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 the 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. The approach is based on *composite likelihood* methods, which recently have gained in popularity for various applications in genetic research, in particular, where the full likelihood cannot be known analytically or is computationally intractable. Coalescent-based composite likelihood methods were 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).

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 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 haplotypes carry the focal allele, it is assumed that they co-inherited the allele 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 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 inferred in a Bayesian framework in which the posterior probability of the time to the most recent common ancestor (T_{MRCA}) between a pair of haplotypes is calculated As is central to Bayesian inference, a likelihood function is calculated from a likelihood function that "updates" updating prior beliefs about the parameters

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

The procedure employed here

"Bayesian empirical likelihood"

The proposed method relies on the estimation of the T_{MRCA} between pairs of haplotypes.

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. Below (Section 1.2.1), I derive the formulations for three 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 $\mathcal{T}_{\mathcal{MR}}$

Each clock model defines a posterior probability distribution for the T_{MRCA} over a prior distribution of the coalescent time. Given larger sample data, this is done for several hundreds or thousands of haplotype pairs, where pairs are formed between carrier haplotypes as well as carrier and non-carrier haplotypes with respect to a single allele, for which it is attempted to estimate the age. To put this approach into practise, posteriors are combined in a fashion similar to existing composite likelihood methods. The approach may therefore be described as a *composite posterior* method, but is more correctly defined as an *ad-hoc* analysis. The procedure of the age estimation method is explained in detail in Section 1.2.2 (page 8).

1.2.1 Coalescent time estimators

Any pair of chromosomes can be seen as a mosaic of haplotype segments that derived from different ancestors who lived at different points in time.

At a given position in the genome, it is assumed that the *breakpoints* of the recombination events that delimit the shared haplotype region around that position are known, such that no recombination has occurred along the sequence in the two haplotypes considered. it is assumed that the length of the shared haplotype region around over which they share a haplotype by descent

Relative to the genealogy seen at a given focal site, a *breakpoint* is defined as the location at which the genealogical relationship between two haplotypes changes due to recombination.

The haplotype region that is shared by descent between a pair of chromosomes is therefore delimited by two breakpoints on the left and right-hand side of the focal position.

A *boundary case* is If no recombination occurred on either the left or the right-hand side such that is recorded where the chromosomal end position is taken as a breakpoint to delimit the length of the interval.

In the following, it is assumed that no recombination has occurred along the sequence in the haplotype segment considered, such that the genealogical relationship between the two haplotypes does not change along the region. This facilitates the analysis under a coalescent process.

The presented age estimation method is based on the computation of the posterior probability of the T_{MRCA} in pairs of haplotypes. The posterior probability is proportional to the prior probability of the time to coalescence multiplied by the likelihood of the time conditional on the observed parameter values defined for a given estimator. 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 are derived from one common ancestral haplotype t generations in the past is

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

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. Further, the probability that the two haplotypes do not share an ancestral haplotype more recently than t generations in the past is given by

$$P(T_c > t \mid N_e) \approx e^{-\frac{t}{2N_e}} \quad \text{CORRECTED}$$

where T_c is the time of the coalescent event between two lineages. It is convenient to use a continuous time approximation and measure time in units of $2N_e$ generations, in the context of the coalescent, such that $\tau = t/2N_e$ Thus, the prior distribution of the coalescent time is $\tau \sim \text{Exp}(1)$ and written as

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

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

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

$$f_S(s) \;=\; P(S = s \mid \theta, h, \tau) \;=\; \frac{(\theta h \tau)^s}{s!}\, \mathrm{e}^{-\theta h \tau}. \tag{1.1}$$

Note that the equation above is the *joint* probability of observing s as the sum of mutational differences along the length h.

The likelihood function for the time parameter τ is proportional to Equation (1.1), but requires only those terms that involve τ and where constant terms can be dropped, such that

$$\mathcal{L}(\tau \mid \theta, h, s) \;\propto\; \tau^s\, \mathrm{e}^{-\theta h \tau}\,. \tag{1.2}$$

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

$$\begin{aligned} p(\tau \mid \theta, h, s) \;&\propto\; \mathcal{L}(\tau \mid \theta, h, s) \;\times\; \pi(\tau) \\ &\propto\; \tau^s\, \mathrm{e}^{-\tau(\theta h + 1)} \end{aligned} \tag{1.3}$$

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

In the above, the density of the posterior probability is specified up to a missing normalising constant. Note that Equation (1.3) 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

$$p(\tau \mid \theta, h, s) \;=\; g(\tau \mid s + 1, \theta h + 1). \tag{1.4}$$

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 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.* Because recombination may occur independently on either of the two haplotypes, distance is modelled such that $D \sim \mathrm{Exp}(2\psi\tau)$, 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

$$P(D = d \mid \psi, \tau) \;=\; 2\psi\tau\; \mathrm{e}^{-2\psi\tau d}. \tag{1.5}$$

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

$$P(D > d \mid \psi, \tau) \;=\; \mathrm{e}^{-2\psi\tau d}. \tag{1.6}$$

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

Equations (1.5) and (1.6) above can be simplified to

$$f_D(d) \;=\; (2\psi\tau)^b \; \mathrm{e}^{-2\psi\tau d} \tag{1.7}$$

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 (1.7), the likelihood function for τ can now be written as

$$\mathcal{L}(\tau \mid \psi, d, b) \;\propto\; \tau^b \; \mathrm{e}^{-2\psi d\tau} \tag{1.8}$$

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

$$\mathcal{L}(\tau \mid \psi, h, b_L, b_R) \;\propto\; \tau^{b_L+b_R} \; \mathrm{e}^{-2\psi h\tau} \tag{1.9}$$

where b_L, b_R indicate the breakpoint on the left and right-hand side, respectively. The posterior probability is obtained as

$$\begin{aligned} p(\tau \mid \psi, h, b_L, b_R) \;&\propto\; \mathcal{L}(\tau \mid \psi, h, b_L, b_R) \;\times\; \pi(\tau) \\ &\propto\; \tau^{b_L+b_R} \; \mathrm{e}^{-\tau(2\psi h+1)}. \end{aligned} \tag{1.10}$$

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 $\beta = 2\psi h + 1$. Thus, the posterior density can be computed as

$$p(\tau \mid \psi, h, b_L, b_R) \;=\; g(\tau \mid b_L + b_R + 1, 2\psi h + 1). \tag{1.11}$$

Importantly, the term ψh refers to the genetic length of the shared haplotype region, but where ψ is rarely constant along the 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.

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 (1.1) and (1.7) on page 5 and page 7 and is given as

$$\mathcal{L}(\tau \mid \theta, \psi, h, s, b_L, b_R) \;\propto\; \tau^{s+b_L+b_R}\, \mathrm{e}^{-\tau h(\theta+2\psi)}.$$

However, it is more convenient to replace the term $h(\theta + 2\psi)$ above with $h_p + h_g$, where $h_p = \theta h$ and $h_g = 2\psi h$, so as to consider the physical and genetic lengths separately; *e.g.* when the recombination rate is not constant and ψh is determined from the distances given in a genetic map. Therefore,

$$\mathcal{L}(\tau \mid h_p, h_g, s, b_L, b_R) \;\propto\; \tau^{s+b_L+b_R}\, \mathrm{e}^{-\tau(h_p+h_g)} \tag{1.12}$$

from which the posterior probability is obtained as

$$\begin{aligned} p(\tau \mid h_p, h_g, s, b_L, b_R) \;&\propto\; \mathcal{L}(\tau \mid h_p, h_g, s, b_L, b_R) \;\times\; \pi(\tau) \\ &\propto\; \tau^{s+b_L+b_R}\, \mathrm{e}^{-\tau(h_p+h_g+1)}. \end{aligned} \tag{1.13}$$

As was done in both the mutation and recombination clock models, the Gamma PDF is used with $\alpha = s + b_L + b_R + 1$ and $\beta = h(\theta + 2\psi) + 1 = h_p + h_g + 1$ to compute the posterior density, *i.e.*

$$p(\tau \mid h_p, h_g, s, b_L, b_R) \;=\; g(\tau \mid s + b_L + b_R + 1, h_p + h_g + 1). \tag{1.14}$$

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

1.2.2 Inference of allele age from coalescent time posteriors

Consider a sample of haplotypes and an allele shared by some of the haplotypes. 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 co-inherited the allele (*carriers*) are distinguished from the other haplotypes which 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.

It follows that all lineages in the X_c subsample coalesce before any of them can coalesce with a lineage in the X_d subsample. Any coalescent event between two lineages in X_c must have occurred *earlier* than the focal mutation event (back in time). On the other

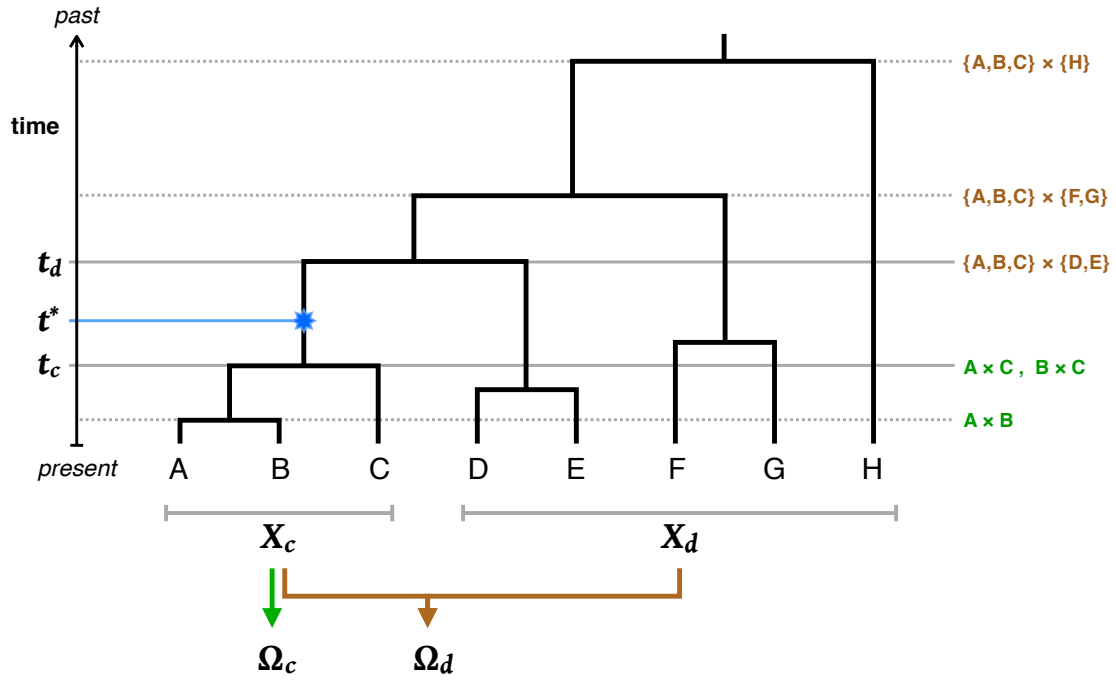


Figure 1.1: Allele age in relation to concordant and discordant pairs. The genealogy of a sample of eight haplotypes is shown of which A, B, and C share a focal allele that derived from a mutation event as indicated in the tree (*star*). These chromosomes constitute the set of *carriers*, denoted by 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.

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). In the following, 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, to distinguish the population-scaled time τ as defined for the T_{MRCA} from the time of the mutation event, let the latter be defined as the likewise population-scaled time t.

The time of a focal 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. Posteriors are obtained for concordant pairs in Ω_c where the *oldest* relation indicates the lower bound in the estimation of the focal allele age. Likewise, posteriors are obtained for discordant pairs in Ω_d where the *youngest* relation indicates the upper bound.

As alluded to earlier,

For simplicity, the approach by which the focal allele age is estimated is referred to as a *composite posterior* method.

1.2.2.1 Cumulative coalescent function (CCF)

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 later arrive at the composite posterior from which the time of the focal mutation event can be estimated, it is of interest to obtain the probability distribution of the T_{MRCA} relative to the time of the focal mutation event. 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

$$\Phi_{ij}(t) \;=\; \begin{cases} P(\tau \leq t) & \text{if } \{i,j\} \subseteq \Omega_c \quad (\textit{i.e.}\ \text{concordant pairs}) \\ P(\tau > t) = 1 - P(\tau \leq t) & \text{if } \{i,j\} \subseteq \Omega_d \quad (\textit{i.e.}\ \text{discordant pairs}). \end{cases} \tag{1.15}$$

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 density using the Gamma distribution, it is straightforward to obtain the CCF from the Gamma CDF; formally given as

$$G(t) \;=\; P(\tau \leq t \mid \alpha, \beta) \;=\; \int_0^t g(u \mid \alpha, \beta)\, du \tag{1.16}$$

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)

$$F(t) \;=\; P(\tau \leq t \mid \alpha, \beta) \;=\; 1 - \mathrm{e}^{-\beta t} \sum_{i=0}^{\alpha-1} \frac{(\beta t)^i}{i!}. \tag{1.17}$$

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

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.

$$\Lambda_k(t) \propto \prod_{i,j \in A_k} \Phi_{ij}(t \mid \alpha, \beta) \tag{1.18}$$

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 to 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 (1.18).

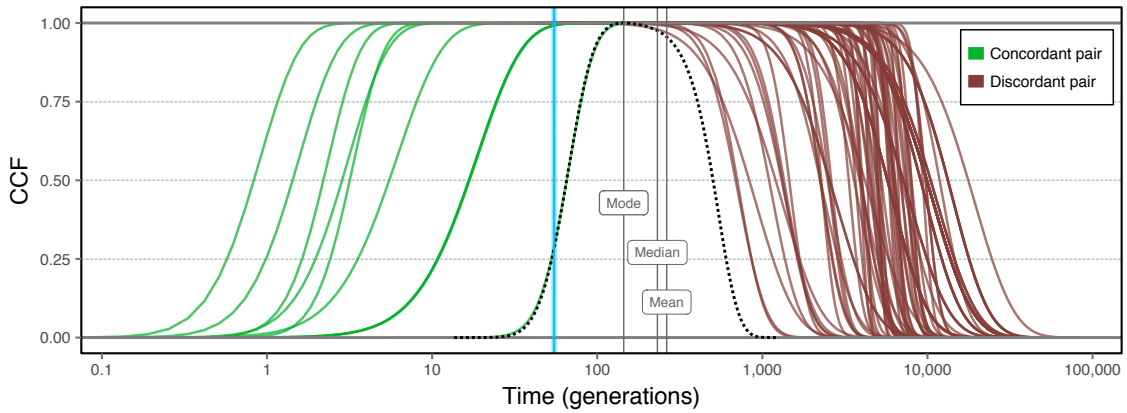


Figure 1.2: Example of the age estimation result for a focal variant. A target variant was randomly selected from simulated data. Each of the possible concordant pairs was formed and analysed using the CCF. A subset of $n_d = 100$ discordant pairs was randomly selected and analysed using the CCF. Vertical lines indicate the mode, median, and mean of the composite likelihood distribution. The *blue* line marks the true age of the mutation, as determined from simulation records.

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. An example is given in Figure 1.2 (this page), showing the CCF posterior distributions for concordant and discordant pairs, as well as the maximised composite posterior distribution. In the following, the mode of the composite posterior distribution is taken as a point estimate of allele age, denoted by $\hat{t}$.

1.2.2.3 Note on the composite likelihood

There is extensive literature on the topic of composite likelihood methods.

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)

$$\mathcal{CL}(\vartheta \mid y) \;=\; \prod_{z \in Z} \mathcal{L}_z(\vartheta \mid y)^{w_z} \tag{1.19}$$

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 composite posterior given in Equation (1.18) on page 11 has a similar form as the above, but where $\Lambda_k(t)$ is proportional to the product of posteriors and, thus, cannot be regarded as a composite likelihood.

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

$$p_{\mathcal{CL}}(\vartheta \mid y) \;\propto\; \pi(\vartheta) \;\times\; \mathcal{CL}(\vartheta \mid y) \tag{1.20}$$

where $\pi(\vartheta)$ is a suitable prior on the parameter.

1.2.2.4 Note on the computational burden

A major caveat of the method presented 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

$$\max[n_c] \;=\; \binom{k}{2} \;=\; \frac{k(k-1)}{2} \tag{1.21}$$

where k is the number of allele copies observed in the sample; *i.e.* the size of X_c. The number of possible discordant pairs is given by

$$\max[n_d] \;=\; k(2N-k) \tag{1.22}$$

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, if $N = 1{,}000$, 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.

1.2.3 Allele age estimation given complete knowledge of the shared haplotype structure

The allele age estimation method presented in this chapter relies on inference of the haplotype region shared by descent between two chromosomes relative to a target site. Several approaches for targeted pairwise shared haplotype have been developed in the previous chapters. But, to first establish *proof of concept* of the age estimation method, the framework was evaluated given complete knowledge of the underlying shared haplotype structure. That is, the "true" shared haplotype segments were taken from simulation records and used to determine the values of model parameters such as the number of pairwise differences or the genetic length.

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. The length of the simulated region was 100 Mb (100 cM), resulting in 326,335 variant sites.

The breakpoints of shared haplotype intervals 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. Note that this allowed overestimation of the actual genetic length of the IBD segment, but thereby provided a realistic benchmark for comparisons with IBD detection methods

1.2.4 Validation of the method under different thresholds

Because an exhaustive analysis of all possible discordant pairs becomes computationally intractable, it is convenient to reduce the number of pairwise analyses that are conducted per target allele. For example, although the sample size of dataset $\mathcal{D}_A$ was modest

(N = 1,000), the total number of possible pairwise analyses for the set of 10,000 selected rare variants would have been 145.725 million. For realistic applications of the method, it is therefore essential to limit the number of discordant pairs, n_d, such that Ω_d consists of a substantially smaller set of randomly formed pairs. In this section, I analyse the impact on the accuracy of estimated allele age under different nominal thresholds of n_d (listed below). Importantly, to focus on the impact resulting from different n_d thresholds, the analysis was conducted using true IBD segments as determined from simulation records. Thus, this section provides a general validation analysis of the age estimation method.

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

Each clock model was considered separately and the same set of 10,000 target sites was analysed under each threshold. This resulted in a total of 276.133 million pairwise analyses in this section alone. None of the analyses returned conflicting results; recall that *conflicts* were defined as invalid estimates resulting from erroneous patterns of coalescent times as computed through the CCF for the set of pairs considered. Note that discordant pairs were formed randomly and therefore differed in each analysis. The results are 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 true age is set at t_m, but t_c and t_d are indicated in Figure 1.3.

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.

Interestingly, the recombination clock, $\mathcal{T}_\mathcal{R}$, showed a tendency to underestimate allele age at higher thresholds, despite using true IBD segments. This observation may be the result of an overestimation of true IBD lengths, since IBD breakpoints were determined from the set of variant sites observed in the data, to provide a realistic

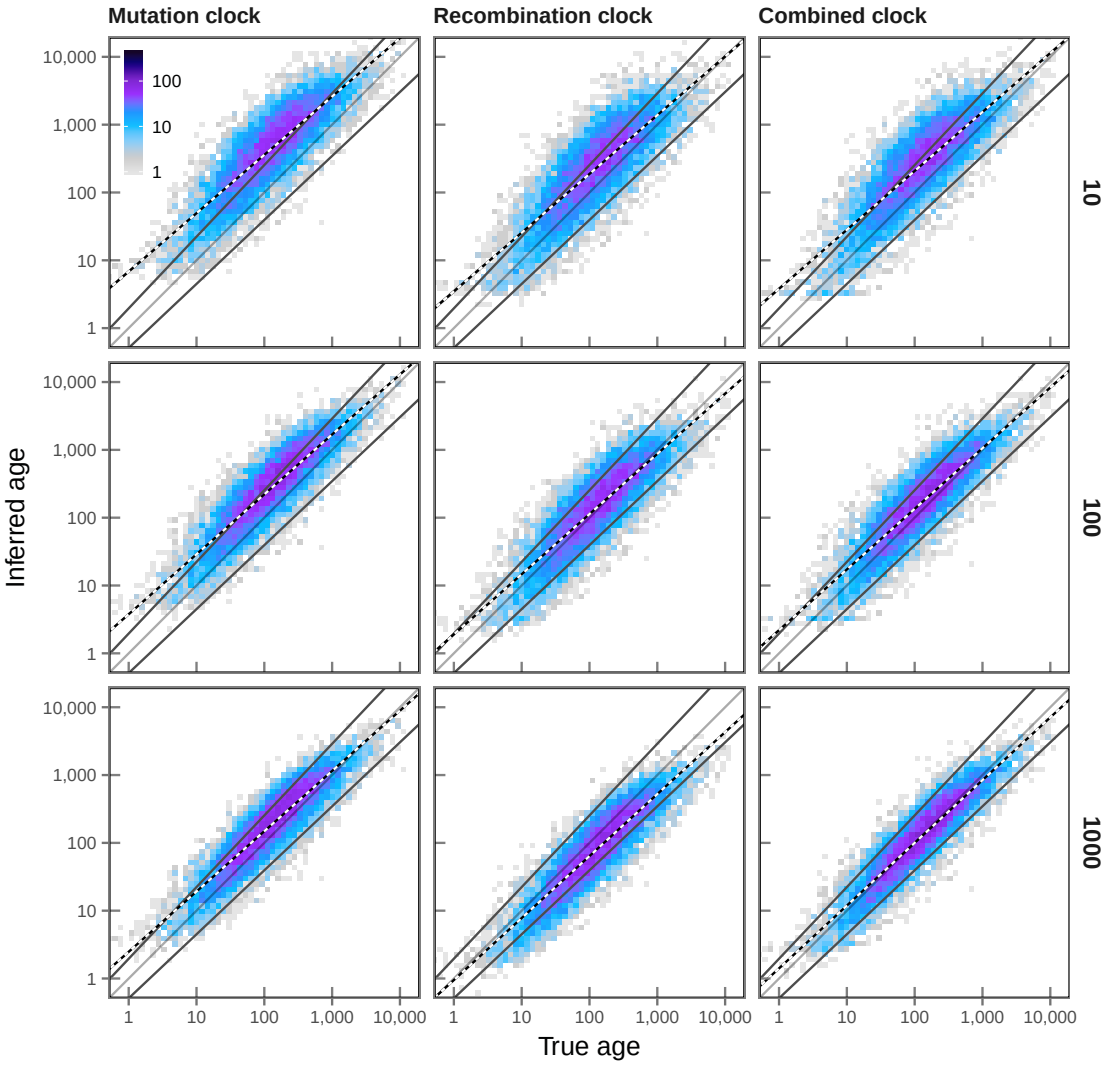


Figure 1.3: True and inferred age under varying numbers of discordant pairs. A set of 10,000 target sites was randomly drawn in $f_{[2,20]}$ (shared allele frequency ≤ 1%) in a simulated sample of 2,000 haplotypes. Different numbers of sampled discordant pairs were analysed on the same set of target 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). The true age of a focal allele was set at 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; these are indicated by their linear 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. Note that both true and inferred age are compared on log-scale, as the time to a coalescent event is expected to increase exponentially back in time.

benchmark for comparisons with IBD detection methods (see next section). Note that allele age is generally expected to be underestimated if genetic lengths in concordant or discordant pairs are overestimated, as a longer IBD segment is indicative for more recent haplotype sharing (*i.e.* recombination had less time the break down the length of a shared haplotype). The average distance between consecutive variant sites in $\mathcal{D}_A$ was 3.064×10^{-4} cM (306.431 basepairs), showing that even small inaccuracies in IBD can affect the estimation of allele age (under the recombination clock).

The proportion of target alleles for which age was correctly estimated 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 (next page). 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. 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.

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, root mean squared logarithmic 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

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 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.

In summary, these results demonstrate that the method as well as the clock models proposed are able to estimate allele age from IBD information alone, without prior knowledge of the demographic history of the sample. However, because data were

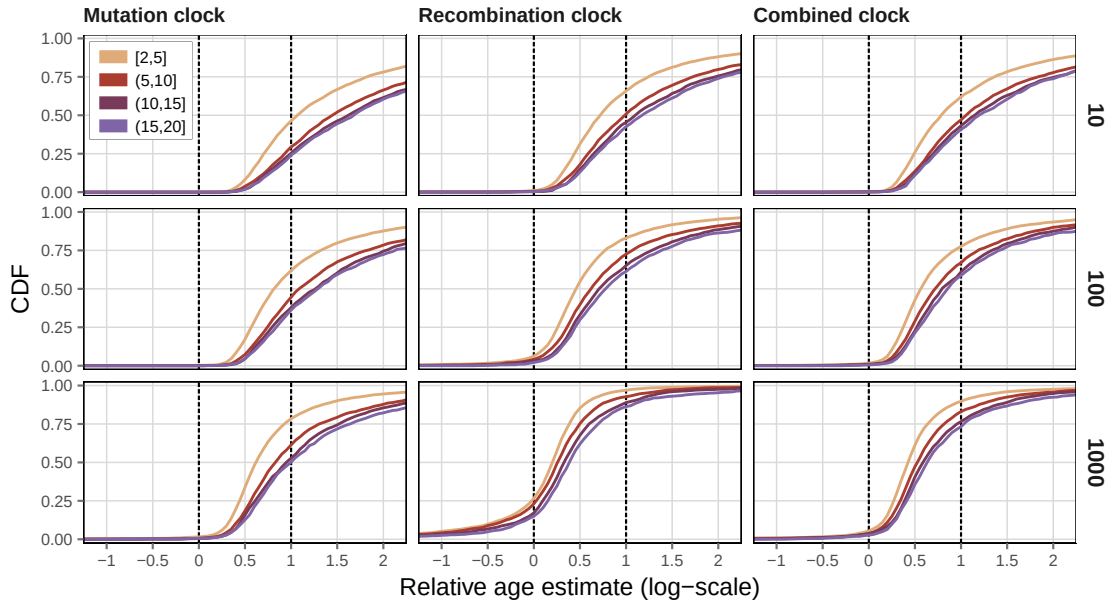


Figure 1.4: Relative age under varying numbers of discordant pairs. A randomly drawn set of 10,000 target sites at allele frequency ≤ 1%, *i.e.* $f_{[2,20]}$, was analysed under each of the three clock models (indicated at the *top* of each column) and with different numbers of sampled discordant pairs; $n_d = 10$, $n_d = 100$, 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 (1.24), 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 $f_{(15,20]}$.

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 further below). 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 process seeks 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 at t_d and are therefore most 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 three 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, the identification of nearest neighbours may be less accurate if only genotype data are available. Both the DGT and the HMM-based approach implemented in `rvage` are able to infer IBD in absence of haplotype information; thus, a method to identify nearest neighbours in genotype data would be required to achieve full compatibility with the algorithm. Regardless, third, a dilemma arises in presence of genotype 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 would bias (or cancel out) the resulting composite likelihood distribution. In such cases, the estimated age is expected to be approximately equal to or smaller than t_c, such that $\hat{t}$ is likely to be underestimated.

It is important to note that the problem of finding false negatives in the data (if genotype error is present) 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. Subsequent analyses were conducted using a threshold equal to the diploid sample size, N; that is $n_d = 1{,}000$ in analyses using $\mathcal{D}_A$, and $n_d = 2{,}500$ using $\mathcal{D}_B$ or $\mathcal{D}_B^*$. Since the results presented in this section were obtained on true IBD information, they serve as a benchmark against which different IBD detection methods are compared in the section below.

1.2.5 Inference of IBD around shared and unshared alleles

The age estimation method relies on the inference of the underlying IBD structure of the sample. In particular, IBD around a given target position is detected in each pair in Ω_c and Ω_d in order to obtain the parameter values required by the clock model used. This is accomplished through the targeted IBD detection methodology incorporated from the `tidy` algorithm; namely the FGT, DGT, and the HMM, which detect IBD in pairs of diploid individuals. However, these methods were originally designed to detect IBD

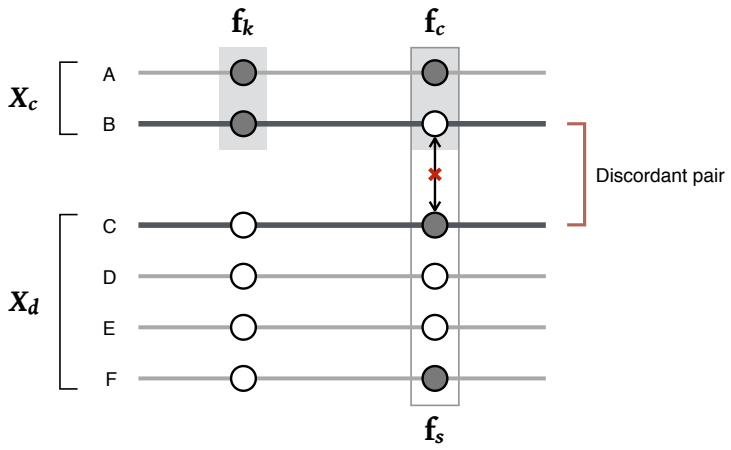


Figure 1.5: Breakpoint detection in discordant pairs. A discordant pair is formed by one haplotype from X_c (which share the focal allele) and one haplotype from X_d (which do not share the focal allele). The lines indicate the chromosomal sequence where the alleles at two sites are indicated; allelic 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.

segments in individuals sharing a focal allele. While this condition is fulfilled when considering concordant pairs, the IBD detection in discordant pairs is problematic as these are defined by not sharing the focal allele.

Recall that 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.

To maintain the variant-centric concept, breakpoints are detected in discordant pairs as follows. Let f_k denote the number of allele copies at the 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, and f_s the number of allele copies 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 20).

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. This may reduce the information available from the sample, but the effect is expected to be negligible, in particular if the focal allele is rare. Since haplotype data are required to determine pairwise differences, S, along haplotype sequences, $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$ cannot be used with genotype data.

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; for example, $(1, 0)$ and $(1, 2)$, where 0 denotes a genotype homozygous for the ancestral allele, 1 a heterozygous genotype, and 2 a genotype homozygous for the derived allele. 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.

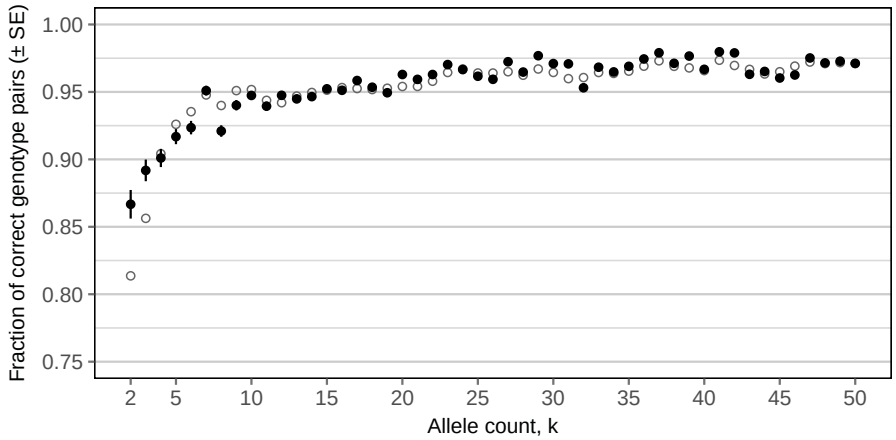


Figure 1.6: Initial state probability of discordant pairs in the Hidden Markov Model (HMM). The proportion of discordant pairs that were correctly identified by their genotypes was empirically determined from data before and after the inclusion of realistic genotype error rates. The mean per f_k was 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).

The HMM-based approach 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 by 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). To extend the model to consider discordant pairs, here, initial state probabilities were estimated as the true positive rate of observing the focal allele as a heterozygous genotype in the X_c individual and not observing the focal allele in a homozygous genotype, g_0, in the X_d individual; again, based on the comparison between genotype data before and after error (using the same dataset as available in Chapter 4). For each f_k category, I randomly selected 1,000 target sites in the dataset before error and randomly selected 1,000 discordant pairs per target site, which I then compared to the genotypes observed 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 probability distribution is shown in Figure 1.6 (page 21); the initial state probabilities used for discordant pairs are indicated for comparison. Notably, the discordant probability of initialisation is similar to the concordant one. A possible explanation is that this is particularly driven by the heterozygous status being false.

REMOVED Section “Anticipated limitations”

1.3 Evaluation

The method was assessed using data generated in coalescent simulations. First, the validity of the method under each clock model was demonstrated based on the true IBD structure of the sample as known from simulation records. Second, the analysis was repeated for each IBD detection method. Third, each approach was then assessed with regard to genotype error, which also considered the effects of phasing error.

1.3.1 Data generation

The performance of the age estimation method was evaluated using several simulated datasets.

Second, the dataset simulated in Chapter 3 was included here to evaluate the age estimation method in presence of genotype 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). Here, this permitted 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. 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 (see Chapter 4).

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. Note that this allowed overestimation of the actual genetic length of the IBD segment, but thereby provided a realistic benchmark for comparisons with IBD detection methods; namely the FGT, DGT, and the HMM-based approach as implemented in the `rvage` algorithm.

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

$$t_m = \sqrt{t_c \, t_d} \,. \quad \text{CORRECTED} \tag{1.23}$$

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.*plotting 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 is 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.. In addition, the 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.

$$\hat{t}_{rel} = \frac{\log\left[\frac{\hat{t}}{t_c}\right]}{\log\left[\frac{t_d}{t_c}\right]} \tag{1.24}$$

As a result, the times of coalescent events at t_c and t_d are mapped to 0 and 1, respectively. An age estimate is defined as being "correct" if $t_c \leq \hat{t} \leq t_d$, which is equal to the condition $0 \leq \hat{t}_{rel} \leq 1$, such 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.

1.4 Results

In each dataset, 10,000 rare variants were randomly selected as target sites for estimation of allele age. These were selected at shared allele frequency $\leq 1\,\%$, *i.e.* $\mathrm{f}_{[2,20]}$ variants, in $\mathcal{D}_A$. Identical sets of target sites were randomly selected in $\mathcal{D}_B$ and $\mathcal{D}_B^*$, at shared allele frequency $\leq 0.5\%$ ($\mathrm{f}_{[2,25]}$ variants). Note that these were sampled from the subset of variants unaffected by genotype error, to ensure that alleles correctly identified haplotype sharing.

1.4.1 Comparison of IBD detection methods

The `tidy` algorithm for targeted IBD detection (see Chapters 3 and 4) was fully integrated in `rvage`, such that the FGT, DGT, and the HMM-based approach were available for the inference of IBD segments around focal variants. Note that genotype data are sufficient for IBD detection using the DGT and HMM, but haplotypes are required for estimation under the mutation clock model; *i.e.* to count pairwise differences, S, along haplotype sequences. Thus, analyses were conducted on the simulated haplotype dataset ($\mathcal{D}_A$), but haplotype phase was ignored during IBD detection in the DGT and HMM. The parameters required by the `rvage` algorithm were specified accordingly with simulation parameters ($N_e = 10{,}000$; $\mu = 1 \times 10^{-8}$ per site per generation; $\rho = 1 \times 10^{-8}$ per site per generation). Here, because simulated data did not include genotype error, theoretical emission model was used in the HMM.

The results presented in this section were obtained on the previously selected 10,000 rare allele target sites, which were analysed using each of the three IBD detection methods and under each clock model, resulting in a total of 93.295 million pairwise analyses. The fraction of conflicting age estimates differed by clock model as well as IBD detection method; no conflicting estimates were returned when true IBD was used. Under the mutation clock, $\mathcal{T}_{\mathcal{M}}$, analyses using the FGT returned 1.809 % conflicts. This fraction was higher using the DGT and HMM, with 2.601 % and 2.327 %, respectively. Conflicts were seen less under the recombination clock, $\mathcal{T}_{\mathcal{R}}$, where none were returned using the FGT, but 0.010 % and 0.030 % using the DGT and HMM. The fraction under the

combined clock, $\mathcal{T}_{\mathcal{MR}}$, was smaller compared to $\mathcal{T}_{\mathcal{M}}$, with 1.097 %, 2.266 %, and 1.819 % of conflicted sites using the FGT, DGT, and HMM, respectively. The remaining sites were intersected to compare clock models and IBD methods on the same set of target sites, retaining 9,434 variants.

The density distribution 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 age distributions in $\mathcal{T}_{\mathcal{M}}$ and under the combined clock model, $\mathcal{T}_{\mathcal{MR}}$, in which alleles 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 the recombination clock, $\mathcal{T}_{\mathcal{R}}$, at 72.610 %, and lowest under the mutation clock, $\mathcal{T}_{\mathcal{M}}$, with 34.460 %; under the combined clock, $\mathcal{T}_{\mathcal{MR}}$, 55.395 % of alleles were correctly estimated when the FGT was used. The HMM achieved similar levels of accuracy, but the accuracy in $\mathcal{T}_{\mathcal{R}}$ was noticeably reduced (10.950 %) compared to $\mathcal{T}_{\mathcal{MR}}$ (51.876 %) and $\mathcal{T}_{\mathcal{M}}$ 32.415 %. Throughout, the lowest proportions of correctly inferred alleles were found for the DGT, which also showed the lowest accuracy in $\mathcal{T}_{\mathcal{R}}$ (8.226 %) and comparatively low levels of accuracy in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$ (14.554 % and 29.659 %, respectively). Overestimation of allele age was highest in $\mathcal{T}_{\mathcal{M}}$, where 65.084 %, 85.277 %, and 66.960 % of alleles were underestimated by the FGT, DGT, and HMM, respectively. Conversely, the proportion of underestimated alleles was lowest in $\mathcal{T}_{\mathcal{M}}$, at $\leq 1\%$ in each method, and similarly low in $\mathcal{T}_{\mathcal{MR}}$ with $\leq 2\%$ in each method. In contrast, alleles were markedly underestimated in $\mathcal{T}_{\mathcal{R}}$; the FGT resulted in 20.140 % of underestimated alleles, whereas 91.753 % and 88.934 % of alleles were underestimated when the DGT and the HMM were used for IBD inference, respectively.

The accuracy measured for each analysis is summarised in Table 1.2 (page 28). The FGT under the recombination clock model, $\mathcal{T}_{\mathcal{R}}$, showed a higher correlation and slightly reduced error with regard to t_d. There, rank correlation was $r_S = 0.899$ for the FGT and $r_S = 0.889$ for true IBD; likewise the magnitude of error (RMSLE) was 0.339 and 0.345 for FGT and true IBD, respectively. However, note that a higher accuracy at t_c does not necessarily reflect an improvement in the estimation of actual allele age. For example, the accuracy with regard to t_m or t_d was lower for the FGT compared to true IBD. In comparison to the other detection methods, the FGT outperformed both the DGT and

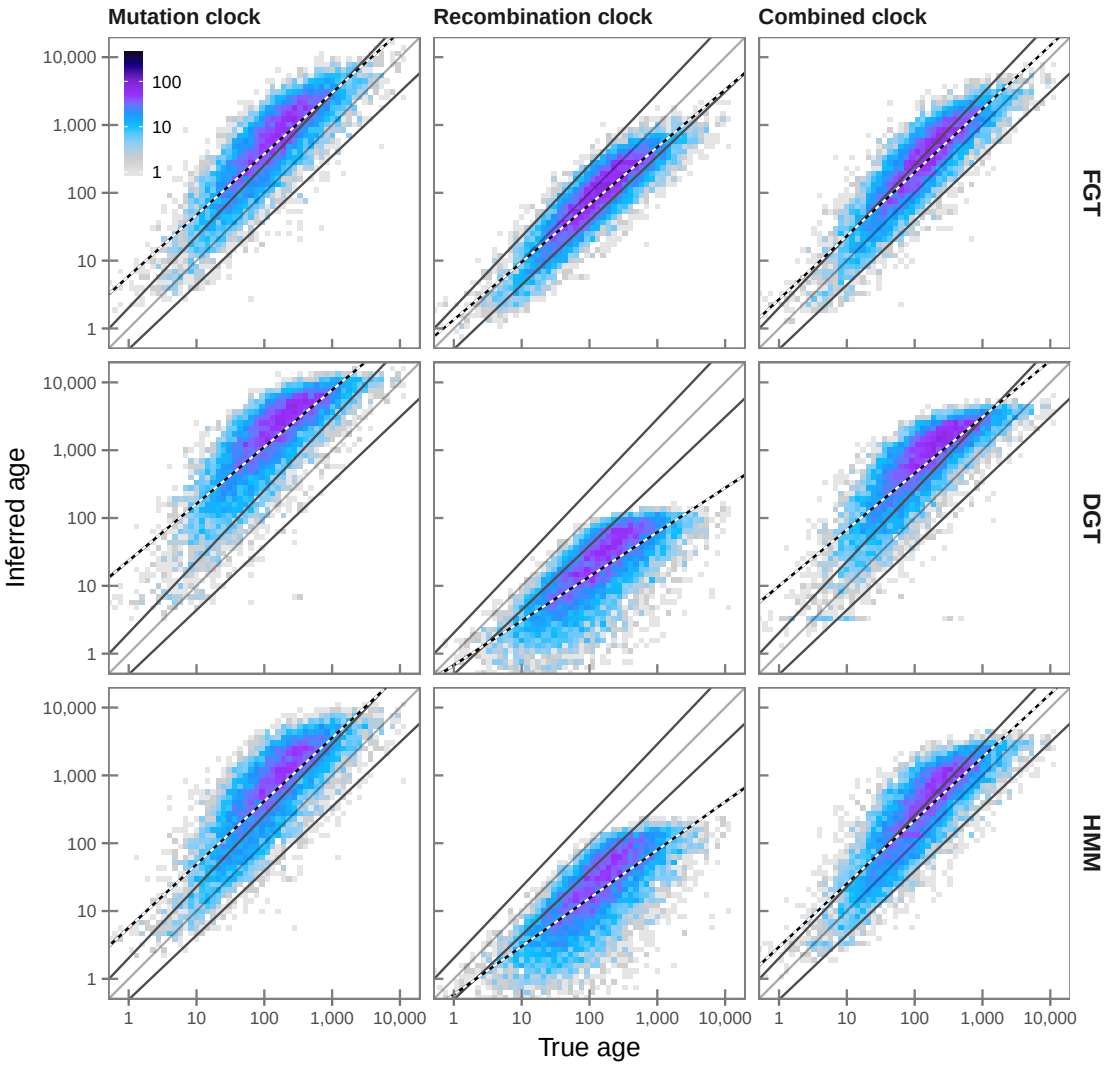


Figure 1.7: Distribution of true and inferred age using different IBD detection methods. The three IBD detection methods implemented in `rvage` were compared, *i.e.* FGT, DGT, and HMM (indicated at the *right* of each row), under each clock model (indicated at the *top* of each column). Analyses were compared on the same set of 9,434 target sites that were drawn from available $f_{[2,20]}$ variants in the simulated dataset of 2,000 haplotypes (allele frequency $\leq 1\%$). Each panel shows the density of true age (t_m) and inferred age (numbers indicated by the colour-gradient). 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 trend line of inferred age ($\hat{t}$) is indicated by the *black-white* line, using the posterior mode of the composite likelihood estimation as the inferred age value.

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 17) 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

HMM with regard to each time measure. The HMM showed slightly higher levels of accuracy than the DGT in $\mathcal{T}_{\mathcal{R}}$, where r_S was higher and RMSLE lower in terms of each time measure for the HMM. Similarly, in both $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$, RMSLE scores were lower for the HMM compared to the DGT, whereas r_S measures were similar.

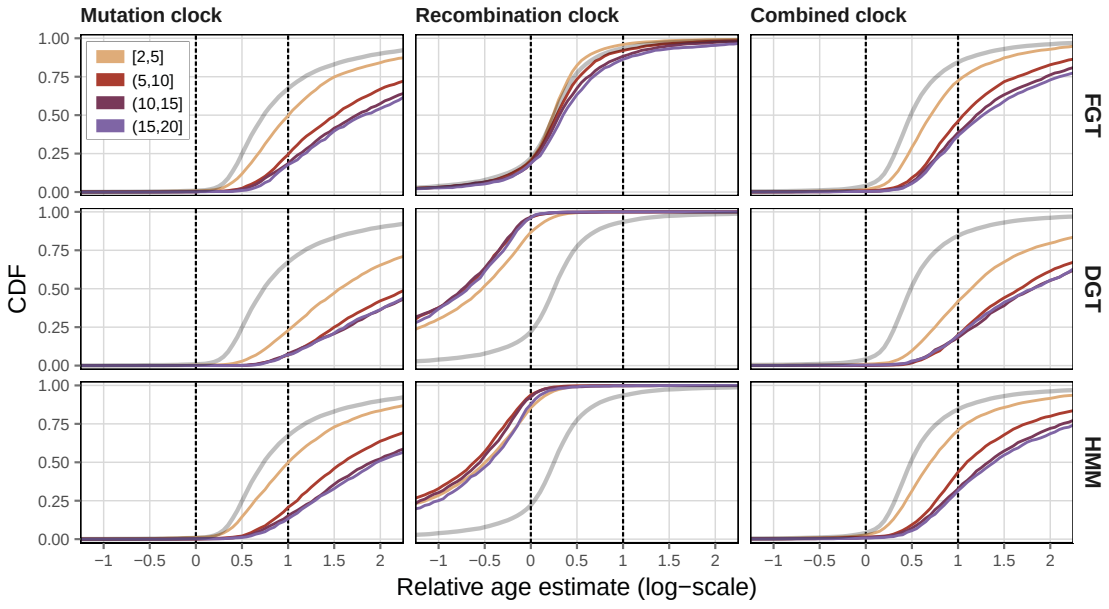


Figure 1.8: Relative age using different IBD detection methods. The three IBD detection methods implemented in `rvage` were compared, *i.e.* FGT, DGT, and HMM (indicated at the *right* of each row), under each clock model (indicated at the *top* of each column). The relative age, $\hat{t}_{rel}$, was calculated as given in Equation (1.24), such that t_c and t_d sit at 0 and 1, respectively (*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 $f_{(15,20]}$.

Relative age estimates are shown for distinct f_k ranges in Figure 1.8 (page 28), where the relative age of true IBD is indicated for comparison per clock model (calculated on the full f_k range). Analyses under the mutation clock and the combined clock models, $\mathcal{T}_{\mathcal{M}}$ 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 the recombination clock model, $\mathcal{T}_{\mathcal{R}}$, but the DGT showed an accuracy decrease for $f_{[2,5]}$ variants.

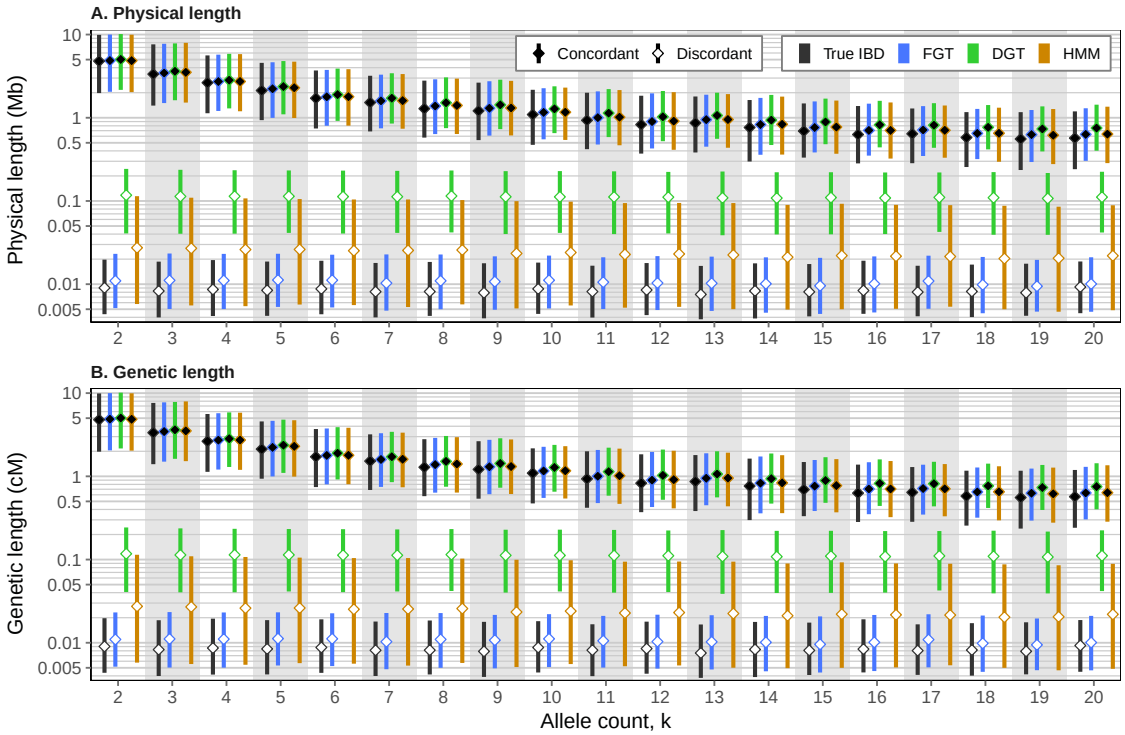


Figure 1.9: Length distribution of inferred IBD segments. Bottom and top of each bar indicate 1st and 3rd quartiles, respectively, between which the median (2nd quartile) is marked (*diamonds*). IBD detected for concordant and discordant pairs is distinguished; *solid* and *hollow* diamonds, respectively.

The distribution of IBD lengths inferred using the FGT, DGT, and the HMM-based approach are shown in Figure 1.9 (this page). Segments inferred using the HMM were close to those detected using the FGT in concordant pairs. However, for discordant pairs, only the FGT produced IBD segments that were close to the length distribution of true IBD segments. The DGT showed the highest degree of overestimation for both concordant and discordant pairs.

In summary, these results suggest that the accuracy of estimated allele age is crucially dependent on correct inference of the underlying IBD structure. The overestimation of IBD lengths, which is generally expected for each method, affected each clock model

differently. While $\mathcal{T}_{\mathcal{M}}$ overall resulted in an overestimation of allele age when IBD is overestimated, this pattern was reversed in $\mathcal{T}_{\mathcal{R}}$. Although both models are combined in $\mathcal{T}_{\mathcal{MR}}$, the impact of mutational differences, seen at the overestimated regions of detected IBD segments, was substantial and could not be mitigated by considering recombinational length. Further, I confirmed that the FGT was the best performing method for the targeted detection of IBD segments, in that the estimation of allele age was similar to the expectations defined by true IBD information. 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.

1.4.2 Impact of genotype error on allele age estimation

The allele age estimation method was evaluated under each clock model and each method for IBD detection, on data before and after the inclusion of genotype error; *i.e.* using datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$, respectively. Each analysis was performed on the same set of 10,000 target sites selected at $f_{[2,25]}$ in a sample of N = 2,500 diploid individuals, using a threshold of n_d = 2,500, which amounted to 25.281 million pairwise analyses per comparison. The FGT was applied to both the simulated (*true*) haplotypes as well as phased haplotype data. The HMM used theoretical emission model in analyses on $\mathcal{D}_B$ and the empirical error model in analyses on $\mathcal{D}_B^*$. To enable direct comparisons, true IBD segments were determined from simulation records and separately analysed on the same number of concordant and discordant pairs in data before and after error. In total, for the results presented in this section, 758.437 million pairwise analyses were conducted.

As in the previous section, some of the analyses returned conflicting estimates; see Table 1.3. Again, no conflicts were seen when true IBD information was used. However, this changed after the inclusion of genotype error; the fraction of conflicting estimates was high in $\mathcal{T}_{\mathcal{M}}$, zero in $\mathcal{T}_{\mathcal{R}}$, and small in $\mathcal{T}_{\mathcal{MR}}$. Before error, the largest fraction of conflicts was seen for the DGT in $\mathcal{T}_{\mathcal{M}}$. Data from analyses before and after error were intersected across results obtained under each clock model and for each IBD method, which retained a set of 5,015 identical target sites. A complete summary of the accuracy per analysis is given below in Table 1.4 (page 38).

Table 1.3: Conflicted estimates in analyses before and after error.

Method	Conflicts before error (%)			Conflicts after error (%)		
	$\mathcal{T}_{\mathcal{M}}$	$\mathcal{T}_{\mathcal{R}}$	$\mathcal{T}_{\mathcal{MR}}$	$\mathcal{T}_{\mathcal{M}}$	$\mathcal{T}_{\mathcal{R}}$	$\mathcal{T}_{\mathcal{MR}}$
FGT*	6.396	0.000	3.695	5.131	0.141	2.189
FGT**	6.587	0.422	4.388	4.940	0.341	3.123
DGT	10.945	0.161	8.384	5.211	1.767	3.956
HMM	5.884	0.392	4.418	13.335	0.823	9.268
True IBD	0.000	0.000	0.000	9.583	0.000	1.030

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

Estimation based on the true IBD structure of the sample is compared before and after error in Figure 1.10a (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, $\mathcal{T}_{\mathcal{MR}}$. Alleles were overestimated because the presence of misclassified alleles substantially increased the number of observed mutational differences, S, along the sequence. For example, accuracy 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 respectively, similarly 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 75.394 % before and 24.068 % after error, which was similar in $\mathcal{T}_{\mathcal{MR}}$, where 80.518 % of alleles were correct before but only 39.402 % after error. The proportion of overestimated alleles was 18.046 % in $\mathcal{T}_{\mathcal{M}}$ and 9.212 % in $\mathcal{T}_{\mathcal{MR}}$ before error, but 74.397 % and 57.926 %, respectively, after error. Note that this did not vary noticeably by focal allele frequency; for example, the proportion of overestimated alleles in $\mathcal{T}_{\mathcal{M}}$ was 75.659 % at lower frequencies ($f_{[2,5]}$) and 79.375 % at higher frequencies ($f_{[20,25]}$), which was also the case in $\mathcal{T}_{\mathcal{MR}}$, where 61.831 % and 1.250 % of alleles were overestimated at $f_{[2,5]}$ and $f_{[20,25]}$, respectively.

In contrast, 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 recombinational 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. As in the previous analysis, alleles showed a tendency to be underestimated in $\mathcal{T}_{\mathcal{R}}$. The average distance between consecutive SNPs was 1.609×10^{-4} cM (93.557 basepairs) in $\mathcal{D}_B$ and $\mathcal{D}_B^*$; *i.e.* the density of variant sites is higher compared to $\mathcal{D}_A$, such that a potential bias resulting from overestimation of true IBD lengths is expected to be reduced. Overall, 42.891 % of alleles were correctly inferred, but this was higher for at $f_{[2,5]}$ and lower at $f_{[20,25]}$; 48.681 % and 39.375 %, respectively. The proportion of underestimated alleles

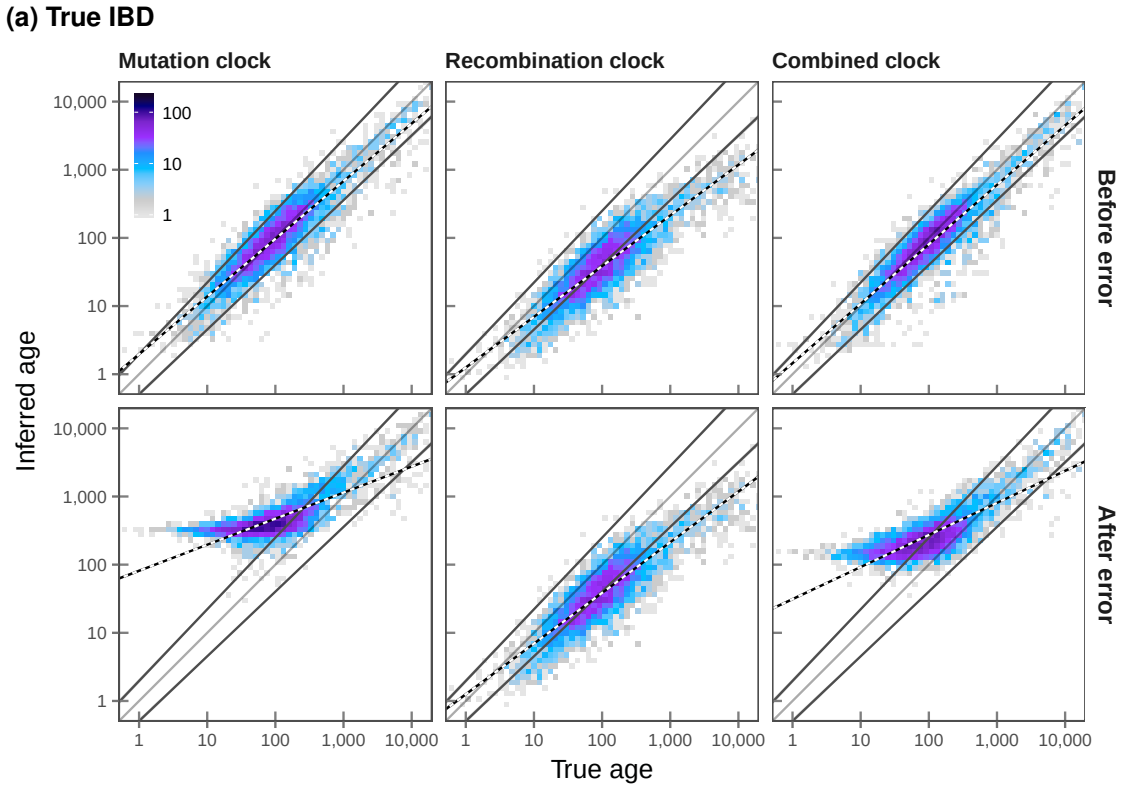
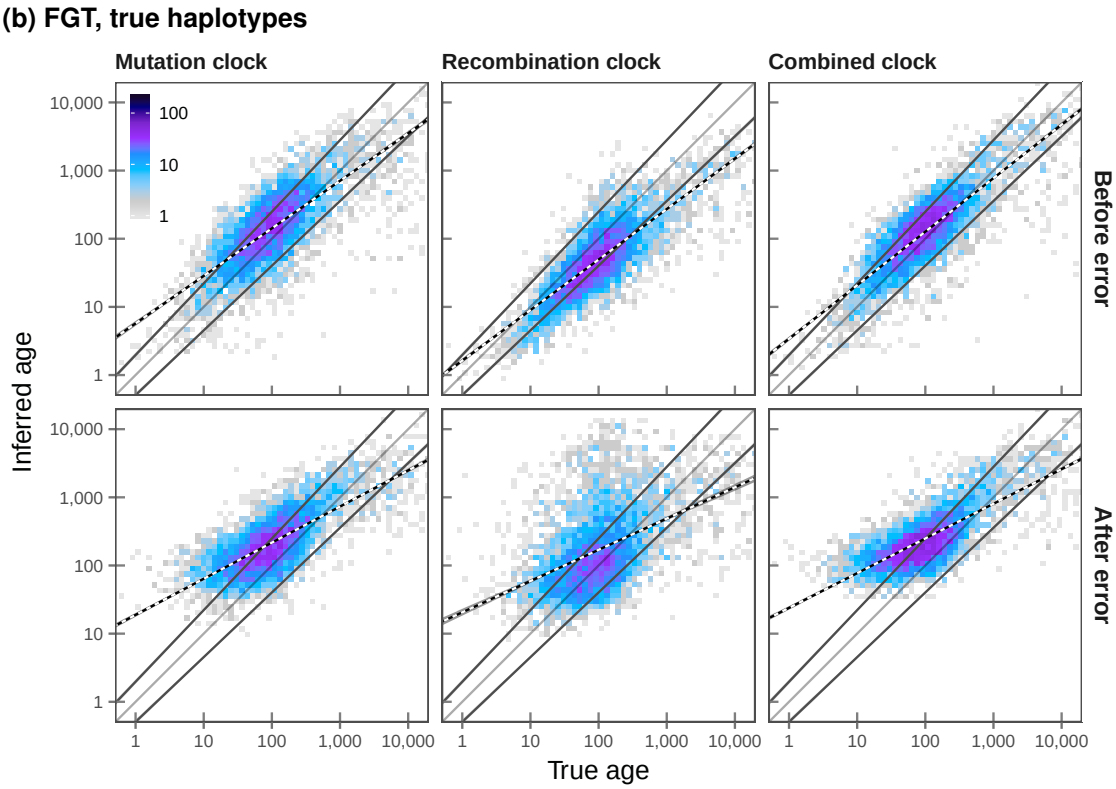


Figure 1.10: Density distribution of allele age before and after the inclusion of genotype error in simulated data. Allele age estimation was conducted on data in which empirical distributions of genotype error were simulated. The effects on the estimation process *before* and *after* error are compared (*top* and *bottom*, respectively). The dividing line is fixed at the true age (t_m), around which the lines *below* and *above* correspond to the regression trend lines of the times of coalescent events delimiting the branch on which focal mutations sit; *i.e.* t_c and t_d, respectively. The *black-white* line indicates the regression trend of 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 35), and the HMM-based approach (**e**; page 36). Each analysis was conducted on the same set of retained 5,015 target variants at allele frequency $\leq 0.5\%$ in simulated data of $N = 2{,}500$ diploid individuals.

was 55.553 %, where 50.528 % and 52.500 % were underestimated at $f_{[2,5]}$ and $f_{[20,25]}$, respectively. The correlation between inferred and true age was generally high (r_S: 0.818, 0.843, and 0.666 at t_c, t_m, and t_d, respectively) but nonetheless slightly lower compared to corresponding results from dataset $\mathcal{D}_A$ (0.889, 0.895, and 0.739, respectively); although, note that these results are not directly comparable as the underlying demographies were different and only half the number of target sites was analysed here.

When IBD was inferred, the accuracy of the estimation analysis was differently affected dependent on the IBD detection method used. Results based on the FGT are shown in Figure 1.10b and 1.10c (next page), which compare age estimates obtained



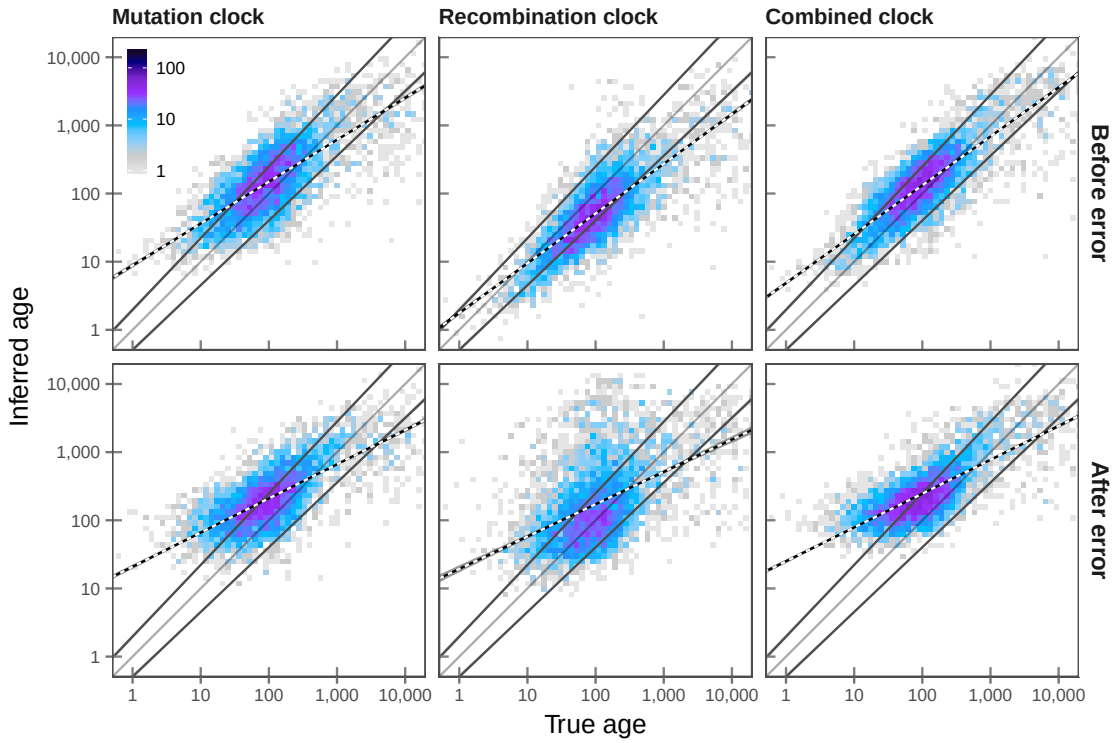


Figure 1.10: Continued.

on the same set of target sites based on IBD detected in true and phased haplotypes, respectively, both before and after error. Without genotype 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 % of correct alleles in $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively. Note that the proportion of correctly inferred alleles increased in $\mathcal{T}_{\mathcal{R}}$ due to phasing error. This is because the underestimation that was generally seen under the recombination clock model may have been mitigated by further reduction of IBD segment lengths resulting from flip or switch errors in phased data. The small difference between true and phased data was further reflected in the accuracy of each analysis, where r_S 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.

When analyses were performed on data with genotype error, the overall proportion of correct alleles was reduced, but again the differences seen from 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 in $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively. Likewise, accuracy was overall reduced but r_S and RMSLE scores did not suggest notable differences between estimation results from true and phased haplotypes; see Table 1.4 (page 38). Notably, the analysis on true IBD suggested that genotype error induces an overall overestimation of allele age in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$. However, this effect was mitigated by underestimating IBD lengths in the FGT, such that the number of pairwise differences, S, may not be elevated as genotype errors that would increase the value of S may also lead to the premature detection of interval breakpoints.

Estimation results based on the DGT for IBD detection are shown in Figure 1.10d (next page). Before error, the proportions of correctly inferred allele age were the lowest in the present comparison in each clock model. Under both the mutation and combined clocks, $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$, DGT-based age estimation resulted in 26.341 % and 36.949 % of correct alleles, respectively, whereas only 2.413 % were correct in $\mathcal{T}_{\mathcal{R}}$. While the majority of alleles in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$ were overestimated, 70.050 % and 57.846 % respectively, 97.587 % were underestimated in $\mathcal{T}_{\mathcal{R}}$ (none were overestimated). The tendency to overestimate allele age was increased after error; the proportions of alleles overestimated were 77.308 % and 67.856 % in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$, respectively. As this was also the case in $\mathcal{T}_{\mathcal{R}}$, the proportion of correctly inferred alleles increased to 15.693 %, but this was an artefact resulting from an overall underestimation of IBD lengths. However, the loss in accuracy was reflected

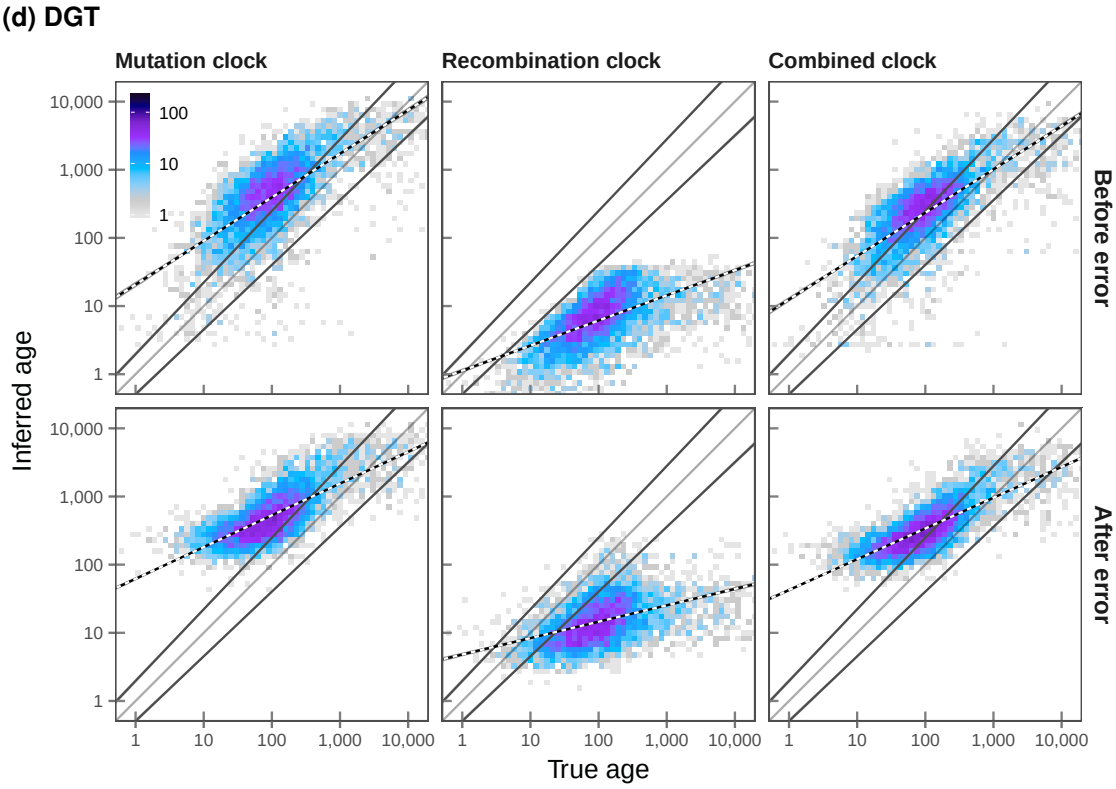


Figure 1.10: Continued.

in the correlation between true and inferred allele age; r_S at t_c, t_m, and t_d was 0.746, 0.628, and 0.406 before error, and 0.588, 0.504, and 0.328 after error. Note that rank correlations at t_m and t_d were higher in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$, both before and after error. However, the same measures taken after error actually suggested that the accuracy increased in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$; see Table 1.4 (page 38). Regardless, rank correlation measured at t_c was decreased after error under each clock model.

The accuracy of age estimation based on IBD inference using the HMM-based approach was overall highly accurate before error; 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 distribution for results obtained using the HMM is given in Figure 1.10e (next 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 $\mathcal{T}_{\mathcal{MR}}$. The majority of alleles was underestimated in $\mathcal{T}_{\mathcal{R}}$ (96.351 %). This was increased after error, *i.e.* 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. For example, RMSLE scores were lowest for the HMM under each clock model after error; see Table 1.4 (page 38). The accuracy before and after error, measured as 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 $\mathcal{T}_{\mathcal{MR}}$. However, importantly, the HMM-based estimation

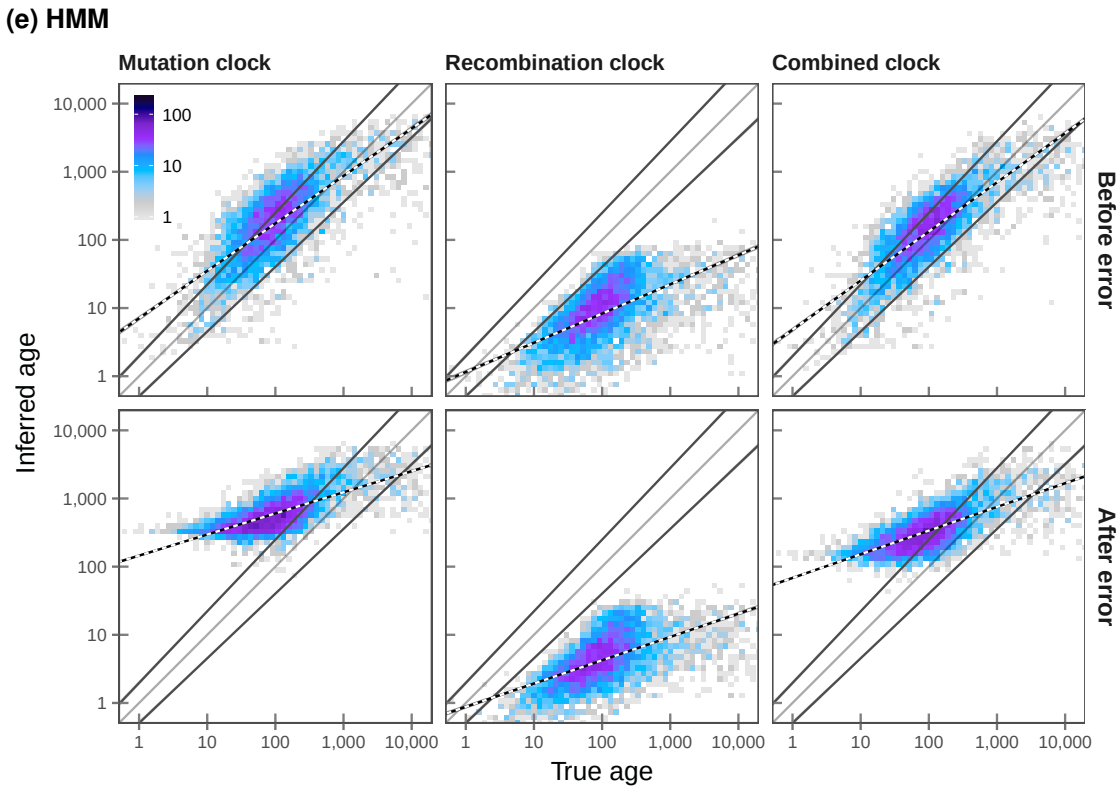


Figure 1.10: Continued.

showed the highest levels of accuracy in $\mathcal{T}_{\mathcal{R}}$ compared to the other methods, *i.e.* r_S at t_c was 0.751 before and 0.737 after error. Although allele age was vastly underestimated, deviations appeared to be consistent.

The distribution of inferred IBD segment lengths for each approach are given in Figure 1.11 (next page). Notably, IBD segments detected using the FGT and DGT were overall underestimated after error; only the HMM maintained similarly accurate lengths before and after error, for both concordant and discordant pairs.

REMOVED Section “Generation of error correction models”

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

1.5 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. The impact of phasing error is small in comparison to genotypic (or allelic) misclassification, which is likely to bias the estimation process.

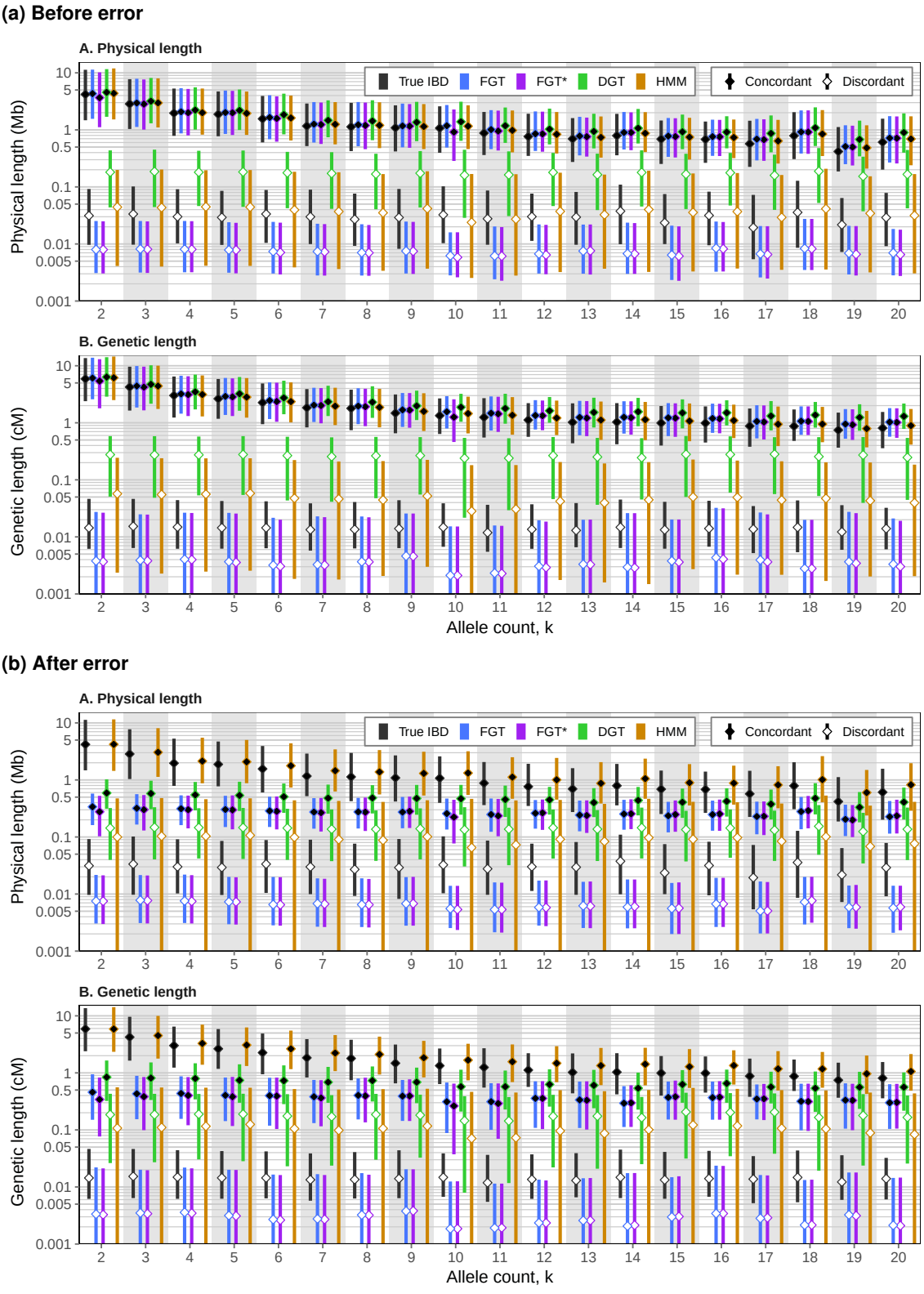


Figure 1.11: Length distribution of inferred IBD segments before and after error. Bottom and top of each bar indicate 1st and 3rd quartiles, respectively, between which the median (2nd quartile) is marked (*diamonds*). IBD detected for concordant and discordant pairs is distinguished; *solid* and *hollow* diamonds, respectively.

Table 1.4: 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 rare allele target sites at shared allele frequency ≤ 0.5% in simulated data of 5,000 haplotypes. The number of discordant pairs was limited to n_d = 2,500 in each analysis. Note that the HMM used the theoretical emission model in the analysis before error (dataset $\mathcal{D}_B$), and the empirical emission model after error ($\mathcal{D}_B^*$). True IBD refers to the first breakpoints that are detectable in the data to both sides of a given target position, which were 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

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. In this regard, the HMM-based approach for IBD inference was shown to maintain consistency even if genotype error is present. However, second, even if IBD is detected with high accuracy, the alleles observed at a focal variant in the sample may wrongly identify haplotype sharing by descent. To account for the possibility that some concordant pairs may actually be discordant pairs, for example, a separate filtering method would be needed to exclude pairs before or after the computation of the CCF, to reduce the chance that the calculation of the composite likelihood is biased. However, because such a method would effectively predict missed alleles in the data, a solution to this problem may not be straightforward. Yet it would be possible, for example, to exclude pairs on basis of patterns of allele sharing or consistency of the inferred IBD structure. Alternatively, instead of excluding pairs, the target site itself would need to be excluded from the analysis if bias is likely. A simple solution was presented in the previous section, where sites are excluded if the lower and upper bounds indicate a reverse order, but further evaluation is required to determine the effectiveness of this filtering criterion.

Lastly, note that both the DGT and the HMM-based approach operate on genotype data to detect IBD, but because the mutation clock model, $\mathcal{T}_{\mathcal{M}}$, requires haplotypes, it would be desirable to estimate pairwise differences, S, in genotype data, so as to make these methods fully compatible with $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$. A possible solution is presented in Chapter 3, where haplotype phase was determined from genotype pairs in detected IBD segments, based on the genealogical constrains that arise under haplotype sharing by descent. Yet, further work is needed to determine the feasibility of such an approach.

1.6 A haplotype-based HMM for shared haplotype inference

1.6.1 Description of the model

1.6.2 Impact of error on allele age estimation

5,000 rare variant sites at $f_{[2,50]}$ were selected at random from the set of sites at which data error was not seen. This ensured that concordant and discordant pairs were 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 in total.

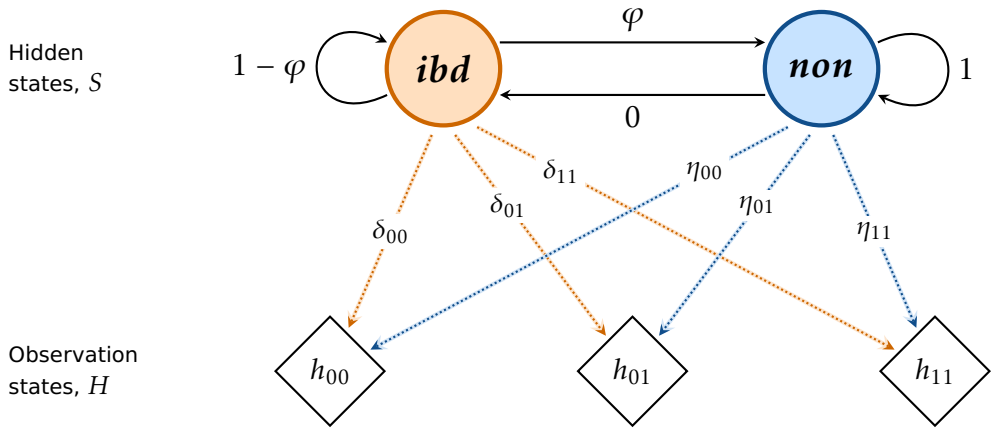


Figure 1.12: Illustration of the Hidden Markov Model for IBD inference. Two hidden states are assumed to generate the observations in a Markov process; *ibd* and *non*. Transitions from each state into any state are indicated by *solid* lines. The probability of transition from *ibd* to *non* is denoted by φ, and from *non* to *ibd* is set to zero; hence, once the Markov chain proceeds into the *non* state it cannot transition back into *ibd*. This is because the IBD process is modelled such that only the innermost IBD segment is inferred, relative to the focal position which sits at the start of the sequence. The input sequence consists of genotype data from a pair of individuals, resulting in six possible observation states; denoted by $g_{k_1 k_2}$, where $k_1, k_2 \in \{0, 1, 2\}$. The probabilities of emitting each possible genotype pair given each hidden state are denoted by $\delta_{k_1 k_2}$ and $\delta_{k_1 k_2}$ 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*.

1.6.2.1 Shared haplotype inference

1.6.2.2 Allele age estimation

1.6.2.3 Discussion

1.6.3 Comparison to the Pairwise Sequentially Markovian Coalescent (PSMC)

Pairwise Sequentially Markovian Coalescent (PSMC) uses an HMM to infer the T_{MRCA} at a locus from the observed sequence of genotypes; *i.e.* (phased) haplotype pairs. emission probabilities transitions represent ancestral recombination events The hidden states are defined as discrete time intervals

1.6.3.1 Implementation to estimate allele age

using the default model parameters; in particular, the number of 64 hidden states.

The boundaries of time intervals in PSMC are calculated as

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

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 $i = 0, 1, \ldots, n$.

Note that I modified the `decode` algorithm implemented in software available for the Multiple Sequentially Markovian Coalescent (MSMC), written in `D`, as it applies the PSMC method when two haplotype sequences are provided as input data. Modifications of `decode` were made to include the option to only return posterior probabilities computed at a specified target position.*

I randomly selected 1,000 target sites 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 in age). At each site, a maximum of 100 concordant and 100 discordant pairs was selected, yielding 187,420 pairwise analyses in total. Pair selection was done randomly, so as to facilitate T_{MRCA} estimation at older genealogical relationships in discordant pairs.

1.6.3.2 Results for the T_{MRCA}

Simulation records were scanned to obtain the true T_{MRCA} for each target site and haplotype pair. The true time was compared to a point estimate taken at the mode of the posterior distribution in PSMC, as well as the mode of the composite posterior in each clock model.

The median was taken as a point estimate from the posterior obtained for a given pair.

Recall that the CCF is defined as the CDF of the posterior; see **??** (page ??).

1.6.3.3 Results for allele age

1.6.3.4 Discussion

1.6.4 Allele age estimation in 1000 Genomes

1.6.4.1 Inferred allele age distribution by population

1.6.4.2 Comparison to PSMC

1.7 Discussion

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

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}	n	Rank correlation (r_S)				RMSLE			
(generations)		$\mathcal{T}_{\mathcal{M}}$	$\mathcal{T}_{\mathcal{R}}$	$\mathcal{T}_{\mathcal{MR}}$	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**

Table 1.6: Accuracy of age inference for different methods. ...

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

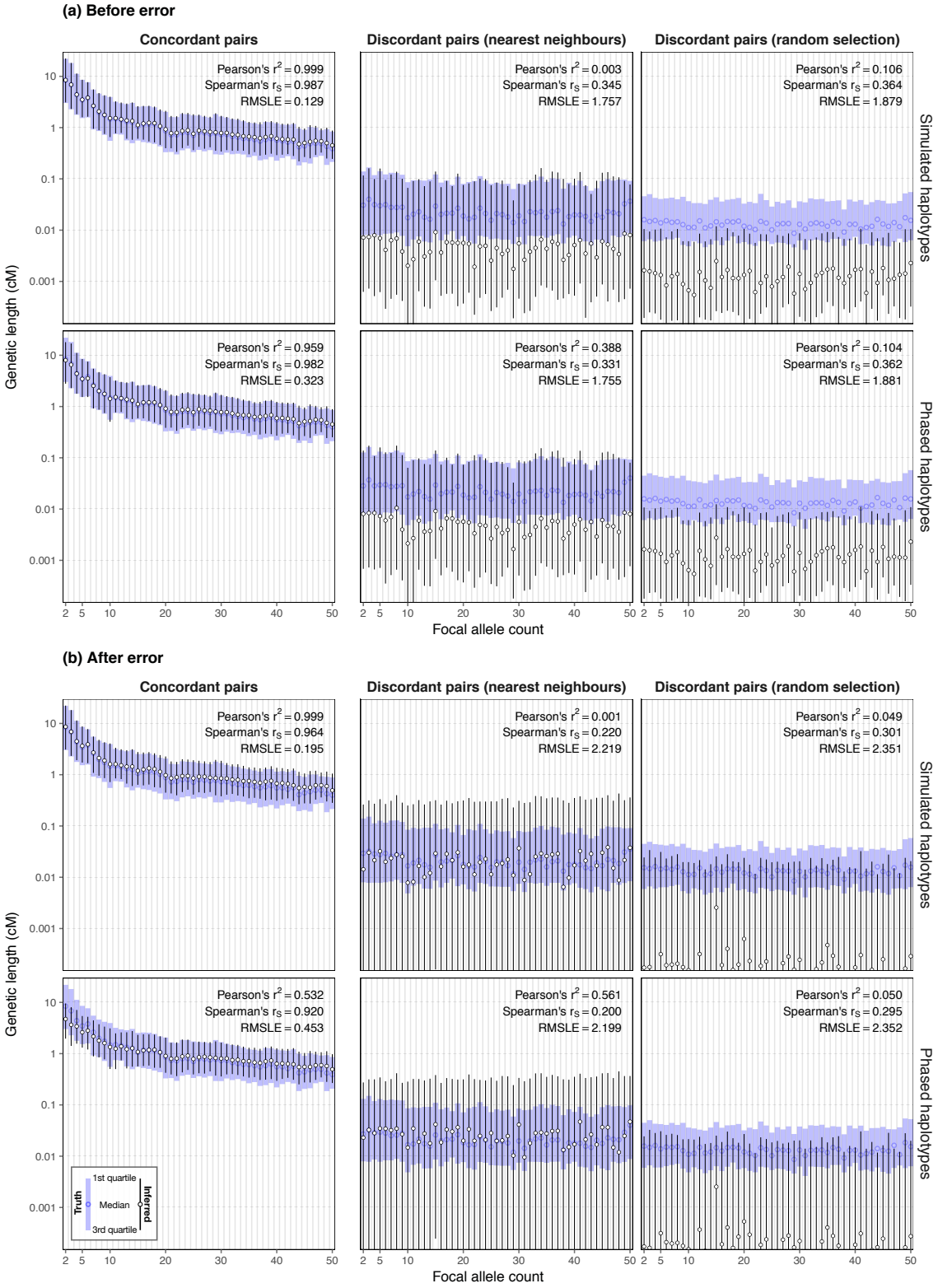


Figure 1.13: Genetic length of shared haplotype segments inferred using the haplotype-based HMM. The distribution of genetic length is shown by allele count of the focal variant in the simulated sample of $N = 5{,}000$ haplotypes, in direct comparison to the corresponding true length at the same set of shared segments (*blue* bars). This is separately shown for concordant discordant pairs. 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)**.

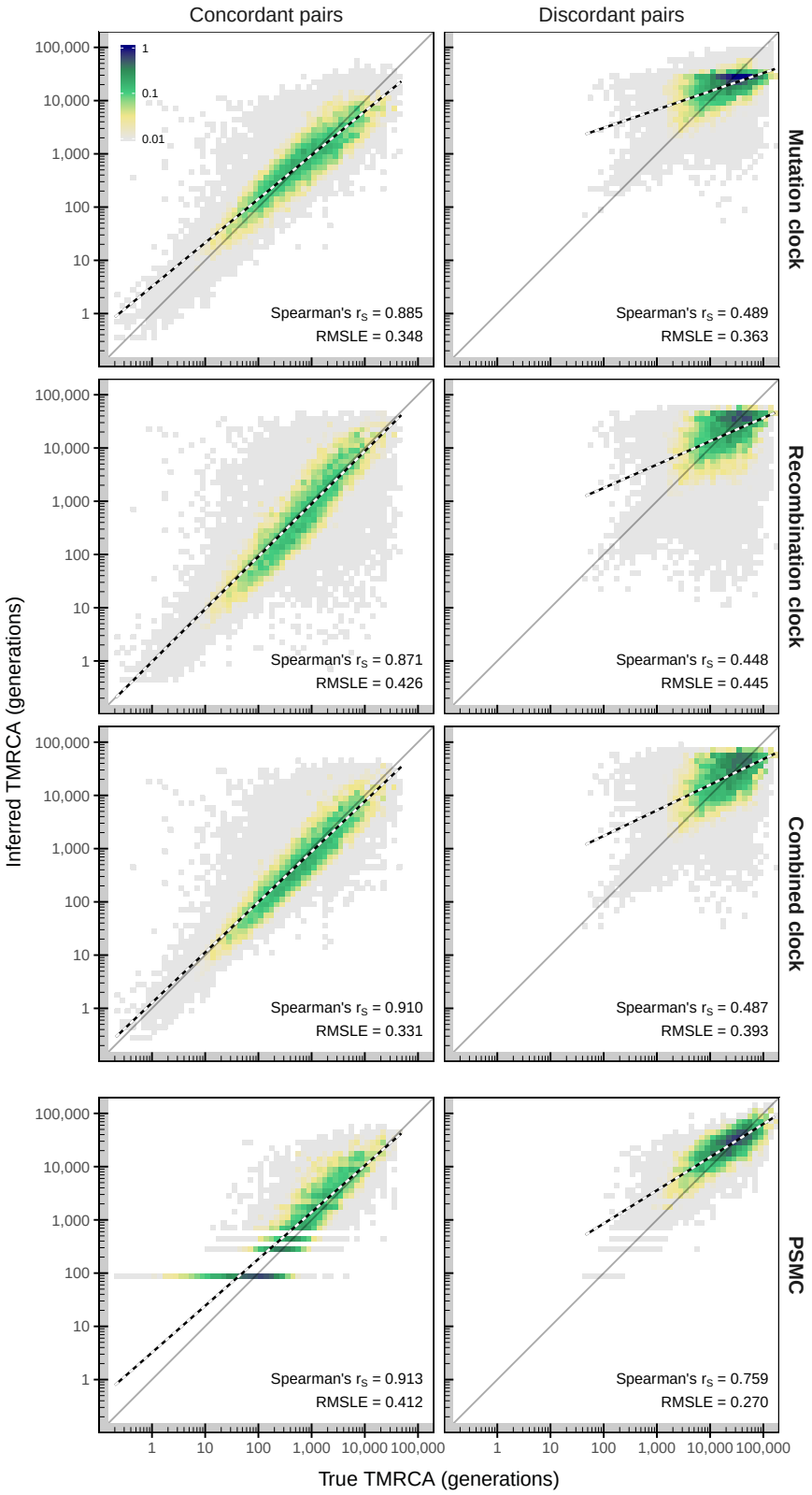


Figure 1.14: True and estimated T_{MRCA} using different methods. ...

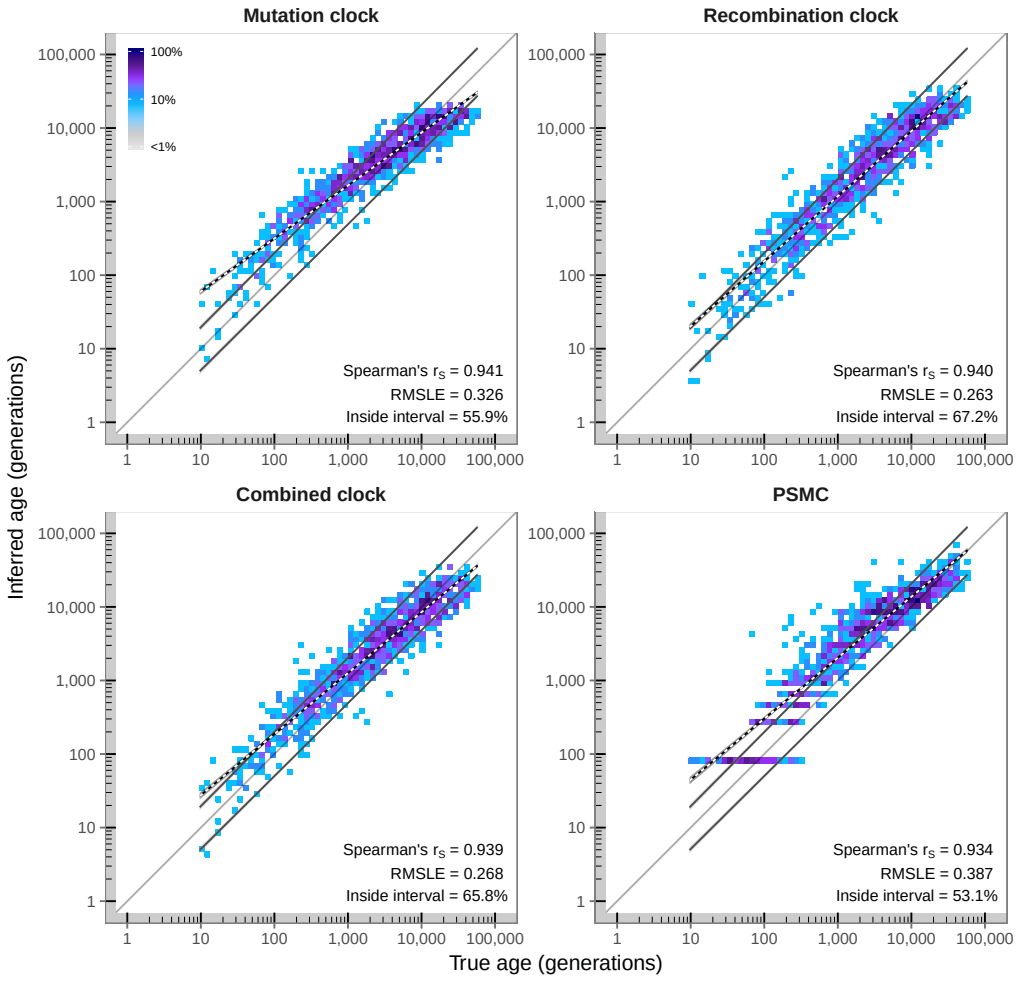


Figure 1.15: Inferred allele age using different methods. ...

Table 1.7: Accuracy of inferred age using the haplotype-based HMM. ...

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

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

— Elon Musk

Abbreviations

1000G	1000 Genomes Project
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
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

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.

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.

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

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.

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.

1. *I have told you more than I know* [...].
2. *What I have told you is subject to change without notice.*
3. *I hope I raised more questions than I have given answers.*
4. *In any case, as usual, a lot more work is necessary.*

– Fuller Albright