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)

5 Estimation of rare allele age

Contents

5.1 Introduction

The inference of the genealogical history of a sample is of interest to a myriad of applications in genetic research, both in population and medical genetics. The "age" of an allele, which simply refers to the time since the allele was created by a mutation

event, is of particular interest; for example, to observe demographic processes and events, or to better understand the effects of disease-related variants by their time of emergence in a population.

In this chapter, I propose a novel method to estimate the age of an allele, which is based on a collection of statistical models that derive from coalescent theory. The approach derives from concepts of the *composite likelihood*, which recently have gained in popularity for various applications in genetic research. In particular, composite likelihood methods can be useful in situations 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 expected to be proportionally constant.

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

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

I begin this chapter by introducing the concept of the method, which is followed by a detailed description of the statistical framework. The method is evaluated in extensive simulation studies, which also consider data error as a source of estimation bias. Although

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

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.

5.2 Approach

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

According to coalescent theory, any two haplotypes that share the allele are expected to have coalesced more recently than the time of the focal mutation event. Conversely, the coalescent event between one haplotype carrying the allele and one haplotype not carrying the allele is expected to date back to a point in time before the mutation event occurred. This insight is of particular interest as it suggests that the actual time of the mutation event lies somewhere in between two such points in time.

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

There are two main sources of information available from sample data which relate to the T_{MRCA}. First, mutation events occur independently in each lineage and mutations accumulate along the sequence as the haplotype is passed on over generations. Second, recombination events break down the length of the haplotype in each generation independently in each lineage. Thus, the T_{MRCA} between a given pair of chromosomes can be estimated from the number of mutations which segregate in two haplotypes, as well as the genetic length of the haplotype region that is shared between two chromosomes in the sample. In the following (Section 5.2.1), I derive the formulations for three T_{MRCA} estimators. These are referred to as follows.

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

I then explain the age estimation method in detail in Section 5.2.2 (page 170).

5.2.1 Coalescent time estimators

The posterior probability is proportional to the prior probability of the time to coalescence multiplied by the likelihood of the time. The derivation of the prior distribution on the coalescent time follows from the results given in Section 1.3.2 (page 14), 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

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

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

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

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 $\tau \sim \mathrm{Exp}(1)$ and written as

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

5.2.1.1 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

$$f_S(s) \;=\; P(S = s \mid \theta, h, \tau) \;=\; \frac{(\theta h \tau)^s}{s!}\, \mathrm{e}^{-\theta h \tau}. \tag{5.1}$$

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

The likelihood function for the time parameter τ is proportional to Equation (5.1), but requires only those terms that involve τ and where constant terms can be dropped, such that

$$\mathcal{L}(\tau \mid \theta, h, s) \;\propto\; \tau^s\, \mathrm{e}^{-\theta h \tau}\,. \tag{5.2}$$

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

$$\begin{aligned} p(\tau \mid \theta, h, s) \;&\propto\; \mathcal{L}(\tau \mid \theta, h, s) \;\times\; \pi(\tau) \\ &\propto\; \tau^s\, \mathrm{e}^{-\tau(\theta h + 1)} \end{aligned} \tag{5.3}$$

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 (5.3) 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

$$p(\tau \mid \theta, h, s) \;=\; g(\tau \mid s + 1, \theta h + 1). \tag{5.4}$$

5.2.1.2 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 $D \sim \text{Exp}(2\psi\tau)$, 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

$$P(D = d \mid \psi, \tau) \;=\; 2\psi\tau \; \mathrm{e}^{-2\psi\tau d}. \tag{5.5}$$

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

$$P(D > d \mid \psi, \tau) \;=\; \mathrm{e}^{-2\psi\tau d}. \tag{5.6}$$

Equations (5.5) and (5.6) above can be simplified to

$$f_D(d) \;=\; (2\psi\tau)^b \; \mathrm{e}^{-2\psi\tau d} \tag{5.7}$$

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

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

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

Considering Equation (5.7), the likelihood function for τ can now be written as

$$\mathcal{L}(\tau \mid \psi, d, b) \propto \tau^{b} \, \mathrm{e}^{-2\psi d\tau} \tag{5.8}$$

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

$$\mathcal{L}(\tau \mid \psi, h, b_L, b_R) \propto \tau^{b_L + b_R} \, \mathrm{e}^{-2\psi h\tau} \tag{5.9}$$

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

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

The posterior probability is obtained as

$$\begin{aligned} p(\tau \mid \psi, h, b_L, b_R) &\propto \mathcal{L}(\tau \mid \psi, h, b_L, b_R) \times \pi(\tau) \\ &\propto \tau^{b_L + b_R} \, \mathrm{e}^{-\tau(2\psi h + 1)}. \end{aligned} \tag{5.10}$$

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

$$p(\tau \mid \psi, h, b_L, b_R) = g(\tau \mid b_L + b_R + 1, 2\psi h + 1). \tag{5.11}$$

5.2.1.3 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 (5.1) and (5.7) on page 167 and page 168 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,

$$\mathcal{L}(\tau \mid h_p, h_g, s, b_L, b_R) \propto \tau^{s+b_L+b_R} \, \mathrm{e}^{-\tau(h_p+h_g)} \tag{5.12}$$

from which the posterior probability is obtained as

$$\begin{aligned} p(\tau \mid h_p, h_g, s, b_L, b_R) &\propto \mathcal{L}(\tau \mid h_p, h_g, s, b_L, b_R) \times \pi(\tau) \\ &\propto \tau^{s+b_L+b_R} \, \mathrm{e}^{-\tau(h_p+h_g+1)}. \end{aligned} \tag{5.13}$$

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

$$p(\tau \mid h_p, h_g, s, b_L, b_R) = g(\tau \mid s + b_L + b_R + 1, h_p + h_g + 1). \tag{5.14}$$

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

5.2.2 Inference of allele age from coalescent time posteriors

Consider a focal allele of interest that is shared by some of the haplotypes in a sample. The time at which this allele was created by a mutation event is bound by the times of the two coalescent events that delimit the length of the branch on which the mutation occurred in the underlying coalescent tree; see the example provided in Figure 5.1 (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 from the coalescent that all lineages in X_c coalesce before any of them can coalesce with a lineage in X_d. Any coalescent event between two lineages in X_c must have occurred *earlier* than the focal mutation event (back in time). On the other hand, any coalescent event between one lineage in X_c and one lineage in X_d must have occurred *later* than the focal mutation event (back in time). Pairs of haplotypes in X_c are referred to as *concordant* pairs, whereas pairs formed by strictly taking one haplotype from X_c and another from X_d are *discordant* pairs. The sets Ω_c and Ω_d are defined to contain all concordant and discordant pairs, respectively.

In the following, I describe how the posterior density of the $\mathrm{T}_{\mathrm{MRCA}}$ is obtained for concordant and discordant pairs to eventually arrive at an estimate of allele age. To distinguish the population-scaled time τ, as defined for the $\mathrm{T}_{\mathrm{MRCA}}$, from the time of

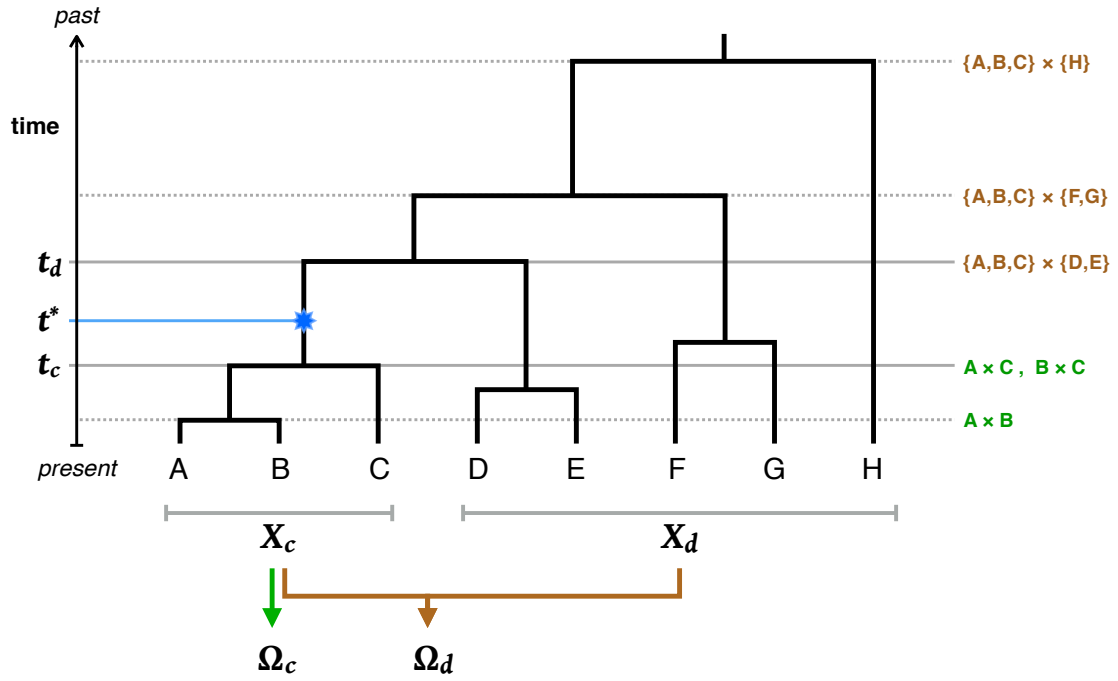

Figure 5.1: Allele age in relation to concordant and discordant pairs. The genealogy of a sample of eight haplotypes is shown of which A, B, and C share a focal allele that derived from a mutation event as indicated in the tree (*star*). These chromosomes constitute the set of *carriers*, denoted by 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.

the mutation event, let the latter be denoted by the likewise population-scaled time t. Informally, t 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 5.1.

5.2.2.1 Cumulative coalescent function (CCF)

CORRECTION Section partially rewritten with revised notation

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

given pair of haplotypes, denoted by i, j. In simple terms, the CCF is expressed as

$$\Phi_{ij}(t) \;=\; \begin{cases} P(\tau \leq t) & \text{if } \{i,j\} \subseteq \Omega_c \quad (\textit{i.e.}\text{ concordant pairs}) \\ P(\tau > t) = 1 - P(\tau \leq t) & \text{if } \{i,j\} \subseteq \Omega_d \quad (\textit{i.e.}\text{ discordant pairs}). \end{cases} \tag{5.15}$$

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

$$G(t) \;=\; P(\tau \leq t \mid \alpha, \beta) \;=\; \int_0^t g(u \mid \alpha, \beta)\, du \tag{5.16}$$

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)

$$F(t) \;=\; P(\tau \leq t \mid \alpha, \beta) \;=\; 1 - \mathrm{e}^{-\beta t} \sum_{i=0}^{\alpha-1} \frac{(\beta t)^i}{i!}. \tag{5.17}$$

Further, to obtain point estimates from the posterior of the $\mathrm{T_{MRCA}}$, 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.

5.2.2.2 Allele age estimation from the composite posterior distribution

CORRECTION Section partially rewritten with revised notation

A at given focal site, the CCF is obtained for concordant and discordant pairs. Because the $\mathrm{T_{MRCA}}$ of concordant pairs would extend to a point below the focal mutation event and the $\mathrm{T_{MRCA}}$ 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.

$$\Lambda_k(t) \;\propto\; \prod_{i,j \in A_k} \Phi_{ij}(t \mid \alpha, \beta) \tag{5.18}$$

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 (5.18).

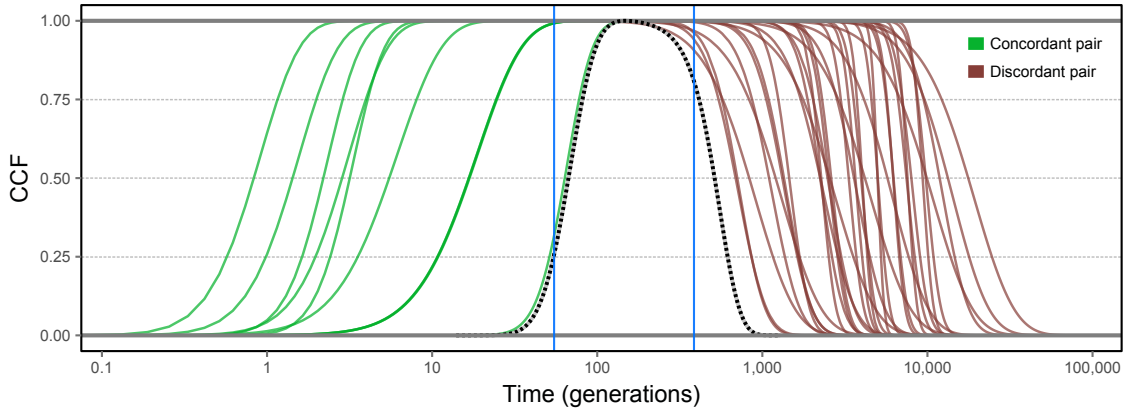

Figure 5.2: Example of concordant and discordant posterior distributions and the resulting composite posterior. A target variant was randomly selected from simulated data. The CCF was obtained for the set of possible concordant pairs and a subset of concordant pairs, which were randomly selected. The thicker *dotted* line shows the distribution of the maximised composite posterior. The *blue* lines mark the times of coalescent events below and above the focal mutation event; *i.e.* t_c (*left*) and t_d (*right*), determined from simulation records. Their distance corresponds to the length of the branch on which the focal mutation event occurred.

The composite posterior distribution can now be obtained over $t \in (0, \infty)$. However, in practise, it is unlikely that the relationship of i, j can be traced back further than a small multiple of N_e (*e.g.* ~ 10). An example is given in Figure 5.2 (this page), showing the CCF 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 for the age of an allele, denoted by $\hat{t}$.

5.2.2.3 Note on composite likelihood methods

ADDITION Section included

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

$$\mathcal{CL}(\vartheta \mid y) = \prod_{z \in Z} \mathcal{L}_z(\vartheta \mid y)^{w_z} \tag{5.19}$$

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

$$p_{\mathcal{CL}}(\vartheta \mid y) \propto \pi(\vartheta) \times \mathcal{CL}(\vartheta \mid y) \tag{5.20}$$

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

The "composite posterior" given in Equation (5.18) on page 172 follows a similar approach in context of the above, but is defined proportional to the product of posteriors. While $\Lambda_k(t)$ itself cannot be regarded as a composite likelihood, it can be argued that the proposed method is an (*ad hoc*) approach equivalent to using the composite likelihood in a Bayesian setting.

5.2.2.4 Note on the computational burden

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

$$\max[n_c] = \binom{k}{2} = \frac{k(k-1)}{2} \tag{5.21}$$

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

$$\max[n_d] = k(2N-k) \tag{5.22}$$

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, for $N = 1{,}000$, each

f_2 variant has $n_c = 2$ and $n_d = 3{,}996$, whereas each f_{20} variant already has $n_c = 190$ and $n_d = 19{,}600$. Therefore, in practise, the computational burden is reduced by employing a sampling regime where, for example, pairs in Ω_c and Ω_d are picked at random.

REMOVED Section “Anticipated limitations”

5.3 Evaluation

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

5.3.1 Data generation

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

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

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

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

5.3.2 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

$$t_m = \sqrt{t_c \, t_d} \,. \quad \text{CORRECTED} \tag{5.23}$$

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.*representation on log-scale).

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

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

$$\hat{t}_{rel} = \frac{\log\left[\frac{\hat{t}}{t_c}\right]}{\log\left[\frac{t_d}{t_c}\right]} \tag{5.24}$$

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

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

5.4 Validation of the method

The allele age estimation method relies on an (ideally) correct inference of the haplotype region shared by descent between two chromosomes relative to a target site. Several approaches for targeted, pairwise inference of the shared haplotype have been developed in the previous chapters, which are applied further below. To first establish *proof of concept* of the age estimation method, its performance was evaluated given complete

knowledge of the underlying shared haplotype structure. That is, the "true" shared haplotype segments were determined from simulation records and analysed using haplotype data in dataset $\mathcal{D}_A$.

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

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

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

5.4.1 Results

An overview is illustrated in Figure 5.3 (next page), which shows the density of true and estimated age under each clock model; results are shown for $n_d = 10$, $n_d = 100$, and $n_d = 1{,}000$, to better distinguish differences visually. Note that, here, true age was set to t_m (the geometric mean of t_c and t_d).

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

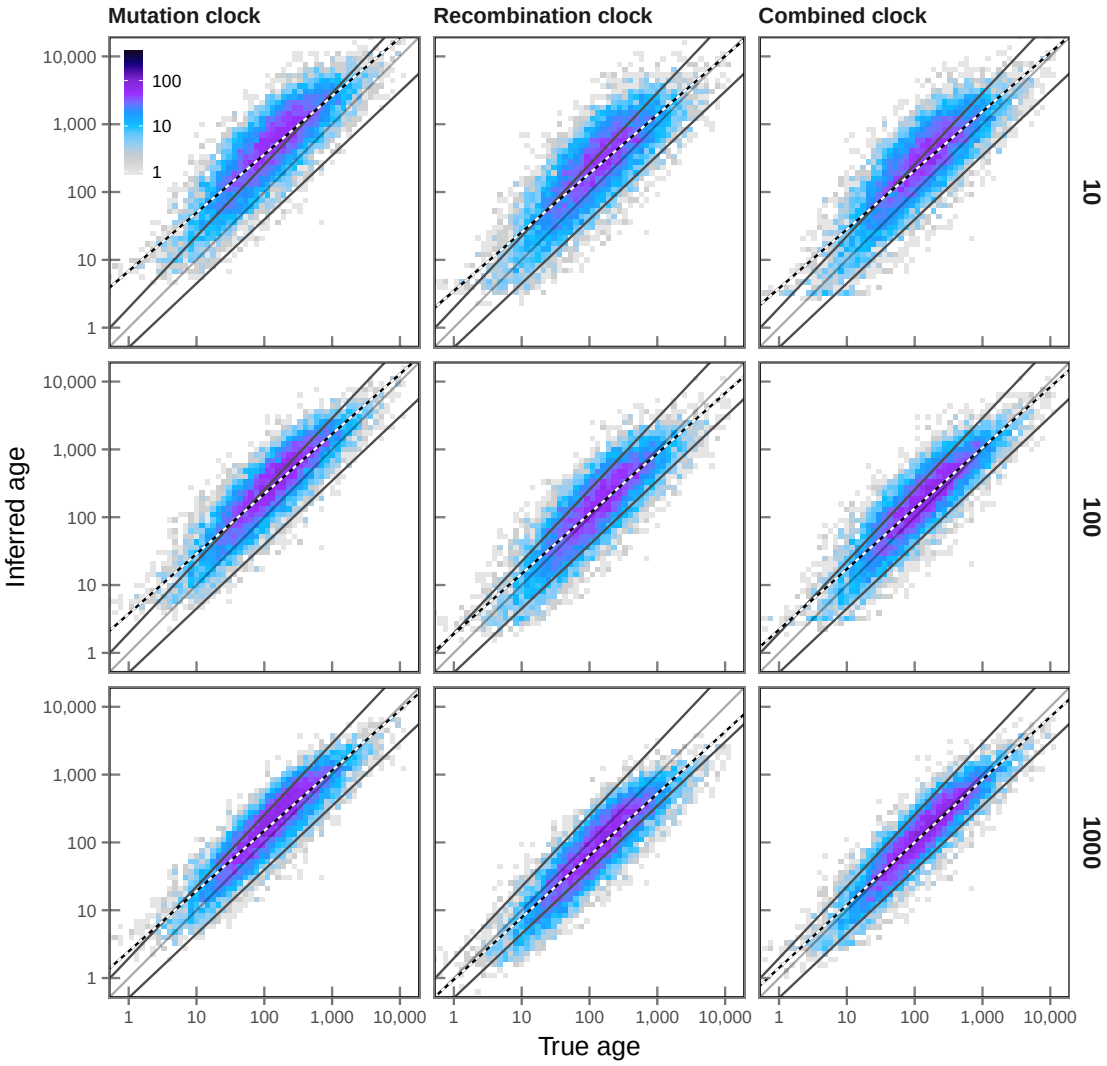

Figure 5.3: True and inferred age under varying numbers of discordant pairs. A set of 10,000 target sites was randomly drawn in $f_{[2,20]}$ (shared allele frequency $\leq 1\%$) in a simulated sample of 2,000 haplotypes. Different numbers of sampled discordant pairs were analysed on the same set of target variants, which is shown for $n_d = 10$, $n_d = 100$, and $n_d = 1{,}000$ (indicated at the *right* of each row). True IBD was used to estimate allele age. IBD breakpoints were determined from simulation records and defined as the first variant sites observed in the data following the two recombination events on each side of a given focal position. Age was estimated under each of the three clock models; *i.e.* mutation clock, $\mathcal{T}_{\mathcal{M}}$, recombination clock, $\mathcal{T}_{\mathcal{R}}$, and combined clock, $\mathcal{T}_{\mathcal{MR}}$ (indicated at the *top* of each column). Each panel shows the density distribution of true and inferred age (numbers indicated by the colour-gradient). Note that the "true age" of a focal allele was set to t_m, which is the geometric mean of t_c and t_d, *i.e.* the true time of the coalescent event from which the focal allele derived (t_c) and the true time of the coalescent event immediately preceding that event (t_d) in the history of the sample, respectively; these are indicated by their regression trend lines *below* and *above* the dividing line at t_m, respectively. The *black-white* line indicates the line of best fit resulting from linear regression of age estimates, using the posterior mode of the composite likelihood distribution as the inferred age value. True and inferred age are both shown on log-scale.

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

Table 5.1: Estimation accuracy under varying numbers of discordant pairs. Different thresholds for the number of randomly formed discordant pairs, n_d, were analysed to evaluate the impact on the accuracy of allele age estimation. Note that all possible concordant pairs were included in each analysis; *i.e.* n_c was not reduced. True IBD segments were used to focus on the differences induced by varying n_d thresholds. Each analysis was conducted on the same set of 10,000 randomly selected rare variants at allele frequency ≤ 1%. Accuracy was measured using the rank correlation coefficient, r_S, and the magnitude of error, RMSLE, between the estimated age, $\hat{t}$ and the the times of coalescent events; *i.e.* the time until all haplotypes in X_c have coalesced, t_c, and the time of the immediately preceding coalescent event, t_d, which joined the lineages in X_c and X_d back in time, as well as the geometric mean of both, t_m.

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

A complete summary of results is given in Table 5.1 (this page). Throughout, rank correlation (r_S) was highest for n_d = 1,000; see Table 5.1. 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 5.1), suggesting that further improvements in accuracy may not be attained by increasing the threshold.

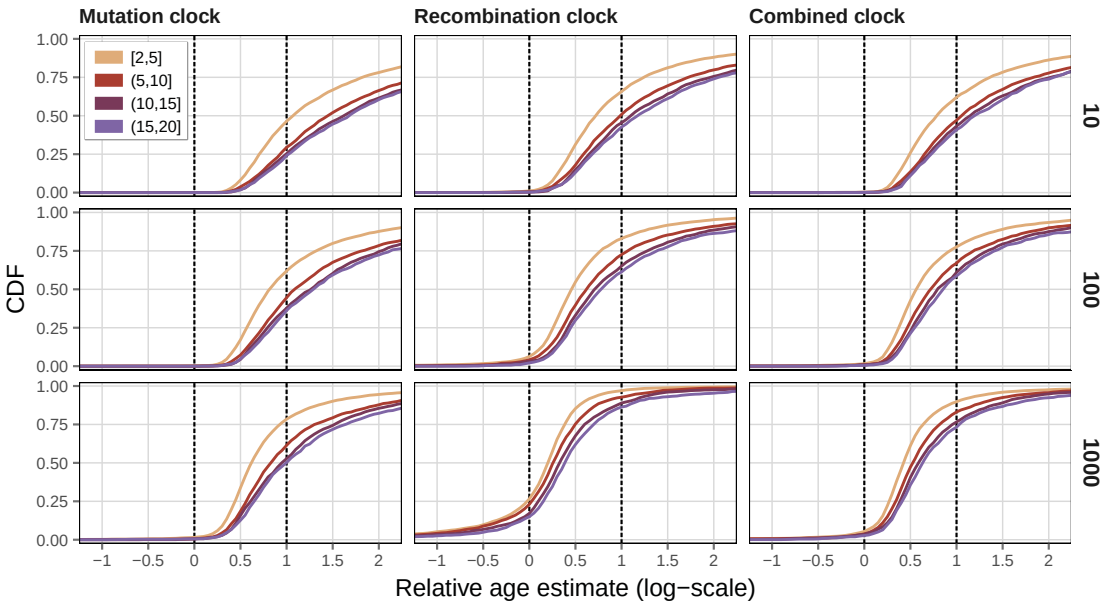

Figure 5.4: Relative age under varying numbers of discordant pairs. A randomly drawn set of 10,000 target sites at allele frequency ≤ 1%, *i.e.* $f_{[2,20]}$, was analysed under each of the three clock models (indicated at the *top* of each column) and with different numbers of sampled discordant pairs; 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 (5.24), 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 $f_{(15,20]}$.

A comparison of the inferred age distributions at distinct f_k ranges is presented in Figure 5.4 (page 181), 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.

5.4.2 Discussion

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

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

For instance, a simple approach would be to compute the Hamming distance between haplotypes in X_c and X_d within a short region around the position of a given target site, such that a subset of presumed nearest neighbours can be selected based on a distance ranking. In practice, however, there are two caveats to such an approach. First, it would

be computationally expensive to conduct an additional pairwise analysis for the (whole) sample at each target site, which may not outweigh the improvement gained through the reduction of n_d. Second, a dilemma arises in presence of data error, as the identification of nearest neighbours is likely to give preference to haplotypes in which the focal allele has been missed. Such *false negatives* distort the estimation of allele age as the CCF computed for false discordant pairs could bias the resulting composite posterior distribution.

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

5.5 Age estimation using inferred shared haplotype segments

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

In this section, two main analyses were conducted. First (Section 5.5.2), dataset $\mathcal{D}_A$ was analysed to provide a comparison to Section 5.4 (page 177), where true shared haplotype information was used to validate the age estimation method. Second (Section 5.5.3), an extensive analysis was performed on datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$ to assess the impact of data error on age estimation. Note that the impact of phasing error was also evaluated in the latter, but which affected only the FGT.

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

5.5.1 Modifications of IBD detection methods

Four-gamete test (FGT). The FGT is applied to the four haplotypes observed in two diploid individuals. A recombination event is inferred to have occurred between two variant sites if all four possible allelic configurations are observed. Let the focal site be denoted by b_i and another, distal site by b_j. In the four haplotypes, the alleles observed at (b_i, b_j) confirm a breakpoint if, for example, $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$ are observed, where 0 denotes the ancestral allelic state and 1 the derived state. Since breakpoints

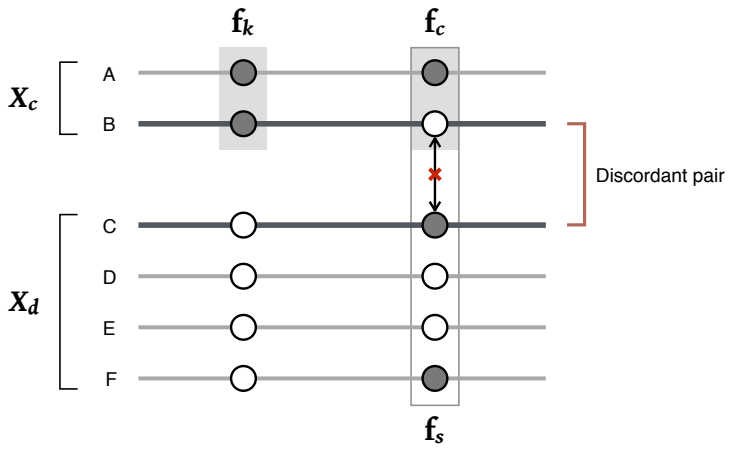

Figure 5.5: Breakpoint detection in discordant pairs. A discordant pair is formed by one haplotype from X_c (which share the focal allele) and one haplotype from X_d (which do not share the focal allele). The lines indicate the chromosomal sequence where the alleles at two sites are indicated; allelic 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.

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.

However, breakpoints in discordant pairs can be detected as based on the allele frequencies observed in the sample. Let f_k denote the number of allele copies at focal site b_i. At a distal site, b_j, let f_c denote the number of allele copies observed only within the subsample X_c (who carry the focal allele at the focal position). Also, let f_s be the number of allele copies at b_j in the whole sample. A recombination breakpoint is indicated at b_j if the two haplotypes carry different alleles and if $f_c < f_k$ and $f_c < f_s$; additionally $f_s > 1$ to exclude singletons and $(f_s - f_c) > (2N - f_k)$ to exclude sites that are monomorphic within X_d, where $2N$ refers to the number of haplotypes in the sample. The condition implies the existence of the four allelic configurations at any of the haplotypes in the sample but is not bound by haplotype occurrence in two diploid individuals. The FGT thereby still holds but is practically inverted. An example is illustrated in Figure 5.5 (this page).

Discordant genotype test (DGT). Recall that the DGT is a special case of the FGT which detects breakpoints at genotypic configurations that would also pass the FGT if haplotypes were available. Given the two heterozygous genotypes at the focal variant, a breakpoint is

found at a distal site if opposite homozygous genotypes are observed. Again, in discordant pairs, such a configuration cannot be observed. The observation of opposite homozygous genotypes nonetheless implies that the two individuals do not share a haplotype at this site and is therefore also applied for breakpoint detection in discordant pairs.

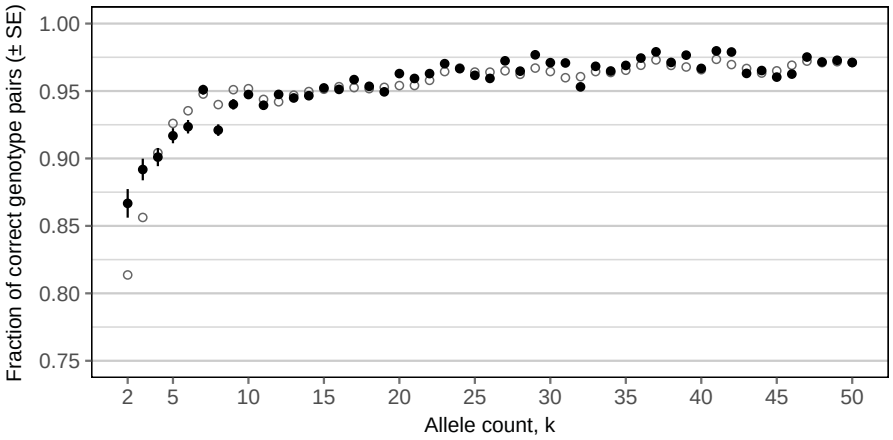

Figure 5.6: Initial state probability of discordant pairs in the Hidden Markov Model (HMM). The proportion of discordant pairs that were correctly identified by their genotypes was empirically determined from data before and after the inclusion of realistic genotype error rates. The mean per f_k was used as the initial state probability of the HMM-based approach for IBD detection around target sites. For comparison, the initial state probability of concordant pairs is shown (*hollow* circles).

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

The initial state probability corresponds to the probability of correctly observing a concordant pair through allele sharing, *i.e.* the true positive rate of observing heterozygous genotypes at a given target site where both individuals share the focal allele, which was determined per focal allele frequency (f_k). The empirical model was extended such that there was an additional initial state probability available and applied to discordant pairs. By comparing the data before and after error ($\mathcal{D}_B$ and $\mathcal{D}_B{}^*$), the initial state probability was determined as follows. For each f_k category, I randomly selected 1,000 target sites in the dataset "before error" for which I randomly selected 1,000 discordant pairs per target site. I then compared these genotypes to the corresponding genotypes in the dataset "after error" to determine the true positive rate. The mean per f_k was taken as the empirical

initial state probability. The resulting distribution is shown in Figure 5.6 (page 185); the initial state probabilities used for discordant pairs are given for comparison. However, the initial state probability for the discordant case is similar to the concordant one. A possible explanation is that this is particularly driven by the heterozygous status being false.

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

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

5.5.2 Comparison of IBD detection methods

Dataset $\mathcal{D}_A$ was used to compare the different IBD detection methods. Age was estimated for each of the three clock models, using a threshold of $n_d = 1{,}000$. The results presented in this section were obtained on the previously selected 10,000 rare allele target sites; see Section 5.4 (page 177). Again, the parameters of the age estimation method were specified according to simulation parameters ($N_e = 10{,}000$; $\mu = 1 \times 10^{-8}$ per site per generation; $\rho = 1 \times 10^{-8}$ per site per generation).

The density of true and inferred allele age is given in Figure 5.7 (next page). In all three methods, a tendency to overestimate allele age was seen, in particular under the mutation clock, $\mathcal{T}_{\mathcal{M}}$. This overestimation was elevated when the DGT was used, and less prominent for the FGT or HMM. The latter methods showed similar distributions in $\mathcal{T}_{\mathcal{M}}$ and under the combined clock model, $\mathcal{T}_{\mathcal{MR}}$, in which age appeared to be less overestimated. Under the recombination clock, $\mathcal{T}_{\mathcal{R}}$, alleles were underestimated in each method, but more severely in both the DGT and HMM.

Specifically, the method with the highest proportion of correctly estimated alleles was the FGT in all three clock models, where accuracy was highest under $\mathcal{T}_{\mathcal{R}}$ (72.6 %), followed by $\mathcal{T}_{\mathcal{MR}}$ (55.4 %) and $\mathcal{T}_{\mathcal{M}}$ (34.5 %). The HMM achieved similar proportions,

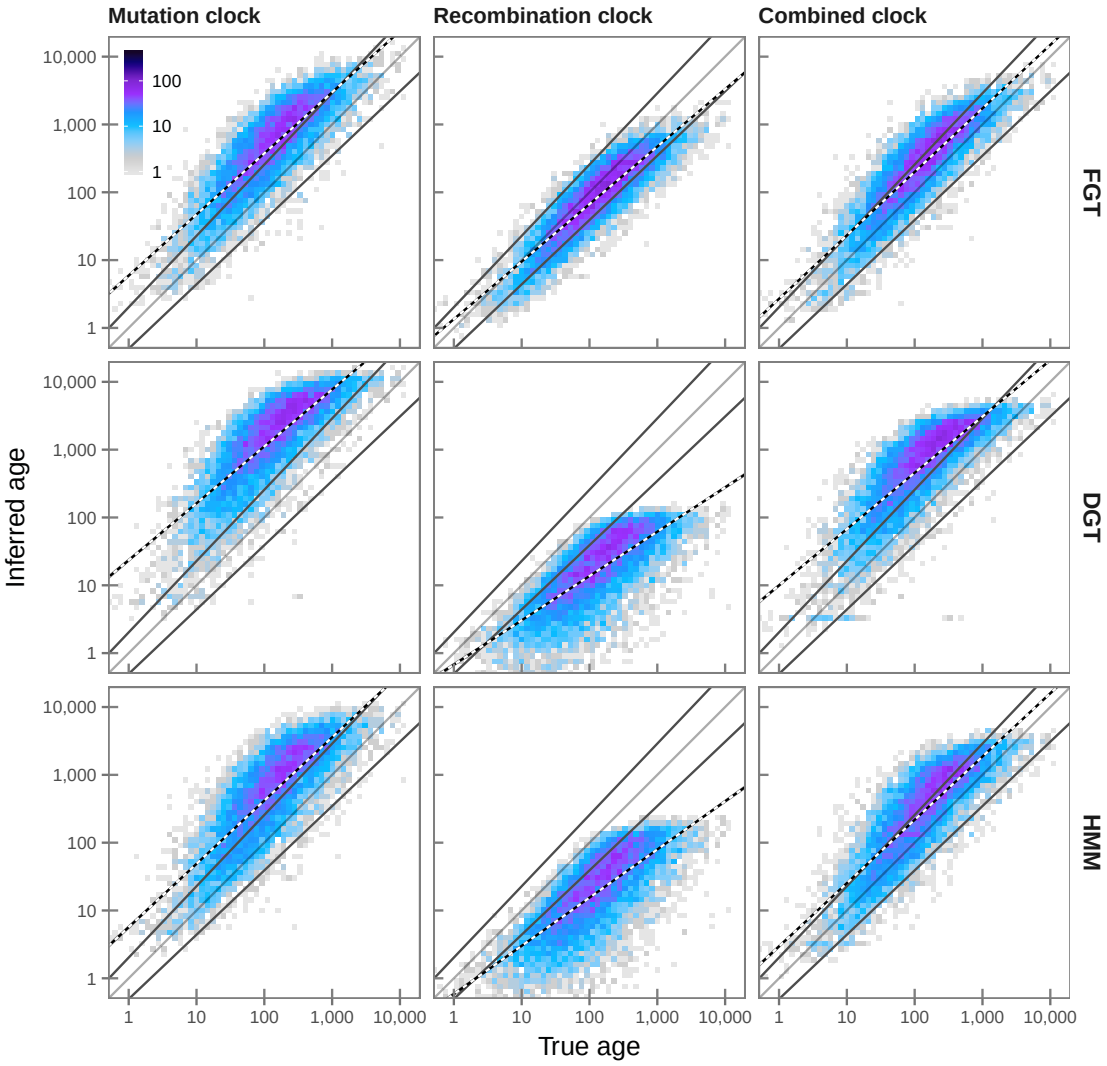

Figure 5.7: Distribution of true and inferred age using different IBD detection methods. The three IBD detection methods FGT, DGT, and HMM were compared under each clock model, on the same set of target sites that were drawn from $f_{[2,20]}$ variants (allele frequency $\leq 1\%$) in $\mathcal{D}_A$. Each panel shows the density of true age (t_m) and inferred age. Lines *below* and *above* the dividing line are regression trend lines of the corresponding true coalescent times around each mutation event, t_c and t_d, respectively. The regression of inferred age ($\hat{t}$) is given by the *black-white* line.

but which was low in $\mathcal{T}_{\mathcal{R}}$ (10.950 %) compared to $\mathcal{T}_{\mathcal{MR}}$ (51.876 %) and $\mathcal{T}_{\mathcal{M}}$ (32.415 %). Throughout, the lowest proportions were found for the DGT (14.554 %, 8.226 %, and 29.659 % for $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively).

Summary metrics for each analysis are given in Table 5.2 (next page). Throughout, the FGT showed a higher accuracy compared to the other IBD detection methods under each clock model. Relative age estimates are shown for distinct f_k ranges in Figure 5.8 (page 189), where the relative age for corresponding results obtained by using true IBD

Table 5.2: 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 5.1 (page 180) 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

information is given for comparison per clock model; see Figure 5.4 (page 181). Analyses under $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$ showed a substantial difference between alleles at lower and higher frequencies; *e.g.* overall accuracy of $f_{[2,5]}$ variants was increased compared to f_k variants at higher frequencies in each method. This difference was reduced under $\mathcal{T}_{\mathcal{R}}$, but the DGT showed an accuracy decrease for $f_{[2,5]}$ variants.

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

I found that the FGT was the best performing method for the targeted detection of IBD segments, as the accuracy of estimated age was similar to the expectations defined by true IBD information in Section 5.4. However, the estimation was more accurate for target sites at lower allele frequencies. The DGT was least accurate in terms of estimated allele age in this comparison.

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

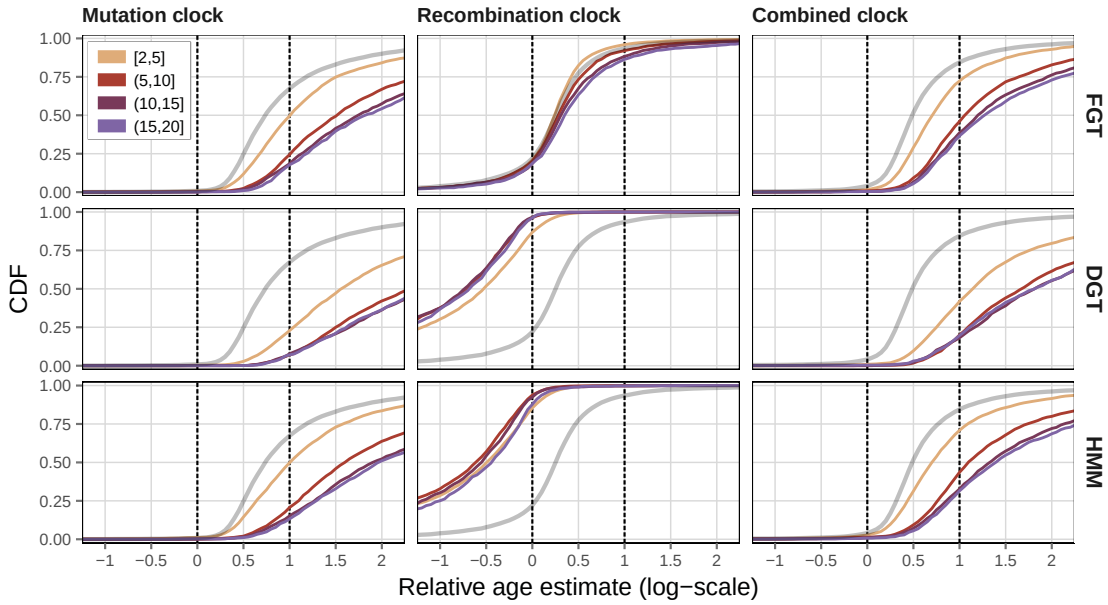

Figure 5.8: Relative age using different IBD detection methods. The three IBD detection methods implemented in `rvage` were compared, *i.e.* FGT, DGT, and HMM (indicated at the *right* of each row), under each clock model (indicated at the *top* of each column). The relative age, $\hat{t}_{rel}$, was calculated as given in Equation (5.24), such that t_c and t_d sit at 0 and 1 (*dashed* lines). The CDF of relative age estimates is shown for different frequency ranges; namely $f_{[2,5]}$, $f_{(5,10]}$, $f_{(10,15]}$, and $f_{(15,20]}$. The *grey* line provides a comparison to age estimated using true IBD information as shown in Figure 5.4 (page 181), but for $f_{[2,20]}$.

5.5.3 Impact of genotype error on allele age estimation

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

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

5.5.3.1 Age estimation using true shared haplotype information

First, estimation based on the true IBD structure of the sample is compared before and after error. Results are shown in Figure 5.9a (next page). The most striking discovery is the extent of overestimation after error under the mutation clock model, $\mathcal{T}_{\mathcal{M}}$, which was

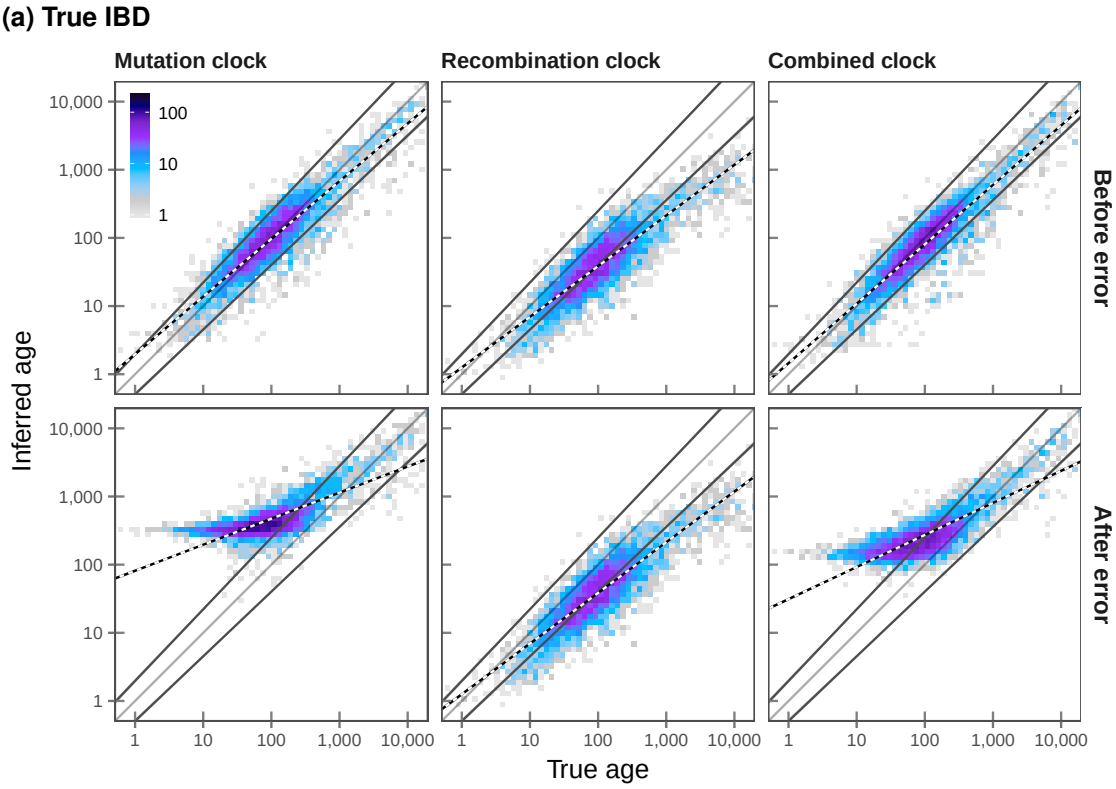

Figure 5.9: Density of allele age before and after error in simulated data. The effects on the estimation process *before* and *after* error are compared. Note that the "true age" was set to t_m, which is the geometric mean of t_c and t_d. Lines *below* and *above* the dividing line correspond to the regression lines over t_c and t_d; *i.e.* of the times of coalescent events delimiting the branch on which a focal mutation occurred. The *black-white* line gives the regression for the inferred age ($\hat{t}$). This panel (**a**) compares the distributions of true and inferred ages, which were estimated on basis of the true IBD structure of the sample as determined from simulation records. The other panels show estimation results based on the different IBD detection methods; FGT on both true and phased haplotypes (**b**, **c**; page 192), DGT (**d**; page 193), and the genotype-based HMM (**e**; page 195). Each analysis was conducted on the same set of 5,000 randomly selected target variants at $f_{[2,25]}$.

similarly high in the combined clock, $\mathcal{T}_{\mathcal{MR}}$. It is suggested that age was overestimated due to misclassified alleles which may have substantially increased the number of observed mutational differences observed within the shared haplotype interval.

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

The estimation under the recombination clock model, $\mathcal{T}_{\mathcal{R}}$, was not affected by genotype error, due to using true IBD information to derive segment lengths. Note that analyses were performed on the same sets of concordant and discordant pairs, which is why the

results in $\mathcal{T}_{\mathcal{R}}$ are identical before and after error. Allele age showed a tendency to be underestimated in $\mathcal{T}_{\mathcal{R}}$. Overall, 42.891 % of alleles were correctly inferred, and rank correlation was relatively high (r_S: 0.818, 0.843, and 0.666 at t_c, t_m, and t_d, respectively).

5.5.3.2 Age estimation based on inferred shared haplotypes

The different IBD detection methods are compared below. Results based on the FGT are shown in Figure 5.9b and 5.9c (next page), which show age estimates based on IBD detected in true (simulated) and phased haplotype data, respectively, both before and after error.

Before error, 53.021 %, 50.847 %, and 60.040 % of alleles were correctly inferred from true haplotype data in $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively. When phased data were used, this changed only slightly; 50.828 %, 51.366 %, and 59.182 % in $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively. Notably, the proportion of correctly inferred alleles increased in $\mathcal{T}_{\mathcal{R}}$ due to phasing error. It is suggested that the tendency for underestimation that was generally seen in $\mathcal{T}_{\mathcal{R}}$ may have been mitigated by further reduction of IBD segment lengths resulting from phasing error. The small difference between true and phased data was further reflected in r_S, which changed from 0.680 to 0.660 in $\mathcal{T}_{\mathcal{M}}$, 0.780 to 0.764 in $\mathcal{T}_{\mathcal{R}}$, and 0.742 to 0.731 in $\mathcal{T}_{\mathcal{MR}}$, with regards to t_d.

After error, the overall proportion of correct alleles was reduced, but again the differences between true and phased data were small. On true haplotypes, the proportion of correct alleles was 44.267 %, 45.025 %, and 42.034 % in $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$, respectively, whereas 43.549 %, 46.002 %, and 41.635 % of alleles were correct using phased haplotypes. Likewise, r_S and RMSLE scores did not suggest notable differences between estimation results from true and phased haplotypes; see Table 5.3 (page 194).

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

Estimation results based on the DGT are shown in Figure 5.9d (page 193). Before error, the proportions of correctly inferred alleles were the lowest in the present comparison in each clock model. For $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$, the estimation resulted in 26.341 % and 36.949 % of correct alleles, respectively, whereas only 2.413 % were correct for $\mathcal{T}_{\mathcal{R}}$. The tendency to overestimate allele age was increased after error. This was also seen for $\mathcal{T}_{\mathcal{R}}$, where the

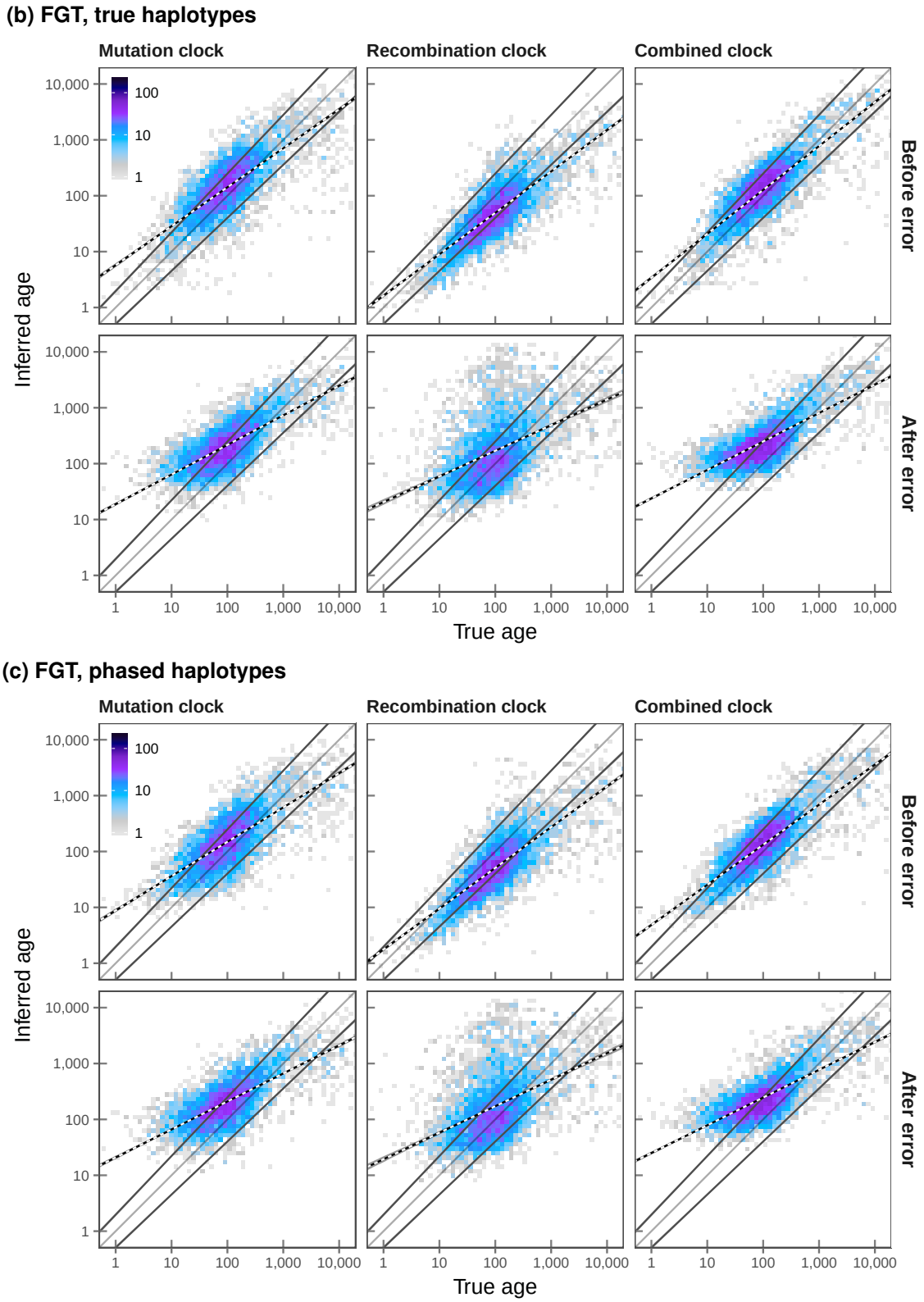

Figure 5.9: Continued.

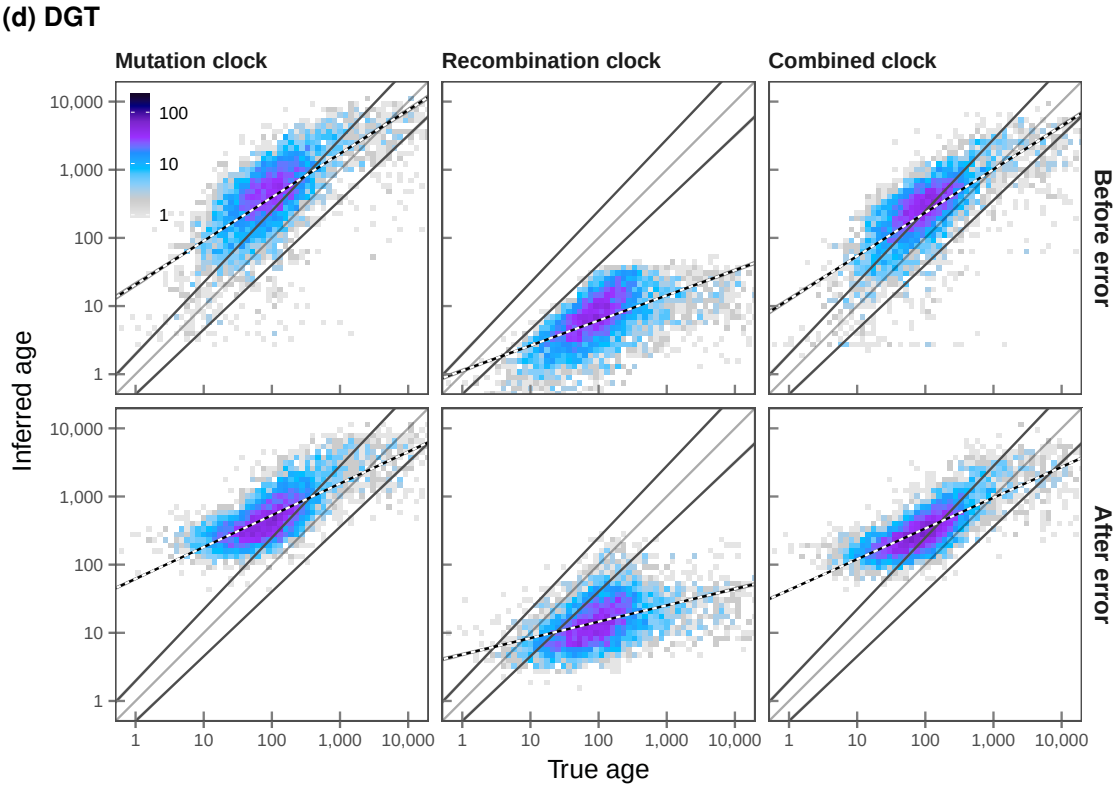

Figure 5.9: Continued.

proportion of correctly inferred alleles increased to 15.693 %, but at a loss of accuracy. Rank correlation, r_S, was 0.746, 0.628, and 0.406 at t_c, t_m, and t_d before error, and 0.588, 0.504, and 0.328 after error; see Table 5.3 (next page).

The accuracy of age estimation using the genotype-based HMM was relatively high before error; that is, more accurate in comparison to the FGT in $\mathcal{T}_{\mathcal{M}}$, similar in accuracy to the DGT in $\mathcal{T}_{\mathcal{R}}$, and similar to the FGT in $\mathcal{T}_{\mathcal{MR}}$. The density of inferred allele age based on the HMM is given in Figure 5.9e (page 195).

Before error, the proportion of correct alleles was 47.537 % in $\mathcal{T}_{\mathcal{M}}$, 3.629 % in $\mathcal{T}_{\mathcal{R}}$, and 57.827 % in $\mathcal{T}_{\mathcal{MR}}$. Allele age was generally underestimated in $\mathcal{T}_{\mathcal{R}}$ (96.351 %). This was increased after error, resulting in an underestimated proportion of 98.305 % in $\mathcal{T}_{\mathcal{R}}$, as the proportion of correct alleles was overall reduced; 16.650 % and 27.657 % in $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$, respectively. Also, RMSLE scores were lowest for the HMM under each clock model after error; see Table 5.3 (next page). Rank correlation before and after error, for r_S at t_c, decreased from 0.702 to 0.535 in $\mathcal{T}_{\mathcal{M}}$, and from 0.733 to 0.569 in $\mathcal{T}_{\mathcal{MR}}$. However, importantly, the HMM-based estimation 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.

Table 5.3: Effect of genotype error on age estimation accuracy. Allele age was estimated based on IBD inferred using the FGT, DGT, and HMM on the same set of 5,000 rare allele target sites randomly selected at allele frequency ≤ 0.5% ($f_{[2,25]}$) in simulated data before and after error (datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$). The number of discordant pairs was limited to n_d = 2,500 in each analysis. True IBD refers to age estimation conducted using knowledge of the actual shared haplotype structure of the sample, as determined from simulation records. CORRECTED

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

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

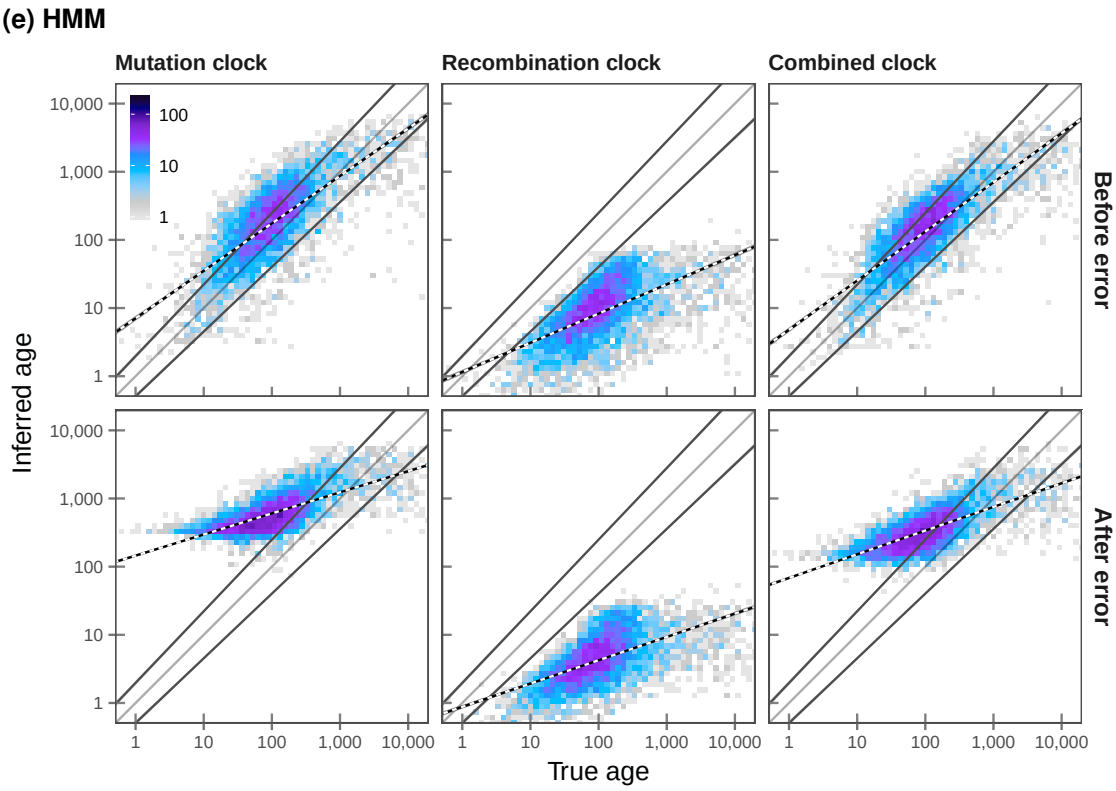

Figure 5.9: Continued.

5.5.4 Discussion

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

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

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

REMOVED Section “Generation of error correction models”

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

5.6 A haplotype-based HMM for targeted shared haplotype inference

5.6.1 Description of the model

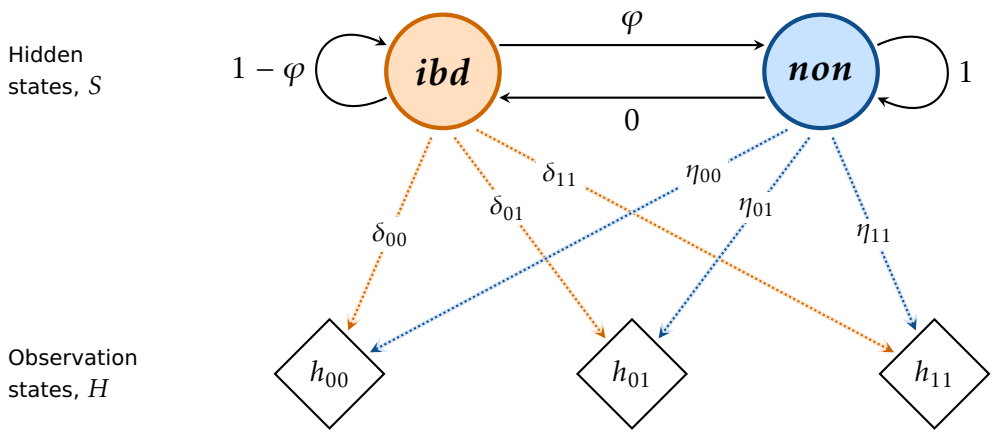

Figure 5.10: Illustration of the Hidden Markov Model for IBD inference. Two hidden states are assumed to generate the observations in a Markov process; *ibd* and *non*. Transitions from each state into any state are indicated by *solid* lines. The probability of transition from *ibd* to *non* is denoted 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*.

5.6.2 Impact of data error

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.

Table 5.4: 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

5.6.2.1 Shared haplotype inference

5.6.2.2 Allele age estimation

5.6.2.3 Discussion

5.6.3 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

5.6.3.1 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

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

where $T_{\max}$ is the maximum $\mathrm{T_{MRCA}}$ 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 $\mathrm{f_{\geq 2}}$ 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 $\mathrm{T_{MRCA}}$ estimation at older genealogical relationships in discordant pairs.

5.6.3.2 Results for the $\mathrm{T_{MRCA}}$

Simulation records were scanned to obtain the true $\mathrm{T_{MRCA}}$ 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 ??).

5.6.3.3 Results for allele age

5.6.3.4 Discussion

5.6.4 Allele age estimation in 1000 Genomes

5.6.4.1 Inferred allele age distribution by population

5.6.4.2 Comparison to PSMC

5.7 Discussion

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

Table 5.5: Accuracy of T_{MRCA} estimation for different methods. The T_{MRCA} estimation conducted using PSMC is compared to estimates obtained using the mutation clock ($\mathcal{T}_{\mathcal{M}}$), recombination clock ($\mathcal{T}_{\mathcal{R}}$), and combined clock ($\mathcal{T}_{\mathcal{MR}}$), where estimates were obtained on identical target sites and haplotype pairs; the median was taken as a point estimate from each posterior. Accuracy was measured using Spearman's rank correlation coefficient (r_S) and root mean squared $\log_{10}$ error (RMSLE) at discrete time intervals defined on the true T_{MRCA} (t) of a given pair at a target site, as determined from simulation records. The number of estimates compared per method at a given time interval is indicated (n).

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

Table 5.6: 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 $t_c \geq 1{,}000$

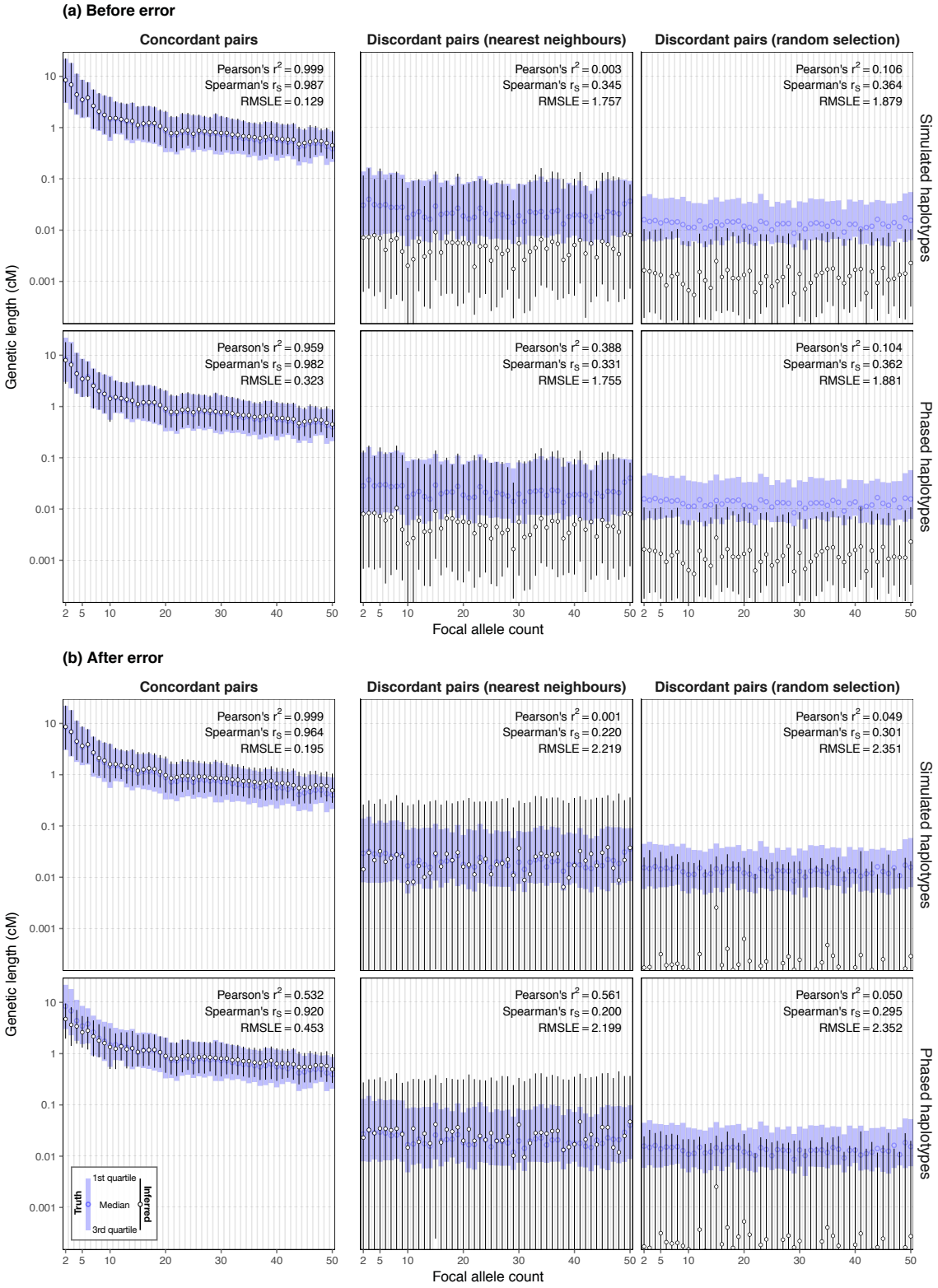

Figure 5.11: Genetic length of shared haplotype segments inferred using the haplotype-based HMM. The distribution of genetic length is shown by allele count of the focal variant in the simulated sample of $N = 5{,}000$ haplotypes, in direct comparison to the corresponding true length at the same set 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)**.

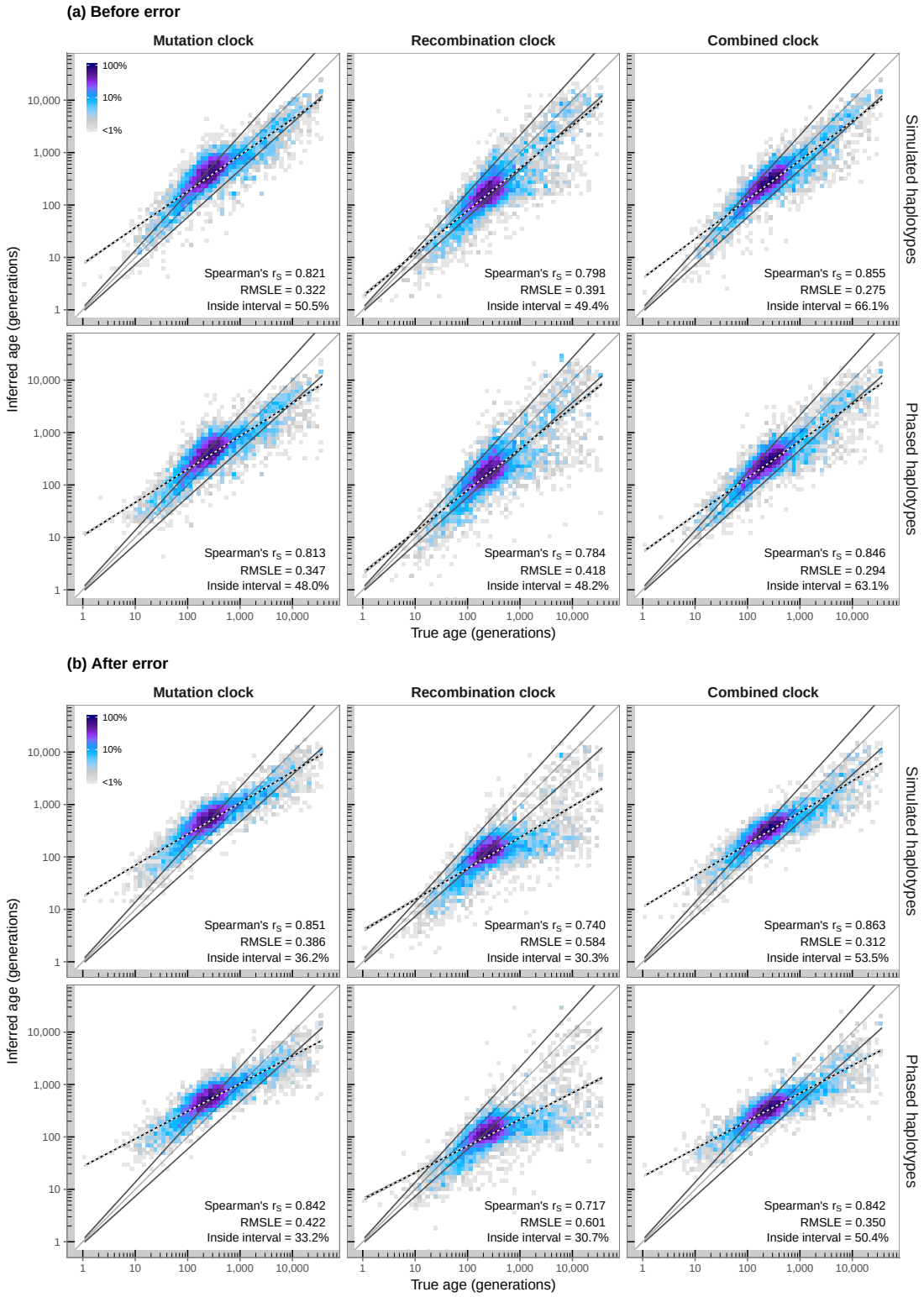

Figure 5.12: Allele age inferred using the haplotype-based HMM. ... Results were obtained on data before **(a)** and after error **(b)**.

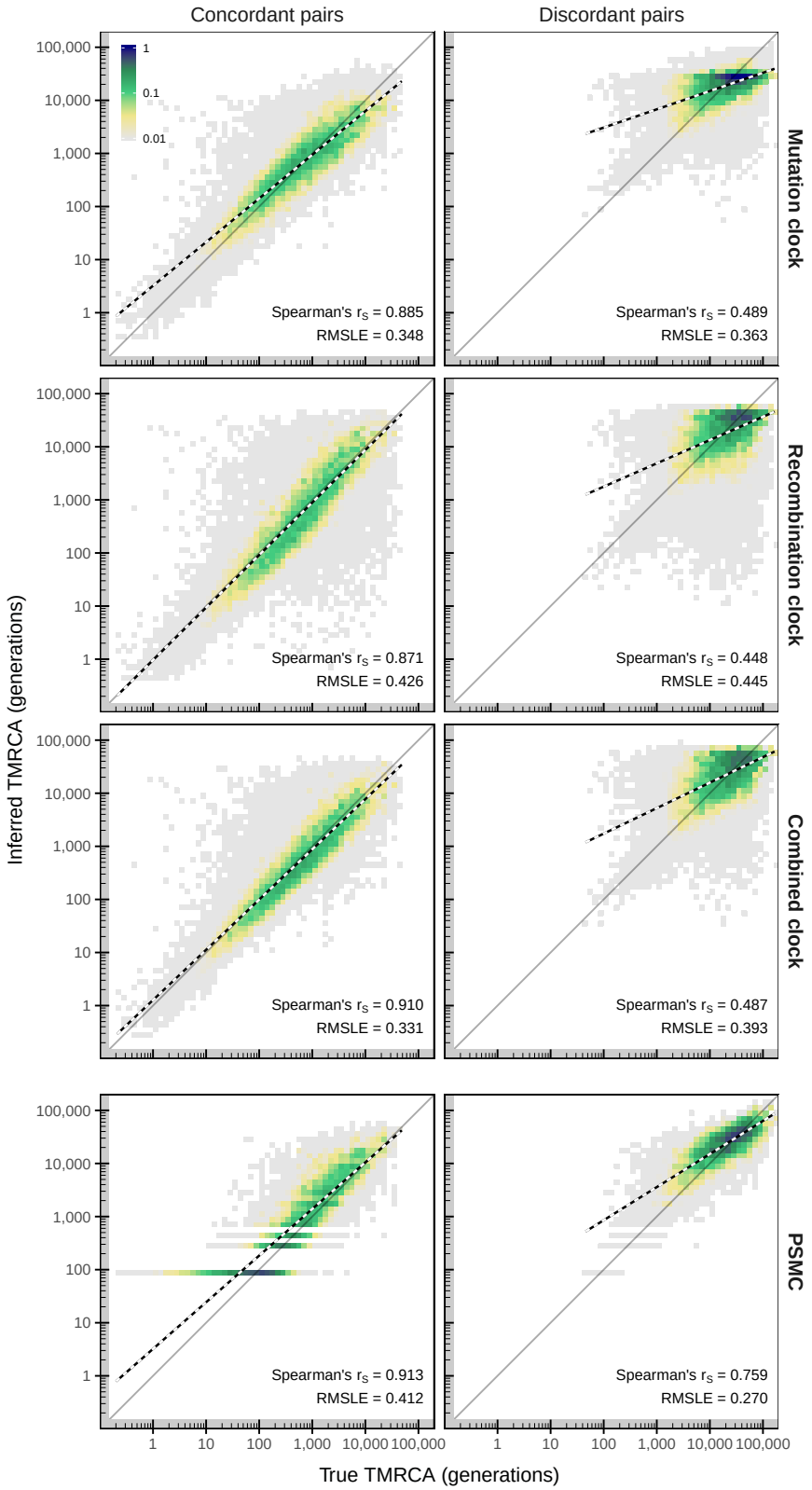

Figure 5.13: True and estimated T_{MRCA} using different methods. ...

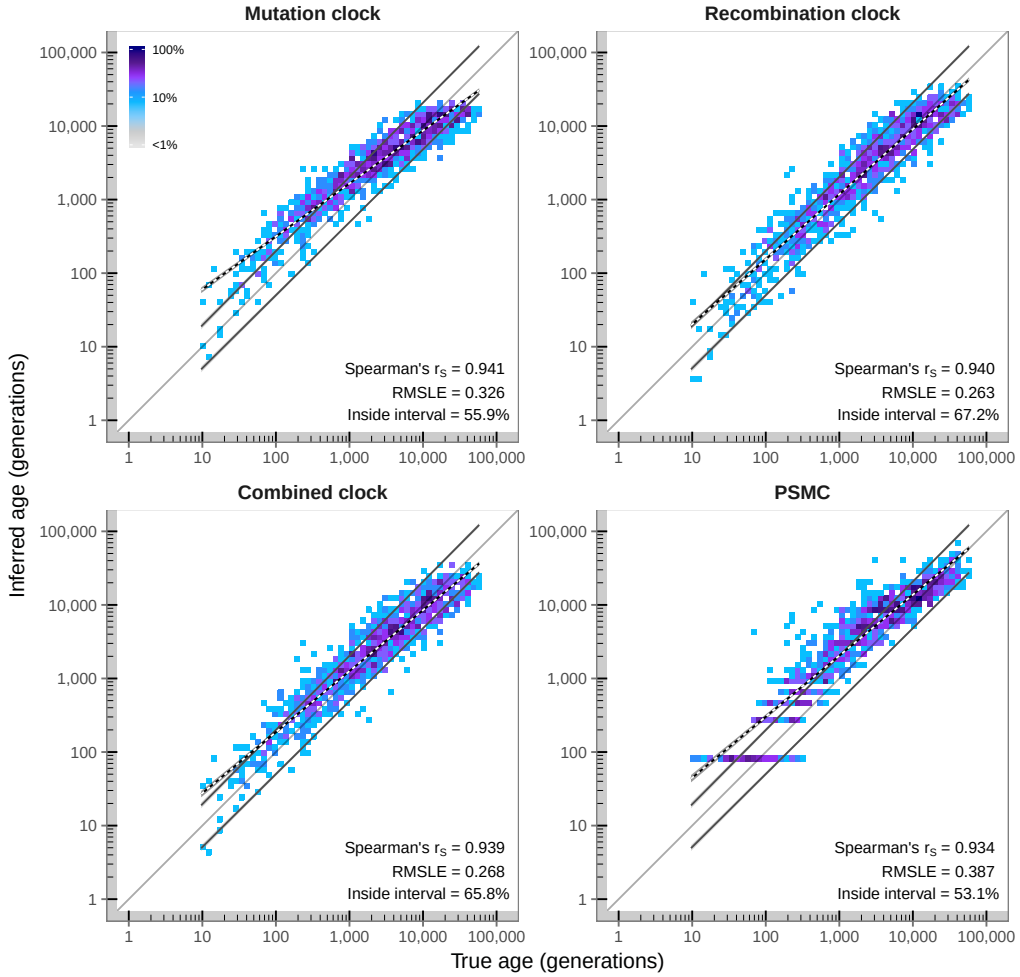

Figure 5.14: Inferred allele age compared to PSMC. ...

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

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65.

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Abecasis, G. R., Cherny, S. S., and Cardon, L. R. (2001). The impact of genotyping error on family-based analysis of quantitative traits. *European Journal of Human Genetics*, **9**(2), 130–134.

Albrechtsen, A., Korneliussen, T. S., Moltke, I., Hansen, T. v. O., Nielsen, F. C., and Nielsen, R. (2009). Relatedness Mapping and Tracts of Relatedness for Genome-Wide Data in the Presence of Linkage Disequilibrium. *Genetic Epidemiology*, **33**(3), 266–274.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic Mapping in Human Disease. *Science*, **322**(5903), 881–888.

Altshuler, D. L., Bentley, D. R., Chakravarti, A., Collins, F. S., Donnelly, P., Gabriel, S. B., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., Nickerson, D. A., Peltonen, L., Wang, J., Wilson, R. K., Gibbs, R. A., Deiros, D., Metzker, M., Wheeler, D., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G., Wang, B., Wang, W., Yang, H., Zhang, X., Zheng, H., Ambrogio, L., Bloom, T., Sougnez, C. L., Gormley, N., Humphray, S., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K. J., Costa, G. L., Ichikawa, J. K., Lee, C. C., Borodina, T. A., Dahl, A., Davydov, A. N., Marquardt, P., Mertes, F., Nietfeld, W., ROSENSTIEL, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Egholm, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Fulton, L., Weinstock, G., Burton, J., Carter, D. M., Churcher, C., Coffey, A., Cox, A., Quail, M., Skelly, T., Stalker, J., Swerdlow, H. P., Turner, D., De Witte, A., Giles, S., Sabo, A., Yu, J., Fang, X., Guo, X., Li, Y., Luo, R., Tai, S., Wu, H., Zheng, H., Zheng, X., Zhou, Y., Wang, J., Huang, W., Indap, A., Kural, D., Lee, W.-P., Stromberg, M. P., Ward, A. N., Lee, C., Mills, R. E., Shi, X., Daly, M. J., DePristo, M. A., Ball, A. D., Browning, B. L., Cibulskis, K., Garimella, K. V., Grossman, S. R., Hanna, M., Jaffe, D. B., Kernytsky, A. M., Li, H., Maguire, J. R., McCarroll, S. A., McKenna, A., Nemesh, J. C., Philippakis, A. A., Poplin, R. E., Price, A., Rivas, M. A., Sabeti, P. C., Schaffner, S. F., Shefler, E., Shlyakhter, I. A., Cooper, D. N., Ball, E. V., Mort, M., Phillips, A. D., Stenson, P. D., Sebat, J., Makarov, V., Ye, K., Yoon, S. C., Clark, A. G., Boyko, A., Degenhardt, J., Gutenkunst, R. N., Kaganovich, M., Keinan, A., Lacroute, P., Ma, X., Reynolds, A., Clarke, L., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Smith, R. E., Zalunin, V., Zheng-Bradley, X., Korbel, J. O., Stütz, A. M., Bauer, M., Keira Cheetham, R., Cox, T., Eberle, M., James, T., Kahn, S., Murray, L., Ye, K., De La Vega, F. M., Fu, Y., Hyland, F. C. L., Manning, J. M., McLaughlin, S. F., Peckham, H. E., Sakarya, O., Sun, Y. A., Tsung, E. F., Batzer, M. A., Konkel, M. K., Walker, J. A., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Herwig, R., Parkhomchuk, D. V., Agarwala, R., Khouri, H. M., Morgulis, A. O., Paschall, J. E., Phan, L. D., Rotmistrovsky, K. E., Sanders, R. D., Shumway, M. F., Xiao, C., Lunter, G., Marchini, J. L., Moutsianas, L., Myers, S., Tumian, A., Desany, B., Knight, J., Winer, R., Craig, D. W., Beckstrom-Sternberg, S. M., Christoforides, A., Kurdoglu, A. A., Pearson, J. V., Sinari, S. A., Tembe, W. D., Haussler, D., Hinrichs, A. S., Katzman, S. J., Kern, A., Kuhn, R. M., Przeworski, M., Hernandez, R. D., Howie, B., Kelley, J. L., Cord Melton, S., Abecasis, G. R., Li, Y., Anderson, P., Blackwell, T., Chen, W., Cookson, W. O., Ding, J., Min Kang, H., Lathrop, M., Liang, L., and Moffatt, M. F. a. (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, **40**(6), 695–701.

Boehnke, M. and Cox, N. J. (1997). Accurate Inference of Relationships in Sib-Pair Linkage Studies. *The American Journal of Human Genetics*, **61**(2), 423–429.

Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., van Dijk, F., Francioli, L. C., Hottenga, J. J., Laros, J. F. J., Li, Q., Li, Y., Cao, H., Chen, R., Du, Y., Li, N., Cao, S., van Setten, J., Menelaou, A., Pulit, S. L., Hehir-Kwa, J. Y., Beekman, M., Elbers, C. C., Byelas, H., de Craen, A. J. M., Deelen, P., Dijkstra, M., den Dunnen, J. T., de Knijff, P., Houwing-Duistermaat, J., Koval, V., Estrada, K., Hofman, A., Kanterakis, A., van Enckevort, D., Mai, H., Kattenberg, M., van Leeuwen, E. M., Neerincx, P. B. T., Oostra, B., Rivadeneira, F., Suchiman, E. H. D., Uitterlinden, A. G., Willemsen, G., Wolffenbuttel, B. H., Wang, J., de Bakker, P. I. W., van Ommen, G.-J., and van Duijn, C. M. (2013). The Genome of the Netherlands: design, and project goals. *European Journal of Human Genetics*, **22**(2), 221–227.

Borodulin, K., Vartiainen, E., Peltonen, M., Jousilahti, P., Juolevi, A., Laatikainen, T., Mannisto, S., Salomaa, V., Sundvall, J., and Puska, P. (2015). Forty-year trends in cardiovascular risk factors in Finland. *European Journal of Public Health*, **25**(3), 539–546.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**(3s), 228–237.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, **32**(3), 314–331.

Browning, B. L. and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American journal of human genetics*, **88**(2), 173–182.

Browning, B. L. and Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**(2), 459–471.

Browning, B. L. and Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *American journal of human genetics*, **98**(1), 116–126.

Browning, S. R. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics*, **178**(4), 2123–2132.

Browning, S. R. and Browning, B. L. (2010). High-Resolution Detection of Identity by Descent in Unrelated Individuals. *The American Journal of Human Genetics*, **86**(4), 526–539.

Browning, S. R. and Browning, B. L. (2012). Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*, **46**(1), 617–633.

Buetow, K. H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *The American Journal of Human Genetics*, **49**(5), 985–994.

Burdett, T., Hall, P., Hasting, E., Hindorff, L., Junkins, H., Klemm, A., MacArthur, J., Manolio, T., Morales, J., Parkinson, H., *et al.* (2016). The nhgri-ebi catalog of published genome-wide association studies. *Available at: www.ebi.ac.uk/gwas. Accessed 2017-01-20, version 1.0.*

Bustamante, C. D., Burchard, E. G., and De La Vega, F. M. (2011). Genomics for the world. *Nature*, **475**(7355), 163–165.

Cai, L., Fisher, A. L., Huang, H., and Xie, Z. (2016). CRISPR-mediated genome editing and human diseases. *Genes & Diseases*, **3**(4), 244–251.

Chakravarti, A. (1999). Population genetics - making sense out of sequence. *Nature Genetics*, **21**, 56–60.

Chen, J., Zhang, J.-G., Li, J., Pei, Y.-F., and Deng, H.-W. (2013). On Combining Reference Data to Improve Imputation Accuracy. *PloS one*, **8**(1).

Choi, Y., Wijsman, E. M., and Weir, B. S. (2009). Case-control association testing in the presence of unknown relationships. *Genetic Epidemiology*, **33**(8), 668–678.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Publishing Group*, **11**(6), 415–425.

Colombo, R. (2007). Dating mutations. *eLS*.

Correns, K. F. J. (1899). Untersuchungen über die Xenien bei *Zea mays*. *Berichte der Deutschen Botanischen Gesellschaft*, **17**, 410–418.

Cotterman, C. W. (1940). *A calculus for statistico-genetics*. Ph.D. thesis, The Ohio State University.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., Wheeler, D. A., Sabo, A., Lusk, C., Weiss, K. G., Akbar, H., Cree, A., Hawes, A. C., Newsham, I., Varghese, R. T., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan, M., Chang, K., Hale IV, W., Templeton, A. R., Boerwinkle, E., Gibbs, R., and Sing, C. F. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature communications*, **1**(8), 131–6.

Cox, D. G. and Kraft, P. (2006). Quantification of the Power of Hardy-Weinberg Equilibrium Testing to Detect Genotyping Error. *Human heredity*, **61**(1), 10–14.

Crow, J. F. (1954). Breeding structure of populations. ii. effective population number. *Statistics and mathematics in biology*, **543**, 556.

Crow, J. F. and Kimura, M. (1970). An introduction to population genetics theory. *An introduction to population genetics theory.*

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**(2), 229–232.

de Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, **37**(11), 1217–1223.

De Vries, H. M. (1900). Sur la loi de disjonction des hybrides. *Comptes rendus de l'Académie des Sciences*, **130**, 845–847.

Deelen, P., Kanterakis, A., van Dijk, F., Medina-Gomez, C., Kreiner-Møller, E., Rivadeneira, F., Gutierrez-Achury, J., van Enckevort, D., Dijkstra, M., Byelas, H., Genome of Netherlands Consortium, de Bakker, P. I. W., and Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *European Journal of Human Genetics*, **22**(11), 1321–1326.

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

Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. *Nature methods*, **9**(2), 179–181.

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.

Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical population biology*, **60**(3), 155–166.

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical population biology*, **23**(1), 34–63.

Douglas, J. A., Boehnke, M., and Lange, K. (2000). A Multipoint Method for Detecting Genotyping Errors and Mutations in Sibling-Pair Linkage Data. *The American Journal of Human Genetics*, **66**(4), 1287–1297.

Douglas, J. A., Skol, A. D., and Boehnke, M. (2002). Probability of Detection of Genotyping Errors and Mutations as Inheritance Inconsistencies in Nuclear-Family Data. *The American Journal of Human Genetics*, **70**(2), 487–495.

Eberle, M. A., Fritzilas, E., Krusche, P., Kallberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H.-Y., Humphray, S. J., Halpern, A. L., Kruglyak, S., Margulies, E. H., McVean, G., and Bentley, D. R. (2016). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, **27**(1), 1–9.

Ewens, W. J. (2012a). *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer Science & Business Media.

Ewens, W. J. (2012b). *Mathematical population genetics 1: theoretical introduction*, volume 27. Springer Science & Business Media.

Fisher, R. (1930a). The genetical theory of natural selection.

Fisher, R. A. (1930b). *The genetical theory of natural selection*. Oxford University Press, Oxford.

Fisher, R. A. (1949). The theory of inbreeding. *The theory of inbreeding*.

Fisher, R. A. (1954). A fuller theory of "junctions" in inbreeding. *Heredity*, **8**(2), 187–197.

Forney, G. D. (1973). The Viterbi Algorithm. In *Proceedings of the IEEE*, pages 268–278.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, **10**(4), 241–251.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, **36**(4), 388–393.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., Project, N. E. S., and Akey, J. M. (2012). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**(7431), 216–220.

Fu, Y. X. (1995). Statistical Properties of Segregating Sites. *Theoretical population biology*, **48**(2), 172–197.

Fu, Y.-X. and Li, W.-H. (1999). Coalescing into the 21st Century: An Overview and Prospects of Coalescent Theory. *Theoretical population biology*, **56**(1), 1–10.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., Rivas, M. A., Perry, J. R. B., Sim, X., Blackwell, T. W., Robertson, N. R., Rayner, N. W., Cingolani, P., Locke, A. E., Tajes, J. F., Highland, H. M., Dupuis, J., Chines, P. S., Lindgren, C. M., Hartl, C., Jackson, A. U., Chen, H., Huyghe, J. R., van de Bunt, M., Pearson, R. D., Kumar, A., Müller-Nurasyid, M., Grarup, N., Stringham, H. M., Gamazon, E. R., Lee, J., Chen, Y., Scott, R. A., Below, J. E., Chen, P., Huang, J., Go, M. J., Stitzel, M. L., Pasko, D., Parker, S. C. J., Varga, T. V., Green, T., Beer, N. L., Day-Williams, A. G., Ferreira, T., Fingerlin, T., Horikoshi, M., Hu, C., Huh, I., Ikram, M. K., Kim, B.-J., Kim, Y., Kim, Y. J., Kwon, M.-S., Lee, J., Lee, S., Lin, K.-H., Maxwell, T. J., Nagai, Y., Wang, X., Welch, R. P., Yoon, J., Zhang, W., Barzilai, N., Voight, B. F., Han, B.-G., Jenkinson, C. P., Kuulasmaa, T., Kuusisto, J., Manning, A., Ng, M. C. Y., Palmer, N. D., Balkau, B., áková, A. S., Abboud, H. E., Boeing, H., Giedraitis, V., Prabhakaran, D., Gottesman, O., Scott, J., Carey, J., Kwan, P., Grant, G., Smith, J. D., Neale, B. M., Purcell, S., Butterworth, A. S., Howson, J. M. M., Lee, H. M., Lu, Y., Kwak, S.-H., Zhao, W., Danesh, J., Lam, V. K. L., Park, K. S., Saleheen, D., So, W. Y., Tam, C. H. T., Afzal, U., Aguilar, D., Arya, R., Aung, T., Chan, E., Navarro, C., Cheng, C.-Y., Palli, D., Correa, A., Curran, J. E., Rybin, D., Farook, V. S., Fowler, S. P., Freedman, B. I., Griswold, M., Hale, D. E., Hicks, P. J., Khor, C.-C., Kumar, S., Lehne, B., Thuillier, D., Lim, W. Y., Liu, J., van der Schouw, Y. T., Loh, M., Musani, S. K., Puppala, S., Scott, W. R., Yengo, L., Tan, S.-T., Taylor, H. A., Thameem, F., Wilson, G., Wong, T. Y., Njølstad, P. R., Levy, J. C., Mangino, M., Bonnycastle, L. L., Schwarzmayr, T., Fadista, J., Surdulescu, G. L., Herder, C., Groves, C. J., Wieland, T., Bork-Jensen, J., Brandslund, I., Christensen, C., Koistinen, H. A., Doney, A. S. F., Kinnunen, L., Esko, T., Farmer, A. J., Hakaste, L., Hodgkiss, D., Kravic, J., Lyssenko, V., Hollensted, M., Jørgensen, M. E., Jørgensen, T., Ladenvall, C., Justesen, J. M., Käräjämäki, A., Kriebel, J., Rathmann, W., Lannfelt, L., Lauritzen, T., Narisu, N., Linneberg, A., Melander, O., Milani, L., Neville, M., Orho-Melander, M., Qi, L., Qi, Q., Roden, M., Rolandsson, O., Swift, A., Rosengren, A. H., Stirrups, K., Wood, A. R., Mihailov, E., Blancher, C., Carneiro, M. O., Maguire, J., Poplin, R., Shakir, K., Fennell, T., DePristo, M., de Angelis, M. H., Deloukas, P., Gjesing, A. P., Jun, G., Nilsson, P., Murphy, J., Onofrio, R., Thorand, B., Hansen, T., Meisinger, C., Hu, F. B., Isomaa, B., Karpe, F., Liang, L., Peters, A., Huth, C., O'Rahilly, S. P., Palmer, C. N. A., Pedersen, O., Rauramaa, R., Tuomilehto, J., Salomaa, V., Watanabe, R. M., Syvänen, A.-C., Bergman, R. N., Bharadwaj, D., Bottinger, E. P., Cho, Y. S., Chandak, G. R., Chan, J. C. N., Chia, K. S., Daly, M. J., Ebrahim, S. B., Langenberg, C., Elliott, P., Jablonski, K. A., Lehman, D. M., Jia, W., and Ma, R. (2016). The genetic architecture of type 2 diabetes. *Nature*, **536**(7614), 41–47.

Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, **46**(8), 818–825.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, **13**(2), 135–145.

Gordon, D., Heath, S. C., Liu, X., and Ott, J. (2001). A Transmission/Disequilibrium Test That Allows for Genotyping Errors in the Analysis of Single-Nucleotide Polymorphism Data. *The American Journal of Human Genetics*, **69**(2), 371–380.

Gordon, D., Finch, S. J., Nothnagel, M., and Ott, J. u. r. (2002). Power and Sample Size Calculations for Case-Control Genetic Association Tests when Errors Are Present: Application to Single Nucleotide Polymorphisms. *Human heredity*, **54**(1), 22–33.

Gore, A., Li, Z., Fung, H.-L., Young, J. E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M. A., Kiskinis, E., Lee, J.-H., Loh, Y.-H., Manos, P. D., Montserrat, N., Panopoulos, A. D., Ruiz, S., Wilbert, M. L., Yu, J., Kirkness, E. F., Izpisua Belmonte, J. C., Rossi, D. J., Thomson, J. A., Eggan, K., Daley, G. Q., Goldstein, L. S. B., and Zhang, K. (2011). Somatic coding mutations in human induced pluripotent stem cells. *Nature*, **471**(7336), 63–67.

Griffiths, R. C. (1991). The Two-Locus Ancestral Graph. In *Selected Proceedings of the Sheffield Symposium on Applied Probability*, pages 100–117. Institute of Mathematical Statistics, Hayward, CA.

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, **3**(4), 479–502.

Griffiths, R. C. and Marjoram, P. (1997a). An ancestral recombination graph. *Institute for Mathematics and its Applications*, **87**, 257.

Griffiths, R. C. and Marjoram, P. (1997b). Progress in population genetics and human evolution.

Griffiths, R. C. and Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models*, **14**(1-2), 273–295.

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, **19**(2), 318–326.

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.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, **8**(29), 299–309.

Hardy, G. H. (1908). Mendelian Proportions in a Mixed Population. *Science*, **28**(706), 49–50.

Harris, K. and Nielsen, R. (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, **9**(6).

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, **13**(4), 635–643.

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

Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., 1000 Genomes Project, Durbin, R. M., Flicek, P., Gabriel, S. B., Lander, E. S., Wheeler, D., Cibulskis, K., Fennell, T. J., Jaffe, D. B., Shefler, E., Egholm, M., Fulton, R., Bainbridge, M., Challis, D., Sabo, A., Wang, J., Lee, C., Korn, J. M., Sudbrak, R., Auton, A., Iqbal, Z., Desany, B., Dooling, D., Hurles, M. E., MacArthur, D. G., Abyzov, A., Zhang, Z., Garrison, E. P., Banks, E., Handsaker, R. E., Hartl, C., De La Vega, F. M., Alkan, C., Snyder, M., Muzny, D., Reid, J., Quinlan, A. R., Stewart, C., Wu, J., Gravel, S., Sherry, S. T., McVean, G. A., Abecasis, G. R., Koboldt, D. C., Palotie, A., Bustamante, C. D., Schafer, A. J., and Brooks, L. D. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, **108**(29), 11983–11988.

Howie, B., Marchini, J., and Stephens, M. (2011a). Genotype imputation with thousands of genomes. *G3 (Bethesda, Md.)*, **1**(6), 457–470.

Howie, B., Marchini, J., and Stephens, M. (2011b). Genotype Imputation with Thousands of Genomes. *G3 (Bethesda, Md.)*, **1**(6), 457–470.

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

Hudson, R. (1990). Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, **7**(1), 44.

Hudson, R. and Kaplan, N. L. (1985). Statistical Properties of the Number of Recombination Events in the History of a Sample of Dna-Sequences. *Genetics*, **111**(1), 147–164.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, **23**(2), 183–201.

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

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**(2), 337–338.

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 (2003). The International HapMap Project. *Nature*, **426**(6968), 789–796.

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.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011), 931–945.

Kaiser, J. (2008). A plan to capture human diversity in 1000 genomes (Science (395)). *Science*, **319**(5868), 1336.

Kang, S. J., Gordon, D., and Finch, S. J. (2004). What SNP genotyping errors are most costly for genetic association studies? *Genetic Epidemiology*, **26**(2), 132–141.

Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**(6082), 740–743.

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

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

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

Kingman, J. F. C. (1982a). Exchangeability and the evolution of large populations. *Exchangeability in Probability and Statistics*, pages 97—-112.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of Applied Probability*, **19**(A), 27–43.

Kingman, J. F. C. (1982c). The coalescent. *Stochastic processes and their applications*, **13**(3), 235–248.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, **308**(5720), 385–389.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., and Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, **40**(9), 1068–1075.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**(2), 139–144.

Kryukov, G. V., Pennacchio, L. A., and Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, **80**(4), 727–739.

Labuda, M., Labuda, D., Korab-Laskowska, M., Cole, D. E., Zietkiewicz, E., Weissenbach, J., Popowska, E., Pronicka, E., Root, A. W., and Glorieux, F. H. (1996). Linkage disequilibrium analysis in young populations: pseudo-vitamin d-deficiency rickets and the founder effect in french canadians. *American journal of human genetics*, **59**(3), 633.

Lander, E. S. (1996). The new genomics: Global views of biology. *Science*, **274**(5287), 536–539.

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, **8**(1), e1002453.

Leutenegger, A.-L., Prum, B., Génin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E. A. (2003). Estimation of the Inbreeding Coefficient through Use of Genomic Data. *American journal of human genetics*, **73**(3), 516–523.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**(4), 2213–2233.

Li, W. H. (1975). The first arrival time and mean age of a deleterious mutant gene in a finite population. *The American Journal of Human Genetics*, **27**(3), 274–286.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annual review of genomics and human genetics*, **10**, 387–406.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34**(8), 816–834.

Lim, E. T., Würtz, P., Havulinna, A. S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., Chan, Y., Salem, R. M., Lek, M., Flannick, J., Sim, X., Manning, A., Ladenvall, C., Bumpstead, S., Hämäläinen, E., Aalto, K., Maksimow, M., Salmi, M., Blankenberg, S., Ardissino, D., Shah, S., Horne, B., McPherson, R., Hovingh, G. K., Reilly, M. P., Watkins, H., Goel, A., Farrall, M., Girelli, D., Reiner, A. P., Stitziel, N. O., Kathiresan, S., Gabriel, S., Barrett, J. C., Lehtimäki, T., Laakso, M., Groop, L., Kaprio, J., Perola, M., McCarthy, M. I., Boehnke, M., Altshuler, D. M., Lindgren, C. M., Hirschhorn, J. N., Metspalu, A., Freimer, N. B., Zeller, T., Jalkanen, S., Koskinen, S., Raitakari, O., Durbin, R., MacArthur, D. G., Salomaa, V., Ripatti, S., Daly, M. J., Palotie, A., and for the Sequencing Initiative Suomi (SISu) Project (2014). Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genetics*, **10**(7), e1004494–12.

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

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of biomedicine & biotechnology*, **2012**(7), 1–11.

Loh, P.-R., Palamara, P. F., and Price, A. L. (2016a). Fast and accurate long-range phasing in a uk biobank cohort. *Nature genetics*.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., and Price, A. L. (2016b). Reference-based phasing using the Haplotype Reference Consortium panel. Technical report.

Malécot, G. (1948). Mathematics of heredity. *Les mathematiques de l'heredite.*

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–753.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. **11**(7), 499–511.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, **39**(7), 906–913.

Mardis, E. R. (2006). Anticipating the 1,000 dollar genome. *Genome biology*, **7**(7), 112.

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, **12**(2), 213–218.

Marth, G. T., Yu, F., Indap, A. R., Garimella, K., Gravel, S., Leong, W. F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., Chen, Y., Challis, D., Clarke, L., Ball, E. V., Cibulskis, K., Cooper, D. N., Fulton, B., Hartl, C., Koboldt, D., Muzny, D., Smith, R., Sougnez, C., Stewart, C., Ward, A., Yu, J., Xue, Y., Altshuler, D., Bustamante, C. D., Clark, A. G., Daly, M., DePristo, M., Flicek, P., Gabriel, S., Mardis, E., Palotie, A., Gibbs, R., and 1000 Genomes Project (2011). The functional spectrum of low-frequency coding variation. *Genome biology*, **12**(9), R84.

Maruyama, T. (1974). The age of an allele in a finite population. *Genetical research*, **23**(2), 137–143.

Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, **44**(3), 243–U29.

Mathieson, I. and McVean, G. (2014). Demography and the Age of Rare Variants. *PLoS Genetics*, **10**(8), e1004528.

Maynard Smith, J. (1989). *Evolutionary genetics.* Oxford University Press.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, **40**(10), 1166–1174.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Publishing Group*, **9**(5), 356–369.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., van Duijn, C. M., Gillies, C. E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J. C., Boomsma, D., Branham, K., Breen, G., Brummett, C. M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F. S., Corbin, L. J., Smith, G. D., Dedoussis, G., Dorr, M., Farmaki, A.-E., Ferrucci, L., Forer, L., Fraser, R. M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O. L., Hveem, K., Kretzler, M., Lee, J. C., McGue, M., Meitinger, T., Melzer, D., Min, J. L., Mohlke, K. L., Vincent, J. B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J. B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P. E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Van den Berg, L. H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M. G., Wilson, J. F., Frayling, T., de Bakker, P. I. W., Swertz, M. A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C. A., Myers, R. M., Boehnke, M., McCarthy, M. I., Durbin, R., Abecasis, G., and Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, **48**(10), 1279–1283.

McClellan, J. and King, M.-C. (2010). Genetic Heterogeneity in Human Disease. *Cell*, **141**(2), 210–217.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., MacLeod, A. K., Farrington, S. M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., Campbell, H., and Wilson, J. F. (2008). Runs of Homozygosity in European Populations. *The American Journal of Human Genetics*, **83**(3), 359–372.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*.

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.

Mendel, G. (1866). Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, **4**, 3–47.

Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Publishing Group*, **11**(1), 31–46.

Milligan, B. G. (2003). Maximum-Likelihood Estimation of Relatedness. *Genetics*, **163**(3), 1153–1167.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, **38**(8).

Morgan, T. H. (1911). Random segregation versus coupling in Mendelian inheritance. *Science*, **34**(873), 384–384.

Morral, N., Bertranpetit, J., Estivill, X., and Nunes, V. (1994). The origin of the major cystic fibrosis mutation (DF508) in European populations. *Nature*.

Morris, A. and Cardon, L. (2007). Whole genome association. *Handbook of Statistical Genetics, Third Edition*, pages 1238–1263.

Moskvina, V. and Schmidt, K. M. (2006). Susceptibility of Biallelic Haplotype and Genotype Frequencies to Genotyping Error. *Biometrics*, **62**(4), 1116–1123.

Moskvina, V., Craddock, N., Holmans, P., Owen, M., and O'Donovan, M. (2005). Minor genotyping error can result in substantial elevation in type i error rate in haplotype based case control analysis. In *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, volume 138, pages 19–19.

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., Albers, P. K., GoT2D Consortium, McVean, G., Boehnke, M., Altshuler, D., and McCarthy, M. I. (2015). The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLoS Genetics*, **11**(4), e1005165–24.

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*.

Naidoo, N., Pawitan, Y., Soong, R., Cooper, D. N., and Ku, C.-S. (2011). Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Human Genomics*, **5**(6), 577–622.

Neuhauser, C. (2001). Mathematical models in population genetics. *Handbook of statistical genetics*.

Nordborg, M. (2001). Coalescent theory. *Handbook of statistical genetics*.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., Navarro, P., Zagury, J.-F., Wilson, J. F., Toniolo, D., Gasparini, P., Soranzo, N., Sandhu, M. S., and Marchini, J. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics*, **10**(4), e1004234–21.

Ott, J. (1999). *Analysis of human genetic linkage*. JHU Press.

Pajunen, P., Rissanen, H., Härkänen, T., Jula, A., Reunanen, A., and Salomaa, V. (2010). The metabolic syndrome as a predictor of incident diabetes and cardiovascular events in the Health 2000 Study. *Diabetes and Metabolism*, **36**(5), 395–401.

Palamara, P. F. and Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics*, **29**(13), i180–i188.

Palamara, P. F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. *The American Journal of Human Genetics*, **91**(5), 809–822.

Palin, K., Campbell, H., Wright, A. F., Wilson, J. F., and Durbin, R. (2011). Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic Epidemiology*, **35**(8), 853–860.

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

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics (Oxford, England)*, **30**(20), 2906–2914.

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

Pe'er, I., de Bakker, P. I. W., Maller, J., Yelensky, R., Altshuler, D., and Daly, M. J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics*, **38**(6), 663–667.

Pennisi, E. (2007). Human Genetic Variation. *Science*, **318**(5858), 1842–1843.

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Publishing Group*, **11**(11), 800–805.

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

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, **69**(1), 124–137.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**(3), 559–575.

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

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*, **10**(5), e1004342–27.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, **411**(6834), 199–204.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, **273**(5281), 1516–1517.

Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield, X., and Bressman, S. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nature Genetics*, **9**(2), 152–159.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, **405**(6788), 847–856.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, D. J. (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*, **328**(5978), 636–639.

Roshyara, N. R. and Scholz, M. (2015). Impact of genetic similarity on imputation accuracy. *BMC Genetics*, **16**(1), 1–16.

Rousset, F. (2002). Inbreeding and relatedness coefficients: What do they measure? *Heredity*, **88**(5), 371–380.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, **94**(3), 441–448.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463–5467.

Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, **13**(10), 745–753.

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, **19**(3), 212–219.

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

Serre, J. L., Simon-Bouy, B., Mornet, E., Jaume-Roig, B., Balassopoulou, A., Schwartz, M., Taillandier, A., Boue, J., and Boue, A. (1990). Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. *Human genetics*, **84**(5), 449–454.

Shields, D. C., Collins, A., Buetow, K. H., and Morton, N. E. (1991). Error filtration, interference, and the human linkage map. *Proceedings of the National Academy of Sciences of the United States of America*, **88**(15), 6501–6505.

Simons, Y. B., Turchin, M. C., Pritchard, J. K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nature Publishing Group*, **46**(3), 220–224.

Slack, J. (2014). *Genes: A Very Short Introduction*. Oxford University Press, Oxford.

Slatkin, M. (2000). Allele age and a test for selection on rare alleles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **355**(1403), 1663–1668.

Slatkin, M. (2008a). Inbreeding coefficients and coalescence times. *doi.org*, pages 1–9.

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

Slatkin, M. and Bertorelle, G. (2001). The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics*, **158**(2), 865–874.

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

Sobel, E., Papp, J. C., and Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics*, **70**(2), 496–508.

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical research*, **35**(02), 131.

Stone, M. (1961). The Opinion Pool. *The Annals of Mathematical Statistics*, **32**(4), 1339–1342.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, **14**(1), 43–59.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**(16), 2304–2305.

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

Tajima, F. (1993). Measurement of dna polymorphism. *Mechanisms of molecular evolution*, pages 37–59.

Takahata, N. (1993). Allelic genealogy and human evolution. *Molecular Biology and Evolution*, **10**(1), 2–22.

Tavaré, S. (2004). Part I: Ancestral inference in population genetics. In *Lectures on probability theory and statistics*, pages 1–188. Springer.

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

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., GO, B., GO, S., and Project, N. E. S. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*, **337**(6090), 64–69.

Thompson, E. A. (1974). Gene Identities and Multiple Relationships. *Biometrics*, **30**(4), 667.

Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of human genetics*, **39**(2), 173–188.

Thompson, E. A. (1976). Estimation of age and rate of increase of rare variants. *The American Journal of Human Genetics*, **28**(5), 442–452.

Thompson, E. A. (2008). The IBD process along four chromosomes. *Theoretical population biology*, **73**(3), 369–373.

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, **194**(2), 301–326.

Tschermak, E. (1900). Über künstliche Kreuzung bei *Pisum sativum*. *Berichte der Deutschen Botanischen Gesellschaft*, **18**, 232–239.

UK10K Consortium, Walter, K., Crooks, L., Memari, Y., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Li, R., Floyd, J., Wain, L. V., Humphries, S. E., Barrett, J. C., Plagnol, V., Richards, J. B., Greenwood, C. M. T., Timpson, N. J., Soranzo, N., Danecek, P., Barroso, I., McCarthy, S., Tachmazidou, I., Durbin, R., Hurles, M. E., Kennedy, K., Palotie, A., Zeggini, E., Cocca, M., Huang, J., and Min, J. L. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, **526**(7571), 82–90.

Vartiainen, E., Laatikainen, T., Peltonen, M., Juolevi, A., Mannisto, S., Sundvall, J., Jousilahti, P., Salomaa, V., Valsta, L., and Puska, P. (2010). Thirty-five-year trends in cardiovascular risk factors in Finland. *International Journal of Epidemiology*, **39**(2), 504–518.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Angela Center, Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., and Majoros... (2001). The Sequence of the Human Genome. *Science*, **291**(5507), 1304–1351.

Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, **13**(2), 260–269.

Voight, B. F. and Pritchard, J. K. (2005). Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLoS Genetics*, **1**(3), e32–10.

Wakeley, J. (2008). *Coalescent Theory: An Introduction*. W. H. Freeman.

Wakeley, J. and Wilton, P. (2016). Coalescent and models of identity by descent. In R. M. Kliman, editor, *Encyclopedia of Evolutionary Biology*, pages 287 – 292. Academic Press, Oxford.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1459), 1395–1409.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**(4356), 737–738.

Watterson, G. (1996). Motoo Kimura's Use of Diffusion Theory in Population Genetics. *Theoretical population biology*, **49**(2), 154–188.

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

Watterson, G. A. (1976). Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theoretical Population Biology*, **10**(3), 239–253.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg.*, **64**, 368–382.

Weissenbach, J. (1993). A second generation linkage map of the human genome based on highly informative microsatellite loci. *Gene*, **135**(1-2), 275–278.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A., and Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**(7189), 872–876.

Winkler, H. (1920). *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreich*. Verlag G. Fischer, Jena.

Wiuf, C. and Hein, J. (1997). On the number of ancestors to a DNA sequence. *Genetics*, **147**(3), 1459–1468.

Wright, S. (1921). Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics*, **6**(2), 111–123.

Wright, S. (1922). Coefficients of Inbreeding and Relationship. *The American Naturalist*, **56**(645), 330–338.

Wright, S. (1931a). Evolution in Mendelian Populations. *Genetics*, **16**(2), 97–159.

Wright, S. (1931b). Evolution in Mendelian Populations. *Genetics*, **16**(2), 97–159.

Yu, N., Zhao, Z., Fu, Y. X., Sambuughin, N., Ramsay, M., Jenkins, T., Leskinen, E., Patthy, L., Jorde, L. B., Kuromori, T., and Li, W. H. (2001). Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Molecular Biology and Evolution*, **18**(2), 214–222.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, **111**(4), E455–64.

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