to retaining a smaller number of pairs. The results presented below were obtained using only $n_d = 100$ discordant pairs (as opposed to several hundreds or thousands of randomly

selected pairs). Note that a nearest neighbour approach could also be applied to select concordant pairs (prioritising higher distances), but which was not done here, as it is assumed that random sampling is sufficient to retain pairs that indicate coalescent events close to the time of the focal mutation event.

1.6.2.2 Restriction of counting pairwise differences in concordant pairs

Misclassification of genotypes or alleles is expected to adversely affect estimation accuracy when the mutation clock or the combined clock model is used. False negatives (missed alleles) and false positives (wrongly called or typed alleles) if not consistent in both haplotypes may artificially inflate the number of pairwise differences seen along the inferred shared haplotype region.

I implemented a simple rule to restrict the count of pairwise differences. At a given site along the sequence, an observed difference is only counted if its frequency in the sample is less than or equal to the frequency of the focal allele. Thereby, allelic differences are restricted to the mutations that have occurred more recently than the focal mutation event; that is, within the sub-tree deriving from the MRCA of carrier haplotypes. This assumes the infinite sites model; *i.e*. excluding back-mutations or recurrent mutations.

The implementation of this rule substantially improved estimation accuracy in presence of data error, and did not affect accuracy when data was error-free (simulated haplotypes); note that these results are not shown for brevity. The restriction rule was used in the analysed presented below.

A similar rule could be implemented for discordant pairs. However, the length of share haplotype segments is expected to be shorter, reducing the possibility to encounter misclassified alleles. Also, overestimation of coalescent time of discordant pairs (using the mutation clock or combined clock model) was found to be less problematic for the subsequent estimation of allele age.

1.6.3 Impact of data error

The haplotype-based HMM was used to infer shared haplotype segments at target sites selected in datasets $\mathcal{D}_B$ and $\mathcal{D}^*_B$, thus evaluating the impact of breakpoint inference and subsequent age estimation before and after error. Note that both datasets were available as "true" (simulated) and phased haplotypes, for which the analysis was conducted additionally to measure the impact of phasing error. The effect of using the relaxed nearest neighbour approach to prioritise discordant pairs was compared to the random pair selection approach.

In total, 5,000 rare variant sites at $f_{[2,50]}$ (allele frequency $\leq 1\%$) were selected at random from the set of sites at which data error was not seen. This ensured that concordant and discordant pairs were correctly formed based on patterns of allele sharing in the sample. A maximum of 100 concordant and 100 discordant pairs was selected per target allele, resulting in 0.894 million pairwise analyses. For each pair at a given target site, simulation records were scanned to determine the true breakpoint interval along the sequence of segregating sites.

1.6.3.1 Shared haplotype inference

In Figure 1.14 (next page), the physical distance between inferred breakpoints and the corresponding focal site is shown relative to the true distance as determined from simulation records. In the following, the genetic lengths of inferred breakpoint intervals around alleles at a given frequency was used to evaluate the accuracy of the HMM; note that boundary cases were removed. Results are shown in Figure 1.15 (page 43), for the analysis before and after data error, for analyses on true and phased haplotypes, and for discordant pairs selected as the nearest neighbours or at random. The summary statistics used to measure accuracy are also given in Figure 1.15.

The accuracy to infer shared haplotype lengths in concordant pairs was high overall. The impact of phasing error were seen when intervals were relatively long; *e.g.* at $f_{\leq 10}$. At higher frequencies, differences between simulated and phased haplotypes became negligible. However, intervals showed a tendency to be overestimated towards higher frequency, which was pronounced after error.

In comparison, accuracy of intervals inferred in discordant pairs was low. While the majority of segments were shorter than those found around concordant pairs, which is expected, inferred lengths were generally underestimated before error. Notably, median genetic length was longer after error, but at a loss of accuracy; *e.g.* measured using Spearman's rank correlation coefficient (r_S) and root mean squared logarithmic error (RMSLE). Also, these metrics did not suggest a notable difference between results obtained on simulated or phased data.

When using the relaxed nearest neighbour approach, median genetic length inferred around selected discordant pairs was longer compared to randomly selected pairs, which suggested that the approach was useful to prioritise the nearest genealogical neighbours in the sample. However, r_S was reduced when compared to the random selection approach, while lower values of RMSLE indicated a smaller magnitude of error.

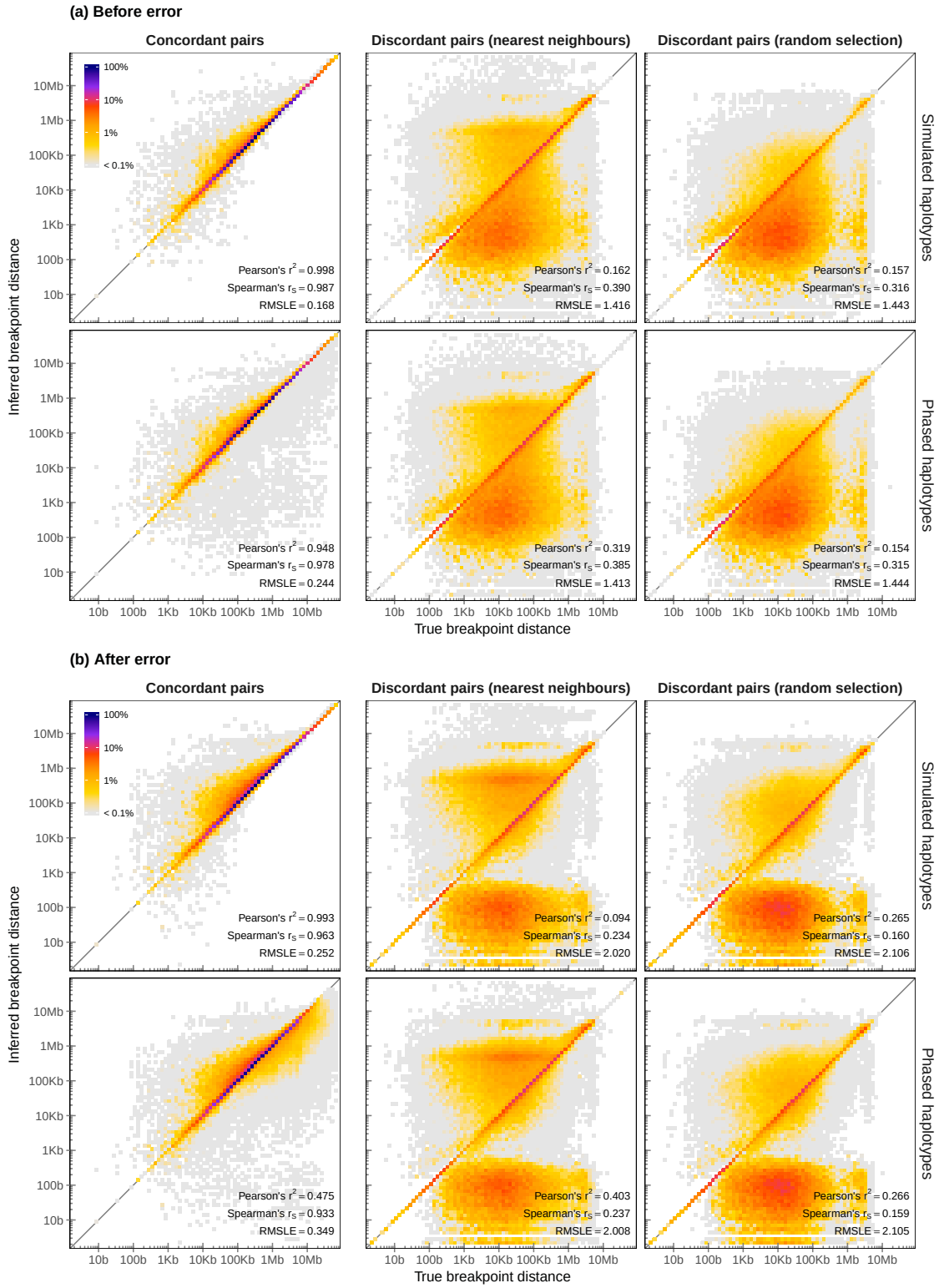


Figure 1.14: Density of breakpoint positions inferred using the haplotype-based HMM. The physical distance between true and inferred breakpoints (either left or right-hand side) is shown, where colours indicate the maximised density of breakpoints at the relative distance to the focal position, around which a shared haplotype segment was detected. Pair selection was done using the relaxed nearest neighbour approach and at random. Note that concordant pairs were selected at random throughout. Results were obtained on data before **(a)** and after error **(b)**, separately on simulated and phased haplotypes.

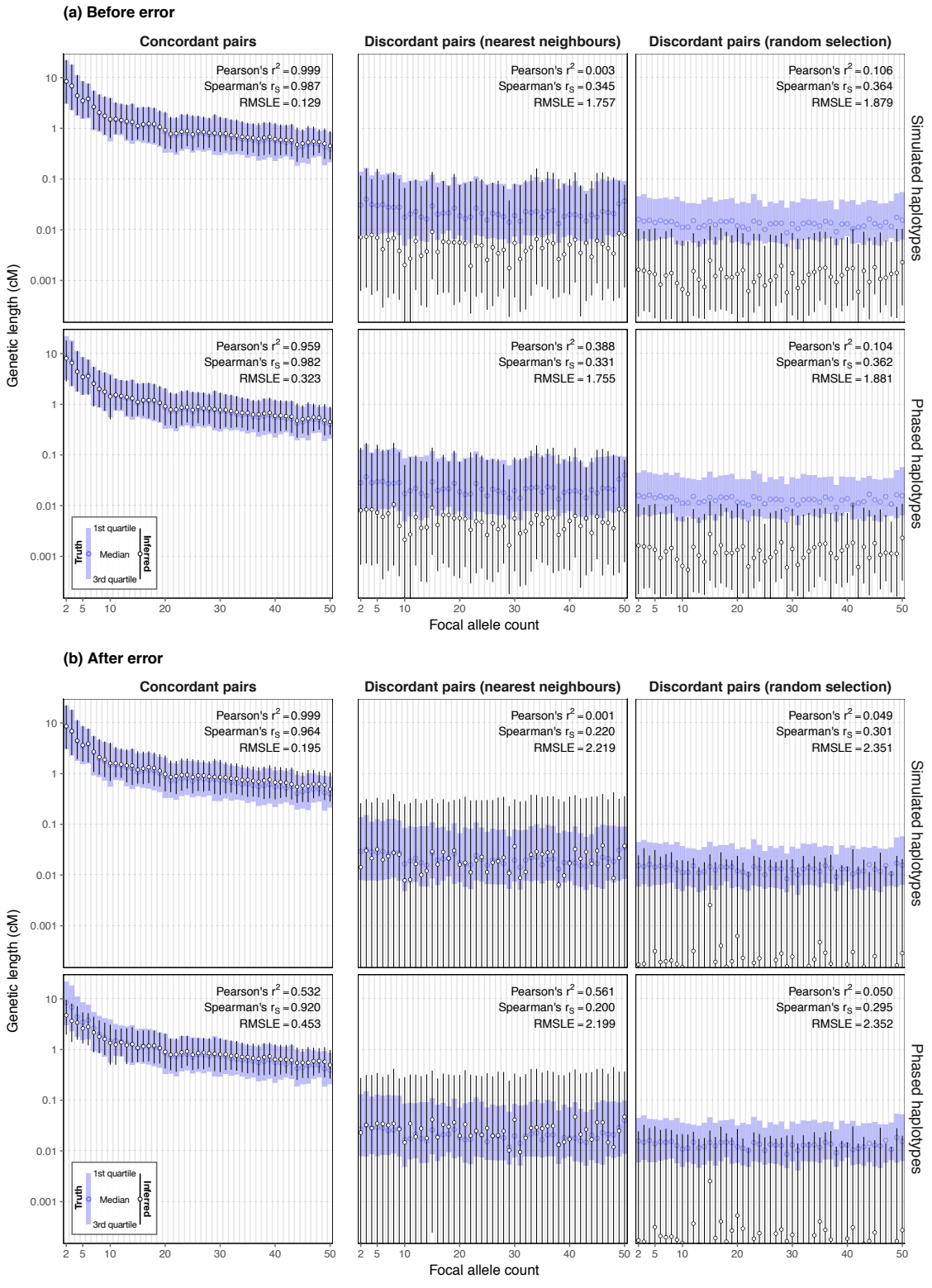


Figure 1.15: Genetic length of shared haplotype segments inferred using the haplotype-based HMM. Inferred genetic length is shown by allele count of the focal variant in the simulated sample of N = 5,000 haplotypes, in direct comparison to the corresponding true segment lengths (*blue*). Pair selection was done using the relaxed nearest neighbour approach and at random. Note that concordant pairs were selected at random throughout. Results were obtained on data before **(a)** and after error **(b)**, separately on simulated and phased haplotypes.

1.6.3.2 Allele age estimation

The shared haplotypes inferred using the haplotype-based HMM were subsequently used to estimate allele age at the selected target sites. Figure 1.16 (next page) shows the results obtained when the nearest neighbour approach was used. Age estimated using the mutation clock ($\mathcal{T}_{\mathcal{M}}$), recombination clock ($\mathcal{T}_{\mathcal{R}}$), and combined clock ($\mathcal{T}_{\mathcal{MR}}$) were compared before and after error, and for simulated and phased data. Summary statistics for both the nearest neighbour and random selection approaches are given in Table 1.4 (page 46). Note that "true" age was set at t_m; *i.e.* the geometric mean between the time of coalescent events below and above the focal mutation event (t_c and t_d), according to which accuracy was measured.

Each clock model was measured at overall high accuracy, both before and after error, but where $\mathcal{T}_{\mathcal{MR}}$ was found to outperform other models throughout. For example, the proportion of alleles "correctly" estimated (such that estimated age was within the interval between t_c and t_d) was highest for $\mathcal{T}_{\mathcal{MR}}$ when the nearest neighbour approach was used, consistently yielding > 50%. The only exception was seen when discordant pairs were selected randomly, where the proportion of correct alleles was higher for $\mathcal{T}_{\mathcal{R}}$. Accuracy was overall lower for $\mathcal{T}_{\mathcal{M}}$ in each comparison.

1.6.3.3 Discussion

The results presented in this section indicate a substantial advancement over previously evaluated methods that were employed in the age estimation method. In particular, the previous genotype-based HMM was outperformed, such that it now can be expected that the method can be applied to real data in a reliable way.

However, it would be useful to compare this method to other, established approaches. For this purpose, in the following section, I used the Pairwise Sequentially Markovian Coalescent (PSMC) to infer the T_{MRCA} at concordant and discordant pairs, from which I derived a posterior distribution that was implemented as described for the CCF in Section 1.2.2.2 (page 10) to estimate allele age.

1.6.4 Comparison to the Pairwise Sequentially Markovian Coalescent (PSMC)

The PSMC model was proposed by Li and Durbin (2011) to infer historic changes of human population size back in time, using sequence data from two haplotypes alone (*i.e.* one diploid genome). The model is based on the Sequentially Markov Coalescent

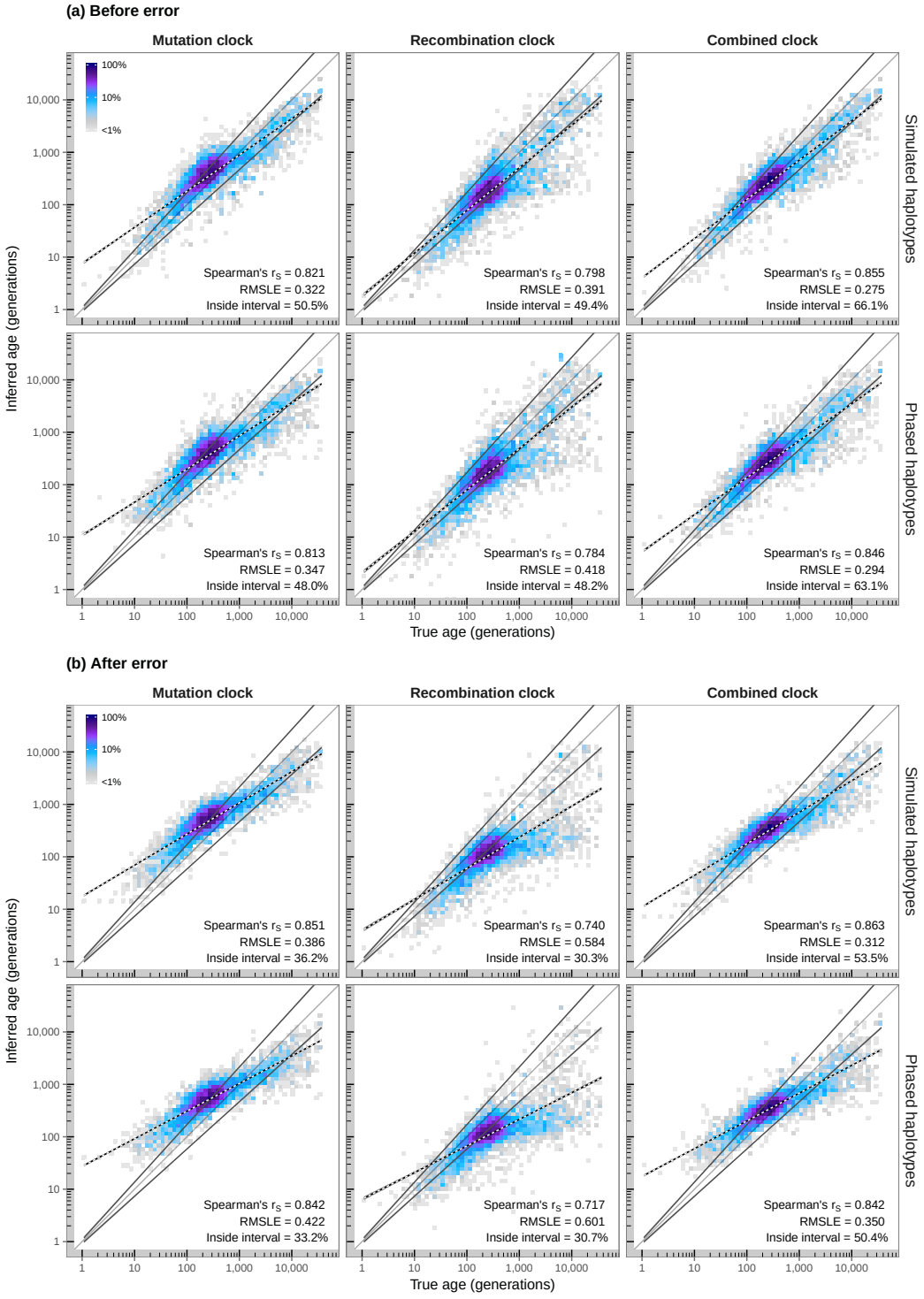


Figure 1.16: Allele age inferred using the haplotype-based HMM. The haplotype-based HMM was used to infer breakpoint intervals in data before **(a)** and after error **(b)**; analyses were conducted on "true" (simulated) haplotypes and phased haplotypes. Age was estimated using each of the three clock models, for a random set of 5,000 target sites at $f_{2,50}$, for which 100 concordant pairs were randomly selected and 100 discordant pairs were selected using the relaxed nearest neighbour approach. Note that "true" age was set at t_m, according to which the summary metrics shown were calculated. Regression lines above and below the dividing line indicate t_c and t_d. The *black-white* line shows the regression over age estimates. Colours indicate the density of true and estimated age; scaled as the maximised density.

Table 1.4: Accuracy of inferred age using the haplotype-based HMM. Age was estimated using the mutation clock ($\mathcal{T}_{\mathcal{M}}$), recombination clock ($\mathcal{T}_{\mathcal{R}}$), and combined clock ($\mathcal{T}_{\mathcal{MR}}$) based on inference of breakpoint intervals using the haplotype-based HMM. Analyses were conducted on dataset $\mathcal{D}_B$ and $\mathcal{D}_B^*$ (*i.e.* before and after error), as well as simulated and phased haplotype data in both. Note that these metrics were calculated with respect to t_m; *i.e.* the geometric mean between t_c and t_d.

Pair selection	Clock model	Before error		After error	
		SIMULATED HAPLOTYPES	PHASED HAPLOTYPES	SIMULATED HAPLOTYPES	PHASED HAPLOTYPES
Spearman's rank correlation coefficient (r_S)					
Nearest neighbour	$\mathcal{T}_{\mathcal{M}}$	0.821	0.813	0.851	0.842
	$\mathcal{T}_{\mathcal{R}}$	0.798	0.784	0.740	0.717
	$\mathcal{T}_{\mathcal{MR}}$	0.855	0.846	0.863	0.842
Randomly selected	$\mathcal{T}_{\mathcal{M}}$	0.789	0.782	0.827	0.826
	$\mathcal{T}_{\mathcal{R}}$	0.822	0.815	0.781	0.781
	$\mathcal{T}_{\mathcal{MR}}$	0.837	0.826	0.863	0.849
Root mean squared logarithmic error (RMSLE)					
Nearest neighbour	$\mathcal{T}_{\mathcal{M}}$	0.322	0.347	0.386	0.422
	$\mathcal{T}_{\mathcal{R}}$	0.391	0.418	0.584	0.601
	$\mathcal{T}_{\mathcal{MR}}$	0.275	0.294	0.312	0.350
Randomly selected	$\mathcal{T}_{\mathcal{M}}$	0.389	0.409	0.427	0.464
	$\mathcal{T}_{\mathcal{R}}$	0.323	0.337	0.342	0.347
	$\mathcal{T}_{\mathcal{MR}}$	0.311	0.329	0.331	0.371
Proportion inside interval (%)					
Nearest neighbour	$\mathcal{T}_{\mathcal{M}}$	50.5	48.0	36.2	33.2
	$\mathcal{T}_{\mathcal{R}}$	49.4	48.2	30.3	30.7
	$\mathcal{T}_{\mathcal{MR}}$	66.1	63.1	53.5	50.4
Randomly selected	$\mathcal{T}_{\mathcal{M}}$	41.5	38.9	34.7	31.2
	$\mathcal{T}_{\mathcal{R}}$	51.1	50.2	55.4	54.6
	$\mathcal{T}_{\mathcal{MR}}$	50.8	48.2	44.1	40.8

(SMC) introduced and further developed by McVean and Cardin (2005) and Marjoram and Wall (2006) for an analytically tractable approximation to the ancestral recombination graph (ARG) in model-based inferences. Li and Durbin (2011) used the PSMC model in HMM methods for inference of the T_{MRCA} between two haplotypes at sites observed along the pairwise sequence, where the observation states are defined as `'0'` (homozygous), `'1'` (heterozygous), and `'·'` (missed) genotypes (*i.e.* allelic pairs). In particular, coalescent time is divided into discrete intervals, which are the hidden states of the HMM, and a posterior probability is obtained for each state using the forward-backward algorithm (*e.g.*, see Rabiner, 1989).

I used the PSMC-HMM as a method to infer the T_{MRCA} of concordant and discordant pairs in the estimation of allele age. For a given pair, I extracted the computed coalescent time posteriors at a focal position, which I then used in the same way as the posteriors

obtained through a "clock" model in the cumulative coalescent function (CCF), so as to compute the composite posterior distribution at discrete time intervals for subsequent age estimation.

I describe the PSMC-based procedure in the section below. This is followed by an analysis using the PSMC method for comparisons to the clock models implemented in `rvage`, where accuracy was compared in terms of the inferred T_{MRCA} and allele age.

1.6.4.1 Implementation to estimate allele age

To establish a baseline comparison with the haplotype-based HMM presented here, I first performed the analysis using `rvage` in which concordant and discordant pairs were selected randomly. Identical sets of pairs per target site were then analysed using the PSMC-based method.

Note that time intervals in PSMC are not scaled on a strict logarithmic scale. The boundaries of intervals are calculated as

$$t_i \;=\; 0.1 \times \mathrm{e}^{\frac{i}{n}\log(1+10T_{\max})} - 0.1 \tag{1.28}$$

where $T_{\max}$ is the maximum T_{MRCA} considered (scaled in units of $2N_e$), n is the number of intervals (*i.e.* hidden states), and $i = 0, 1, \ldots, n$. Here, I performed the analysis with 64 coalescent time intervals, from which I computed the composite posterior at the mean between consecutive boundaries.

To conduct the analysis, I modified the `decode` algorithm implemented in software available for the Multiple Sequentially Markovian Coalescent (MSMC) method (Schiffels and Durbin, 2014), written in `D`, as it specifically applies the PSMC-HMM when two haplotype sequences are provided as input data. Modifications of `decode` were made to include the option to only return posterior probabilities at a specified target position (without affecting the computation of posteriors).*

The above modification facilitated faster computations such that I was able to obtain posterior probability distributions for a reasonably large set of target sites for a relatively large number of haplotype pairs. Yet, it must be noted that PSMC (as implemented in `decode` and embedded in the analytical pipeline used here) is slow in comparison to the haplotype-based HMM in `rvage`, which is why such analyses on a larger scale would be computationally prohibitive.

* Modified `decode` algorithm: https://github.com/pkalbers/msmc2 [Date accessed: 2017-11-04]

Dataset $\mathcal{D}_A$ was used (N_e = 10,000; $\mu = 1 \times 10^{-8}$; $\rho = 1 \times 10^{-8}$; N = 1,000), in which I selected 1,000 target sites at random at $f_{\geq 2}$ and allele frequency below 50%, so as to include alleles that could be relatively old (as opposed to only selecting rare alleles that are presumed to be relatively young). At each site, a maximum of 100 concordant and 100 discordant pairs was selected, yielding 187,420 pairwise analyses in total. The same parameters used for simulating the data were specified for inference in `rvage` and PSMC.

1.6.4.2 Results for the T_{MRCA}

Simulation records were scanned to obtain the true T_{MRCA} for each haplotype pair at a given target site. The true time of coalescent events was compared to point estimates taken at the median of posterior distributions. The median was chosen because the Gamma distribution used to compute posteriors in the clock models may equal zero at the mode, such that the median was seen as a more reliable estimate. Results for concordant and discordant pairs are shown in Figure 1.17 (next page).

The combined clock, $\mathcal{T}_{\mathcal{MR}}$, showed the highest accuracy overall among the clock models when concordant pairs were considered; measured using rank correlation (r_S) and root mean squared logarithmic error (RMSLE). The mutation clock, $\mathcal{T}_{\mathcal{M}}$, was slightly more accurate for discordant pairs. The PSMC-based approach, however, was highest in terms of r_S in both concordant and discordant pairs, but where RMSLE indicated a higher magnitude of error compared to $\mathcal{T}_{\mathcal{MR}}$ and $\mathcal{T}_{\mathcal{M}}$. Notably, as seen in Figure 1.17, because of the discrete time intervals in PSMC it was problematic to infer recent T_{MRCA}. While it would be possible to increase the number of hidden states to include intervals at more recent points in time, a consequence would be that computations would become substantially slower.

An additional analysis was conducted using the results obtained above by sorting pairs by their true T_{MRCA} into broader time intervals at which accuracy was measured (r_S and RMSLE). The results are given in Table 1.5 (page 50). The PSMC-HMM was notably more accurate compared to each clock model when discordant pairs where considered. But for concordant pairs, $\mathcal{T}_{\mathcal{MR}}$ outperformed PSMC when the true coalescent time was more recent that 10,000 generations.

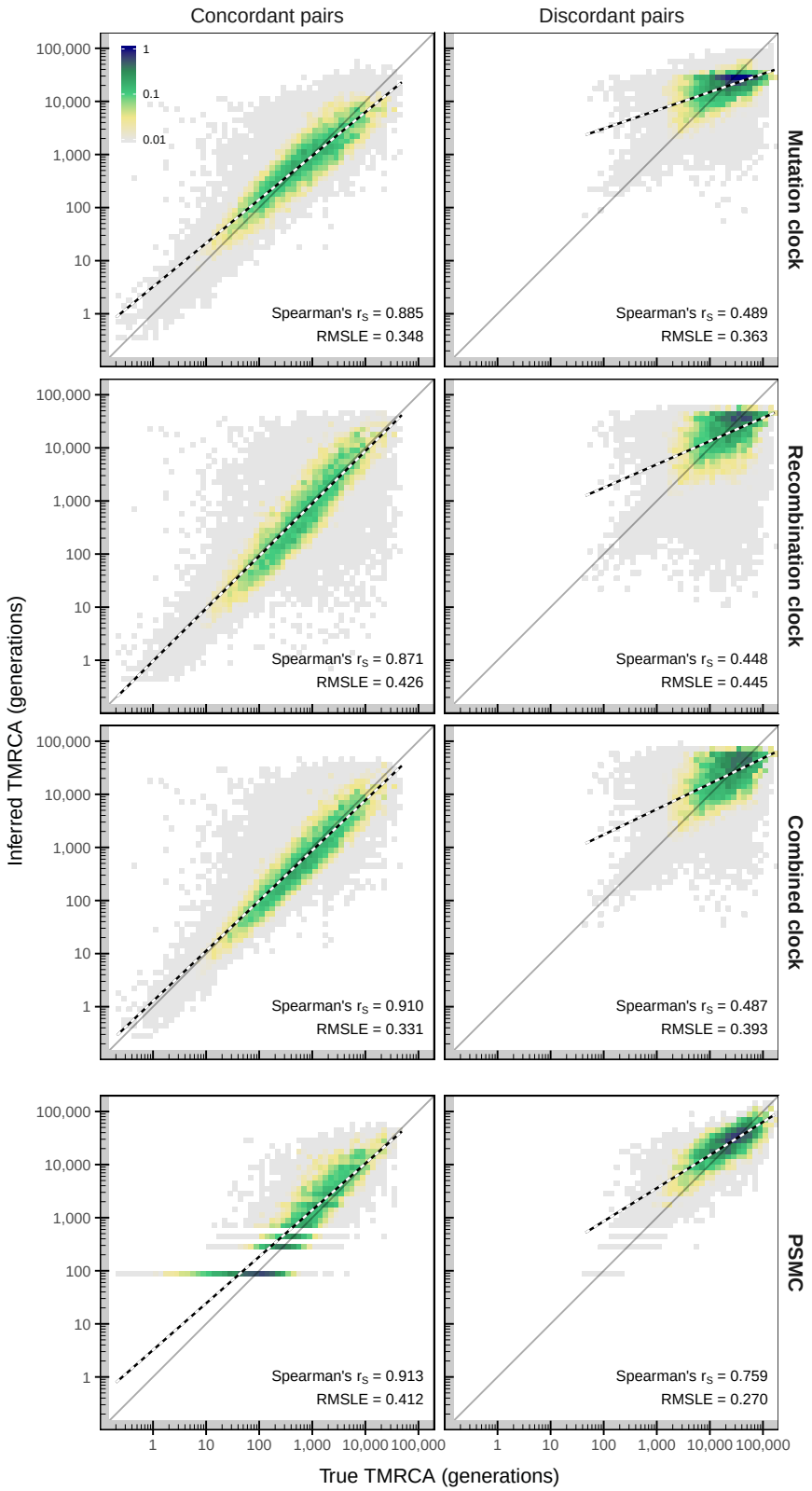


Figure 1.17: True and estimated T_{MRCA} using PSMC. The clock models developed in this chapter were compared to the PSMC-based method for inference of the T_{MRCA}. The posterior distribution on the coalescent time was obtained for the same sets of concordant and discordant pairs at 1,000 randomly selected target positions in simulated data ($\mathcal{D}_A$). Point estimated were taken at the mode of posterior distributions. Colours indicate the density of true and estimated coalescent time; scaled as the maximised density.

Table 1.5: Accuracy of T_{MRCA} estimation for different methods. The T_{MRCA} estimation conducted using PSMC is compared to estimates obtained using the mutation clock ($\mathcal{T}_{\mathcal{M}}$), recombination clock ($\mathcal{T}_{\mathcal{R}}$), and combined clock ($\mathcal{T}_{\mathcal{MR}}$), where estimates were obtained on identical target sites and haplotype pairs; the median was taken as a point estimate from each posterior. Accuracy was measured using Spearman's rank correlation coefficient (r_S) and root mean squared $\log_{10}$ error (RMSLE) at discrete time intervals defined on the true T_{MRCA} (t) of a given pair at a target site, as determined from simulation records. The number of estimates compared per method at a given time interval is indicated (n).

True T_{MRCA} (generations)	n	Rank correlation (r_S)				RMSLE			
		$\mathcal{T}_{\mathcal{M}}$	$\mathcal{T}_{\mathcal{R}}$	$\mathcal{T}_{\mathcal{MR}}$	PSMC	$\mathcal{T}_{\mathcal{M}}$	$\mathcal{T}_{\mathcal{R}}$	$\mathcal{T}_{\mathcal{MR}}$	PSMC
Concordant pairs									
$t \leq 100$	13,854	0.724	0.664	**0.740**	0.227	0.393	0.435	**0.363**	0.612
$100 < t \leq 1{,}000$	37,505	0.655	0.633	**0.714**	0.713	0.328	0.408	**0.320**	0.390
$1{,}000 < t \leq 10{,}000$	32,563	0.547	0.581	0.645	**0.656**	0.330	0.426	**0.323**	0.341
$10{,}000 < t \leq 100{,}000$	3,698	0.277	0.269	0.327	**0.525**	0.491	0.549	0.389	**0.220**
$t > 100{,}000$	0	–	–	–	–	–	–	–	–
Discordant pairs									
$t \leq 100$	16	0.159	0.245	0.214	**0.473**	1.415	1.401	1.400	**0.225**
$100 < t \leq 1{,}000$	944	0.204	0.177	0.197	**0.518**	1.017	1.021	1.029	**0.577**
$1{,}000 < t \leq 10{,}000$	21,469	0.314	0.280	0.308	**0.547**	0.488	0.523	0.529	**0.400**
$10{,}000 < t \leq 100{,}000$	75,713	0.298	0.272	0.301	**0.605**	0.291	0.402	0.326	**0.211**
$t > 100{,}000$	1,658	0.369	0.320	**0.382**	0.329	0.624	0.625	0.464	**0.337**

1.6.4.3 Results for allele age

Next, allele age was estimated by calculating the composite posterior distribution from the T_{MRCA} posteriors obtained in pairwise analyses per approach. The mode of the resulting composite posterior was taken as a point estimate for allele age. Results are shown in Figure 1.18 (next page), in which the relevant summary statistics to quantify accuracy are indicated; that is, r_S, RMSLE, and the proportion of "correct" alleles (estimated to sit between t_c and t_d). As before, "true" age was set at t_m (geometric mean between t_c and t_d).

Each method achieved high levels of accuracy overall; for example, $r_S > 0.9$ in each method. The proportion of correctly estimated alleles was > 65% in $\mathcal{T}_{\mathcal{R}}$ and $\mathcal{T}_{\mathcal{MR}}$. However, rank correlation measured for the PSMC-based approach was lowest in this comparison. Likewise, RMSLE indicated a higher magnitude of error for PSMC, and the proportion of correct alleles was also smaller.

Additionally, these results were again sorted into broader time intervals. Three intervals were distinguished at a nominal value of 1,000 generations, so as to distinguish relatively "young" alleles from "old" ones, considering that the distribution of true allele age was defined at overlapping age intervals at t_c and t_d. Exact definitions and accuracy results are given in Table 1.6 (page 52). Notably, rank correlation measured at alleles

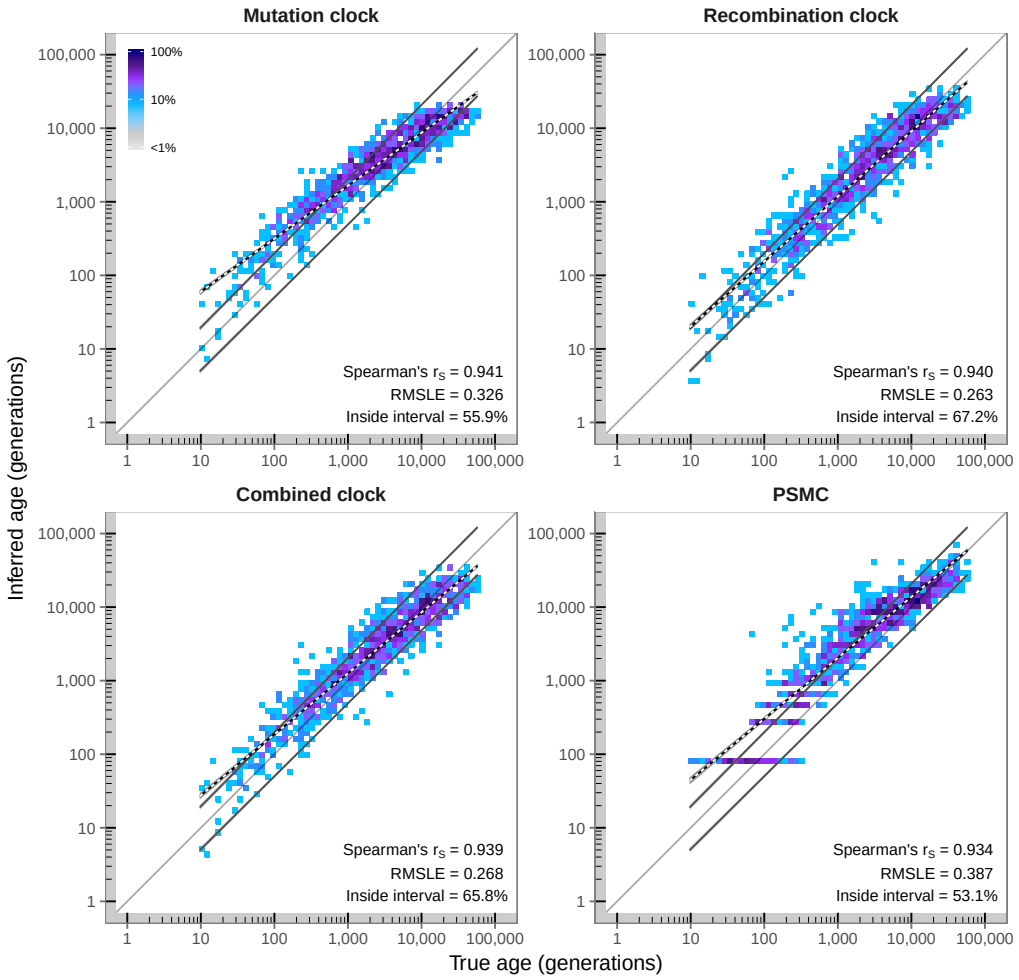


Figure 1.18: Allele age inferred using PSMC. The three clock models ($\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$) were compared to PSMC. Posterior distributions obtained on the same set of concordant and discordant pairs were used to compute the composite posterior distribution, from which point estimates were taken at the mode in each approach. The results shown compare the "true" age of an allele (set at t_m) to the estimated age at 1,000 target sites in dataset $\mathcal{D}_A$, which were randomly selected at allele frequency $\leq 50\%$. Regression lines above and below the dividing line indicate the regression over t_c and t_d. The *black-white* line shows the regression over age estimates. Colours indicate the density of true and estimated age; scaled as the maximised density.

estimated based on PSMC was highest for older alleles, but where r_S was higher for $\mathcal{T}_{\mathcal{M}}$ when younger alleles were considered. The proportion of correct alleles and the magnitude of error, however, favoured $\mathcal{T}_{\mathcal{R}}$ and $\mathcal{T}_{\mathcal{MR}}$ overall.

1.6.4.4 Discussion

Allele age estimated using $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, or $\mathcal{T}_{\mathcal{MR}}$ was overall comparable to the estimates obtained using the PSMC-based approach. However, it must be noted that the integration of PSMC as a method to subsequently arrive at the composite posterior distribution may not have

Table 1.6: Accuracy of allele age inferred using PSMC. The true times of the delimiting coalescent events t_c and t_d are used to distinguish alleles of young, intermediate, and old age; see definitions provided at the bottom of the table.

Method	Rank correlation (r_S)			RMSLE			Inside interval (%)		
	Set A	Set B	Set C	Set A	Set B	Set C	Set A	Set B	Set C
$\mathcal{T}_{\mathcal{M}}$	**0.888**	0.671	0.791	0.477	0.296	0.249	23.6	67.0	65.2
$\mathcal{T}_{\mathcal{R}}$	0.840	0.673	0.802	**0.307**	0.272	**0.238**	**53.2**	**82.6**	66.9
$\mathcal{T}_{\mathcal{MR}}$	0.854	0.676	0.801	0.333	**0.262**	0.238	45.9	81.7	**67.8**
PSMC	0.817	**0.747**	**0.828**	0.525	0.444	0.277	30.9	55.9	61.5

Set A: "Young" age, $n = 233$, defined at $t_d \leq 1{,}000$
Set B: "Intermediate" age, $n = 222$, defined at $t_c < 1{,}000$, $t_d > 1{,}000$
Set C: "Old" age, $n = 543$, defined at $t_c \geq 1{,}000$

been an ideal baseline for comparisons to the clock models developed in this chapter. In particular, the current procedure considered 64 coalescent time intervals, which may not directly compare to the continuous scale used by the estimation method.

Nonetheless, the results presented in this section suggested that the haplotype-based HMM used for targeted shared haplotype inference achieved comparable estimates of T_{MRCA}, despite incorporating a (biased) empirical emission model. I further showed that the haplotype-based HMM was able to estimate age of alleles that occurred at relatively high frequencies in the data. Also, note that the decreased accuracy in inferences at discordant pairs had only minor effects on subsequent estimation of allele age. Thus, I used the haplotype-based HMM in the following section to estimate allele age in real data.

1.6.5 Allele age estimation in 1000 Genomes

The haplotype-based HMM was used for allele age estimation in an extensive analysis of data from the 1000 Genomes Project (1000G) Phase III. I selected 50,000 sites at random in chromosomes 1-22, but only at positions within high confidence regions as defined in the strict accessibility mask available for the Phase III dataset. The diploid sample size was $N = 2{,}504$. Note that all sites at $f_{\geq 2}$ were considered. Model parameters in `rvage` were $N_e = 10{,}000$, $\mu = 1.2 \times 10^{-8}$, and recombination rates according to genetic maps available from HapMap Phase II, Build 37.*

In total, 2.370 million concordant and 5.204 million discordant pairs were analysed. Below, I provide a descriptive analysis of the haplotype structure at alleles shared within and between populations, followed by an overview of allele age as estimated per population group; namely African (AFR), Ad-Mixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). I conclude this section with selected examples where allele age was estimated at specific loci.

* HapMap recombination map: ftp://ftp.ncbi.nlm.nih.gov//hapmap/recombination/2011-01_phaseII_B37/genetic_map_HapMapII_GRCh37.tar.gz [Date accessed: 2016-11-12]

1.6.5.1 Shared haplotypes inferred by population

The set of pairs was divided into pairs at which the focal allele was shared among the individuals within the same population and those at which it was shared between different groups. This implied that only concordant pairs were considered, of which 2.357 million were retained after removing boundary cases. Median physical and genetic lengths are shown in Figure 1.19 (next page); the proportion of pairs at which alleles were shared within and between populations is indicated.

The average ratio of genetic and physical length was $1.680\,\mathrm{cM\,Mb^{-1}}$ (±0.002 SE). Also, the ratio was similar for discordant pairs; $1.625\,\mathrm{cM\,Mb^{-1}}$ (±0.002 SE). Segment lengths were seen to decrease towards higher focal allele frequencies when alleles were shared within the same population, suggesting an almost linear trend on log-log scale in each population. However, note that the number of target sites found at higher frequencies was also low. Segments inferred around rare alleles (*e.g.* $\mathrm{f}_{<10}$) were relatively long in AFR and AMR, indicating that carrier haplotypes were separated by very recent coalescent events. In contrast, segments where focal alleles were shared between different populations showed notably shorter lengths at lower frequencies.

1.6.5.2 Allele age estimated by population

Age was estimated using the relaxed nearest neighbour approach to select discordant pairs, and using the same model parameters per population group. Note that N_e is expected to differ across populations, but which may only lead to inappropriate scaling of the population-scaled age estimates per group. Here, I focused on alleles that were shared only within populations, which retained 31,906 target sites. The results for each clock model are shown in Figure 1.20 (page 55).

A general difference was seen between clock models, where age was highest on average in $\mathcal{T}_{\mathcal{M}}$, and lowest $\mathcal{T}_{\mathcal{R}}$. But as suggested in previous analyses, $\mathcal{T}_{\mathcal{MR}}$ would be the preferred choice among clock models. Further, it is not straightforward to compare estimated age distributions between populations, as the number of alleles retained differed substantially among populations and where appropriate scaling of time would be needed to facilitate such comparisons.

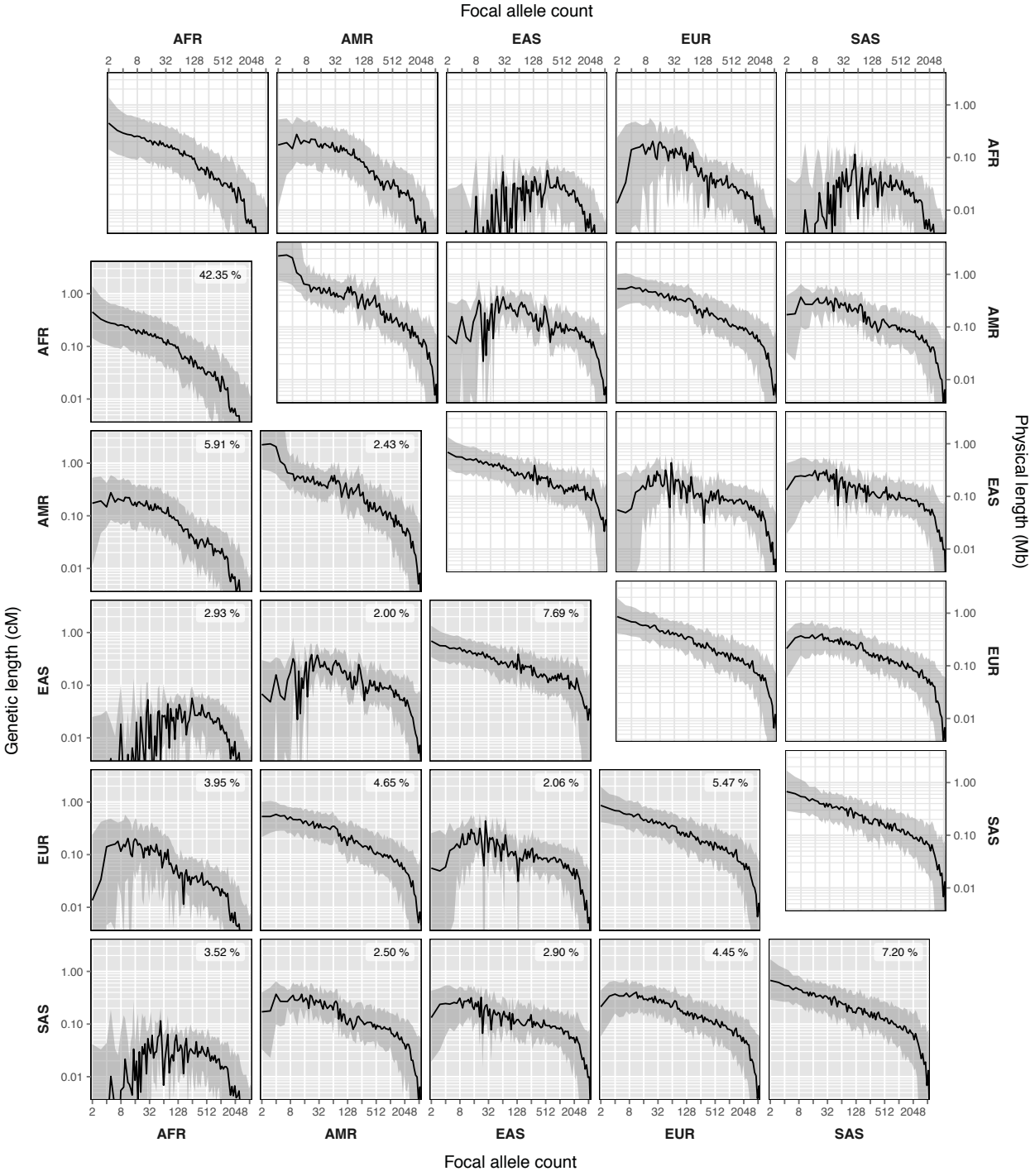


Figure 1.19: Shared haplotype length by population in 1000 Genomes, chromosome 20. Median physical and genetic lengths are shown for haplotype pair where individuals share a focal allele within the same or between different populations. Note that the condition of seeing an allele shared between individuals implied that only concordant pairs were considered. Physical lengths are given in the upper triangle (*white* panels) and genetic lengths in the lower triangle (*grey* panels). The median (*black* lines) is drawn between the 1st and 3rd quartiles, where lengths were binned by focal allele count (log-scale). The lower triangle also shown the proportions of pairs at which the focal allele was shared within or between populations.

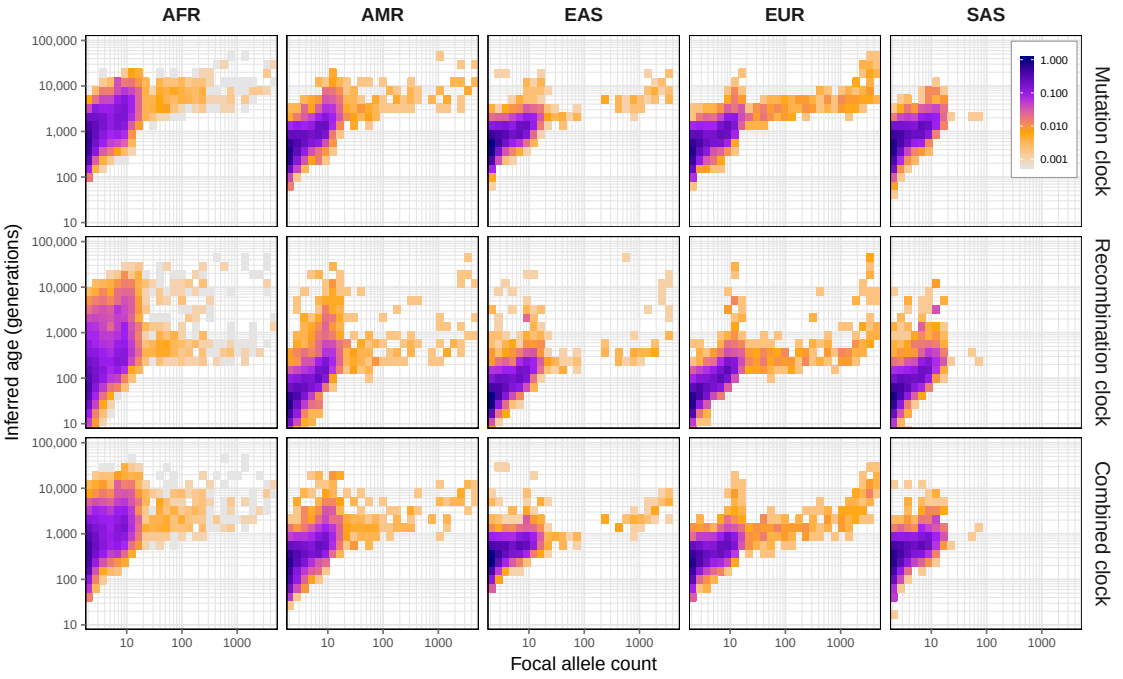


Figure 1.20: Allele age estimated by population in 1000 Genomes. Age was estimated for alleles shared within the same population group contained in the 1000G sample, using the relaxed nearest neighbour approach to select discordant pairs. Shared haplotype detection was performed using the haplotype-based HMM and age was estimated under each of the three clock models ($\mathcal{T}_M$, $\mathcal{T}_R$, and $\mathcal{T}_{MR}$). Colours indicate the maximised density of allele age by focal allele frequency. Note that results are shown on log-log-scale.

1.6.5.3 Selected allele age profiles

The results presented above produced a holistic picture of the broader distribution of mutational origin at population level. However, given the amount of work that went into the characterisation of the underlying genealogy at a single locus, such a representation may not capture the relevant aspects that could be explored by focusing on one allele alone.

Figure 1.21 (next page) shows the *profiles* of the inferred shared haplotype distribution and allele age at two selected target sites in 1000G. Pairs were sorted by T_{MRCA} to indicate the shape of the underlying genealogy around a given focal site. Estimated allele age is shown in between the sorted concordant and discordant pairs.

The example shown in Figure 1.21a is an intron variant the the MCM6 locus on chromosome 2, which sits approximately 22 kb upstream of the LCT gene (encoding the lactase enzyme) and has been associated with lactase persistence in Europeans. For example, Enattah *et al.* (2002) found that the variant occurs in distantly related populations and therefore concluded that it is relatively old. Research on the LCT gene

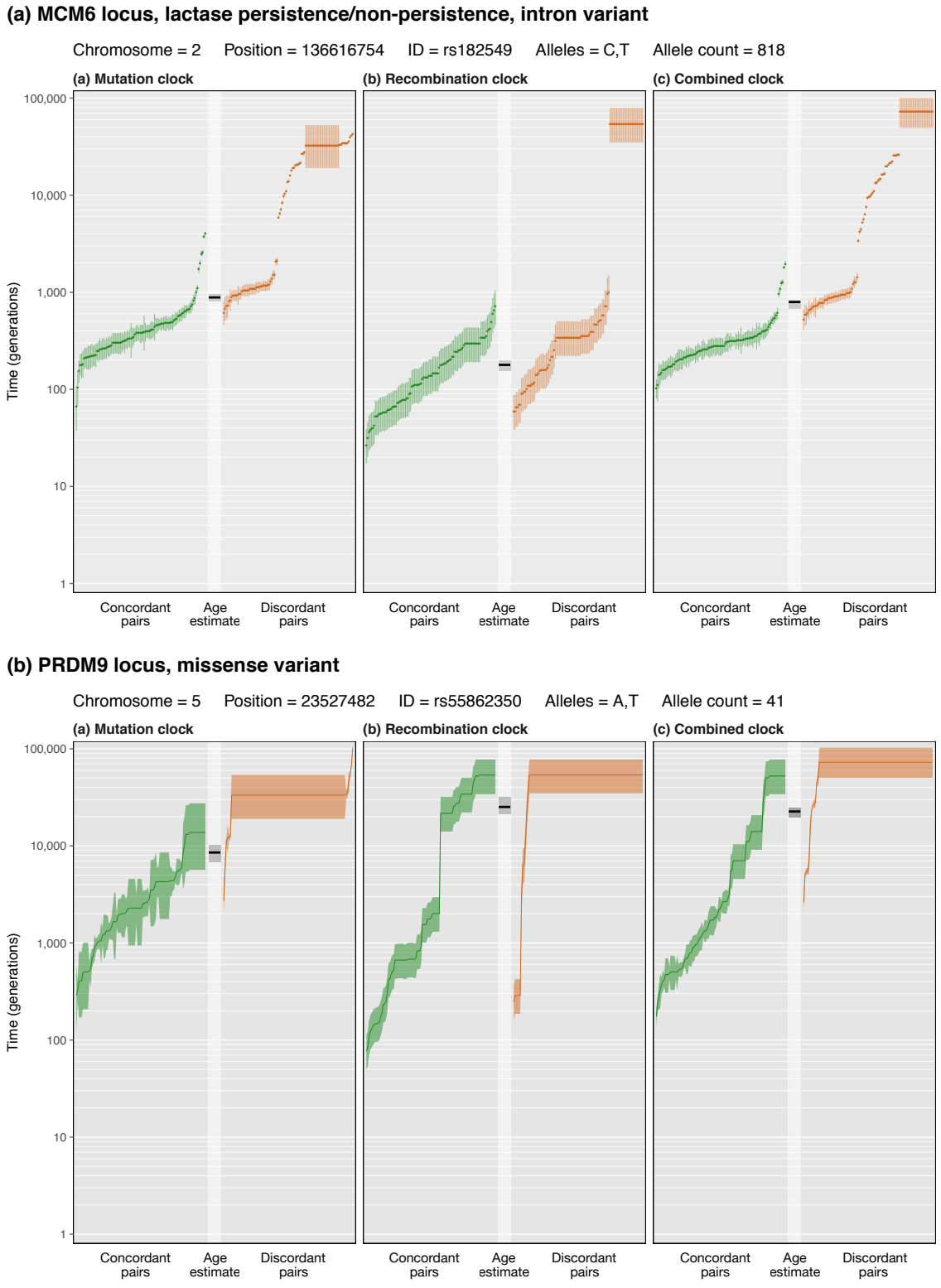


Figure 1.21: Example profiles of estimated allele age in 1000 Genomes, chromosome 20. Allele age profiles are shown for two selected loci in 1000G data. Each profile is composed of the T_{MRCA} posterior distributions inferred for concordant pairs (*left*) and discordant pairs (*right*), which are sorted by their inferred T_{MRCA}. The estimated age obtained from the resulting composite posterior distribution is shown in the *middle*. Each T_{MRCA} distribution is shown as the median, around which the 1st and 3rd quartiles are drawn. The mode of the composite posterior distribution is shown within a 95% pseudo-confidence interval. Time was scaled at $2N_e$, where $N_e = 10{,}000$.

has indicated signatures of recent positive selection around years ago 5,000 – 10,000 (Bersaglieri *et al.*, 2004). Here, age estimation suggested that the variant originated at around 800 generations ago (according to $\mathcal{T}_{\mathcal{MR}}$).

A missense variant at the PRDM9 locus in chromosome 5 is shown in Figure 1.21b. Interestingly, the variant is low in frequency in the sample, but its age was estimated to be surprisingly old, and far older in comparison to Figure 1.21a. Note that I also analysed other SNPs at PRDM9 which also indicated an old age and a similar haplotype structure. However, here, I only attempted to demonstrate one possible application of the methodology, but further research would be needed to arrive at conclusive results for either of the examples provided.

1.6.6 Discussion

The results in this section provided a general overview of the haplotype structure and allele age distributions that can be expected when the methodology is applied to larger, diverse sample data. In must be noted that the methods were previously tested on target sites at which no data error was present, but it should be expected that the inclusion of false positive or false negative sites may substantially bias the results. Although, here, I included sites that were called with high confidence in 1000G, it was nonetheless possible that allele sharing was flawed at a considerable fraction of sites.

An overview such as presented here can only be limited, because the most interesting observations would arise from looking at particular variant sites; *e.g.* due to effects such as selection, migration, or population stratification. I presented selected allele age profiles as an example. In this variant-centric view, ideally, the history of an allele can be observed back in time and followed through different lineages to arrive at the time and place of its origin. Potential applications include, for example, analyses of alleles previously implicated in disease status, or the identification of previously unnoticed alleles at which indicators of selection are suggested.

The key test for an acronym is to ask whether it helps or hurts communication.

— Elon Musk

Abbreviations

1000G	1000 Genomes Project
ARG	Ancestral recombination graph
CCF	Cumulative coalescent function
CDF	Cumulative distribution function
cM	CentiMorgan
DGT	Discordant genotype test
FGT	Four-gamete test
HapMap	International HapMap Project
HMM	Hidden Markov Model
Mb	Megabase
MRCA	Most recent common ancestor
MSMC	Multiple Sequentially Markovian Coalescent
PDF	Probability density function
PMF	Probability mass function
PSMC	Pairwise Sequentially Markovian Coalescent
RMSLE	Root mean squared logarithmic error
SMC	Sequentially Markov Coalescent
SNP	Single-nucleotide polymorphism
$\mathbf{T_{MRCA}}$	Time to the most recent common ancestor

My definition of a scientist is that you can complete the following sentence: 'he or she has shown that . . . '

— E. O. Wilson

Bibliography

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., and Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, **74**(6), 1111–1120.

Delaneau, O., Coulonges, C., and Zagury, J.-F. (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. **9**(1), 540.

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, **10**(1), 5–6.

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nature genetics*, **30**(2), 233–237.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, **5**(10), e1000695–11.

Hein, J., Schierup, M., and Wiuf, C. (2004). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, Oxford.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, **159**(4), 1805–1817.

International HapMap 3 Consortium, Schaffner, S. F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Keinan, A., Montgomery, S. B., Pollack, S., Soranzo, N., Gonzaga-Jauregui, C., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Zhang, Q., Ghori, M. J. R., McGinnis, R., McLaren, W., Price, A. L., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., and McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**(7311), 52–58.

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Onofrio, R. C., Richter,

D. J., Birren, B. W., Daly, M. J., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–861.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS computational biology*, **12**(5), e1004842–22.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**(4), 893–903.

Kimura, M. and Ota, T. (1973). The age of a neutral mutant persisting in a finite population. *Genetics*, **75**(1), 199–212.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357), 493–U84.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics*, **80**(1), 221–239.

Marjoram, P. and Wall, J. D. (2006). Fast" coalescent" simulation. *BMC genetics*, **7**(1), 16.

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

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**(5670), 581–584.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*.

Papoulis, A. and Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education.

Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, pages 149–164.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286. Alcatel-Lucent Bell Labs, Murray, United States.

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

Schroff, M. H. (2016). *Genealogical properties of rare variation and their implications for demographic inference*. Ph.D. thesis, University of Oxford.

Slatkin, M. and Rannala, B. (2000). Estimating allele age. *Annual review of genomics and human genetics*, **1**(1), 225–249.

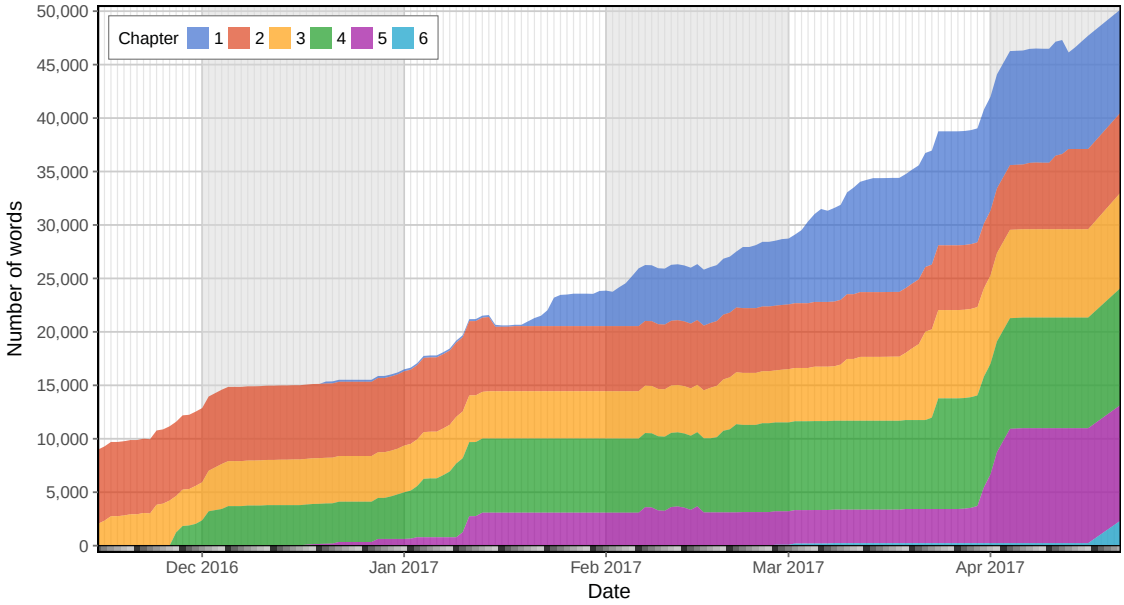
Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**(2), 437–460.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, **145**(2), 505–518.

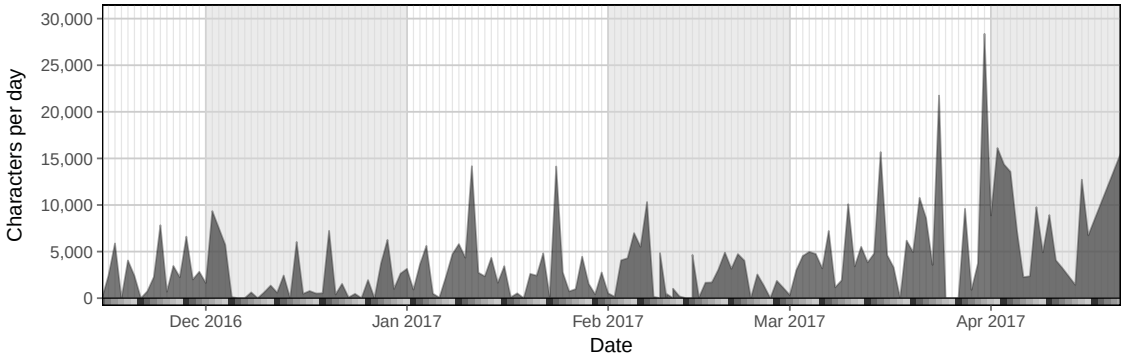
Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, **7**(2), 256–276.

Remember kids, the only difference between
screwing around and science
is writing it down.

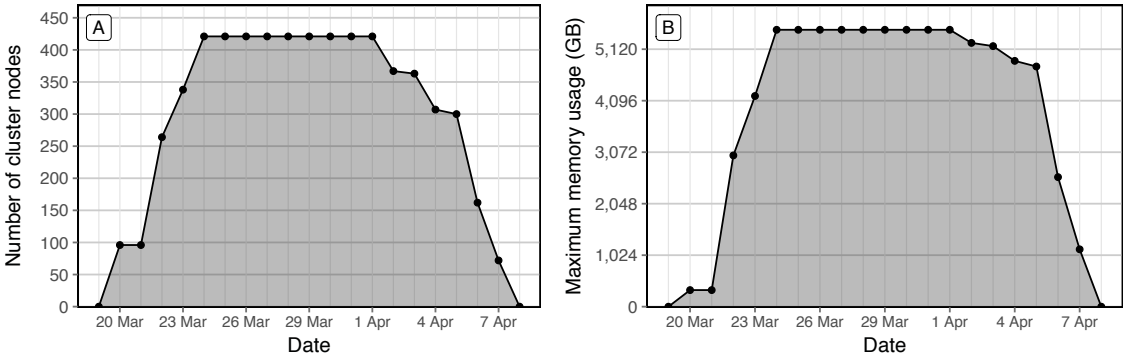
— Adam Savage



Supplementary Figure 1: Word count over time during thesis writing period. Shown for the time since I automatically generated daily backups and until the submission of this thesis.



Supplementary Figure 2: Number of characters written per day. Note that all characters in each LaTeX file were counted.



Supplementary Figure 3: Computer cluster usage one month before the submission date of this thesis. Indicated by the **(A)** number of nodes used and **(B)** daily maximum of computer memory on the cluster of the Wellcome Trust Centre for Human Genetics.