Rare and low-frequency variants and predisposition to complex disease

Patrick K. Albers

Wellcome Trust Centre for Human Genetics
Big Data Institute
Medical Sciences Division

Green Templeton College

University of Oxford

Supervised by

Professor Gil McVean

Professor Mark McCarthy

Submitted in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy (DPhil)

Hilary 2017

Abstract

Advances in high-throughput genomic technologies have facilitated the collection of DNA information for thousands of individuals, providing unprecedented opportunities to explore the genetic architecture of complex disease. One important finding has been that the majority of variants in the human genome are low in frequency or rare. It has been hypothesised that recent explosive growth of the human population afforded unexpectedly large amounts of rare variants with potentially deleterious effects, suggesting that rare variants may play a role in disease predisposition. But, importantly, rare variants embody a source of information through which we may learn more about our recent evolutionary history. In this thesis, I developed several statistical and computational methods to address problems associated with the analysis of rare variants and, foremost, to leverage the genealogical information they encode.

First, one constraint in genome-wide association studies is that lower-frequency variants are not well captured by genotyping methods, but instead are predicted through imputation from a reference dataset. I developed the *meta-imputation* method to improve imputation accuracy by integrating genotype data from multiple, independent reference panels, which outperformed imputations from separate references in almost all comparisons (mean correlation with masked genotypes $r^2 > 0.9$). I further demonstrated in simulated case-control studies that meta-imputation increased the statistical power to identify low-frequency variants of intermediate or high penetrance by 2.2–3.6%.

Second, rare variants are likely to have originated recently through mutation and thereby sit on relatively long haplotype regions identical by descent (IBD). I developed a method that exploits rare variants as identifiers for shared haplotype segments around which the breakpoints of recombination are detected using non-probabilistic approaches. In coalescent simulations, I show that such breakpoints can be inferred with high accuracy ($r^2 > 0.99$) around rare variants at frequencies $\leq 0.05\%$, using either haplotype or genotype data.

Third, I show that technical error poses a major problem for the analysis of whole-genome sequencing or genotyping data, particularly for alleles below 0.05% frequency (false positive rate, FPR = 0.1). I therefore propose a novel approach to infer IBD segments using a Hidden Markov Model (HMM) which operates on genotype data alone. I incorporated an empirical error model constructed from error rates I estimated in publicly available sequencing and genotyping datasets. The HMM was robust in presence of error in simulated data ($r^2 > 0.98$) while non-probabilistic methods failed ($r^2 < 0.02$).

Lastly, the age of an allele (the time since its creation through mutation) may provide clues about demographic processes that resulted in its observed frequency. I present a novel method to estimate (rare) allele age based on the inferred shared haplotype structure of the sample. The method operates in a Bayesian framework to infer pairwise coalescent times from which the age is estimated using a composite posterior approach. I show in simulated data that coalescent time can be inferred with high accuracy (rank correlation > 0.91) which resulted in a likewise high accuracy for estimated age (> 0.94). When applied to data from the 1000 Genomes Project, I show that estimated age distributions were overall conform with frequency-dependent expectations under neutrality, but where patterns of low frequency and old age may hint at signatures of selection at certain sites. Thus, this method may prove useful in the analysis of large cohorts when linked to biomedical phenotype data.

Acknowledgements

This work would not have been possible without the patience of my supervisors, Gil McVean and Mark McCarthy. If my work will prove to be useful, it is because of their guidance.

I dedicate this thesis to my parents, for their unconditional love and their relentless efforts to support me in my studies. I also want to express my gratitude to my mentors, who have guided me over the years, and those who taught me valuable lessons, during my time as a Bachelor student, during my time as a Master student, and in between; to Nico Michiels (Eberhard Karls Universität Tübingen, Germany) who was the first to spark my interest in statistics, to Christopher Bartlett (James Cook University, Australia) who made my first research project possible, to the people of Vanuatu who will always have a special place in my heart (I was the scientific advisor to a marine protected area), to the people of Laos who taught me to appreciate the small things in life (I was a teacher in a Buddhist temple school), to Mikael Thollesson (Uppsala Universitet, Sweden) who fostered my enthusiasm for Evolutionary Biology, to Sophie Caillon (Université de Montpellier II, France) who showed me that communication is key, to Scott Edwards (Harvard University, United States) who stopped working on a crucial grant application to write a recommendation letter for my application to the University of Oxford, to Dirk Metzler (Ludwig-Maximilians Universität Munich, Germany) who taught me to not blindly trust any statistical results. Lastly, I want to thank my friends for their support and their understanding.

Contents

List of Figures

List of Tables

Two distinct elements are included under the term "inheritance" –
the transmission, and the development of characters.

— Charles Darwin, The Descent of Man

1 Introduction

Contents

The human genome consists of 23 chromosome pairs which harbour more than 20 thousand genes embedded in a filigree molecular filament that encodes a sequence which is more than 3 billion nucleotides long and which itself is the result of an ongoing evolutionary process that began when life emerged on this planet around 3.5 billion years ago; yet all of this information is compacted into the 10 µm wide nucleus of a cell. The genetic material contained within this microscopic dot determines the development of an organism, its ability to interact with and react to the environment, as well as its predisposition to disease.

One of the main goals of modern genetic research is to learn about the genetic architecture that underpins heritable disease traits. Early efforts in disease research had been directed towards the identification of genetic variants with highly penetrant effects on disease traits; *e.g.* mutations that contribute to distinct phenotypes, such as cystic fibrosis or Huntington's disease, which typically segregate within families (*i.e.* *monogenic* or *Mendelian* diseases). The classical approach to locating (or *mapping*) the genetic factors involved in such 'simple' diseases is linkage analysis within affected families (*e.g.*, see Morris and Cardon, 2007). While linkage studies have been successful in the identification of genetic factors underlying Mendelian diseases (Altshuler *et al.*, 2008), they have been less powerful with regard to locating variants that influence complex disease risk, such as type 2 diabetes, because each variant individually may only contribute modestly to disease susceptibility (Risch, 2000; Botstein and Risch, 2003). Genome-wide association (GWA) studies have become the preferred method to interrogate common variants in the context of complex traits; they have uncovered significant associations between thousands of genetic factors and major common diseases, and have been a driving force in the ongoing accumulation of more, larger, and denser genomic datasets.

One major insight gained from the extensive study of the (human) genome is that the genetic variation between individuals is mostly determined by *common* variants (*e.g.* $\geq$ 5% frequency), but most variant sites in the genome are *rare*; that is, a particular allele is shared by only few individuals in the population (*e.g.* 1 in 1,000). This abundance of rare variants in the human genome can be seen as a predictable consequence of a recent, exponential growth of the human population (Fu, 1995). In general, low-frequency variants tend to be population-specific, but may also be highly differentiated between demographic groups on a finer scale (Henn *et al.*, 2011; Bustamante *et al.*, 2011; Mathieson and McVean, 2014). This is because rare variants are likely to have a relatively recent origin through mutation; *i.e.* they are "young" in age and therefore have less time to spread. Conversely, genetic factors that contribute to substantial disease risk (particularly with early onset) are likely to be under purifying selection, which implies that they should be observed at relatively low frequencies, *e.g.* despite being "old". However, recent research has indicated that the human genome harbours an excess of rare, functional variants, which may entail deleterious consequences, due to the combined effects of recent, explosive growth and weak purifying selection (*e.g.*, see Kryukov *et al.*, 2007; Marth *et al.*, 2011; Coventry *et al.*, 2010; Keinan and Clark, 2012; Tennessen *et al.*, 2012).

Rare variants are now considered to be potentially involved in the predisposition to complex disease (Bodmer and Bonilla, 2008; Schork *et al.*, 2009; McClellan and King, 2010; Cirulli and Goldstein, 2010), though their contribution has been hypothesised for more than a decade (Pritchard, 2001). Notably, it has been hypothesised that rare variants may help understand the problem of *missing heritability*, where the genetic loci detected through GWA studies can only explain a small fraction of the genetic variance inferred for a disease trait (Manolio *et al.*, 2009; Gibson, 2012; Zuk *et al.*, 2014). However, the interrogation of alleles found at lower frequencies is not straightforward. For instance, rare variants may not exert large enough effects to be detected by family-based linkage studies. Conversely, rare alleles are generally too low in frequency to achieve statistical significance in association tests. An additional complication applies, namely that genotyping arrays are typically not designed to capture low-frequency variants and, on the other hand, sequencing coverage may be insufficient to call rare variants with confidence. Hence, there are considerable challenges to be addressed.

1.1 Aims and structure of this thesis

The overall aim of this thesis is to develop novel strategies and computational methods to harness the information represented by rare and low-frequency variants, and to demonstrate that these methods provide workable solutions for application to existing genomic datasets. In particular, I address the problems typically associated with the analysis of rare variants but, primarily, I focus on the opportunity that arises from the genealogical properties of alleles found at lower frequencies. Thus, the aims of this work relate to the "heads and tails" of rare variants and can be summarised as follows.

- To increase the statistical power to detect significant signals in GWA studies by developing a method that integrates information from multiple, independently obtained reference datasets for imputation into a given study sample; thus attempting to optimise the ability to implicate low-frequency and rare variants as contributing factors to disease risk.

- To utilise rare variants as a source of information about relatedness and haplotype sharing by descent, which aligns with two objectives; first, to develop a method for the inference of the underlying identity by descent (IBD) structure in which a given allele of interest is embedded, and second, from this, to develop a method to reconstruct the sequence of coalescent events such that the age of the allele can be estimated.

These two main goals entail distinct analytical paradigms, both being motivated by the overarching purpose to learn more about the genetic architecture that predisposes disease risk. Under the first paradigm, the genetic variation observed in a sample is examined in order to discern variants that associate with a certain phenotypic (disease) trait; this approach can therefore be described as being *phenotype-focused*, which I cover only in the first chapter following this introduction. In the chapters thereafter, I advocate a *variant-centric* approach, which aims to better understand the patterns of descent that led to the emergence of a disease phenotype in a population. In particular, knowledge about allele age is of interest to a wide range of problems studied in both population and medical genetics, as it allows us to observe demographic changes over time and to learn more about past events and evolutionary processes which came into effect somewhere along the branches of a genealogical tree.

In the following, I further describe the structure of this thesis by briefly presenting the objectives as addressed in each chapter. In the remainder of this introduction (**Chapter 1**), I explain the basic terminology and provide further information about the subjects touched upon below.

Chapter 2. I focus on the population or cohort-specific coverage of genetic variation as a limiting factor to the imputation and subsequent interrogation of low-frequency and rare variants in GWA studies. I propose a new method which integrates genotype data after performing separate imputations from multiple reference panels into a given study sample, such that the combined set of variants across references is available for association analysis.

Chapter 3. I propose a non-probabilistic method for the detection of recombination events around target sites in either haplotype or genotype data. The method capitalises on the presumed young age of rare variants to identify (recent) relatedness in samples of reportedly unrelated individuals, thereby facilitating the detection of relatively long stretches of pairwise shared haplotypes that are identical by descent (IBD).

Chapter 4. I characterise the extent of genotype error in data obtained on different genotyping and sequencing platforms, so as to investigate the impact of error on IBD detection. The results of this analysis are incorporated in a probabilistic model that is enabled for the inference of IBD tracts using a Hidden Markov Model (HMM), thereby improving on the method presented in Chapter 3.

Chapter 5. I propose a novel method for the estimation of (rare) allele age, *i.e.* the time since a mutation event gave rise to a particular allele that is observed in sample data. The method represents a composite Bayesian analysis and operates on insights gained from the inferred shared haplotype structure of the sample; thus, prior knowledge of the demographic history of the population or the genealogy of the sample is not required. I apply this method to rare variants in the 1000 Genomes Project (1000G) Phase III.

Lastly, I conclude this thesis by providing a summary of the relevant results and by highlighting the implications of the presented methodology for future research (**Chapter 6**).

In the following, I outline the biological concepts relevant to define basic terminology (Section 1.2, this page), as well as the principal definitions in population genetics that underpin the methodology developed in this thesis (Section 1.3, page 11). I then provide a summary of available genomic technologies which are the essential tools for the exploration of the human genome (Section 1.4, page 25). I reserve the remaining sections of this chapter to provide an introduction to genome-wide association (Section 1.5, page 30), the definition of identity by descent (Section 1.6, page 32), and the implications of allele age estimation (Section 1.7, page 35).

1.2 Basic concepts and terminology

The term *genome*, which was coined almost a century ago (Winkler, 1920), refers to the totality of the genetic hereditary information and its organisation into *chromosomes*. The number of chromosomes is characteristic for an organism, as is the number of chromosome sets, referred to as *ploidy*. Cells with only one set of chromosomes, are said to be *haploid*. In most animal species, including humans, somatic cells typically carry two sets of chromosomes, where one set is derived from each parent; *i.e.* they are said to be *diploid*. Chromosomes can be further distinguished into *autosomes* and *allosomes* (or "sex chromosomes") in sexually reproducing organisms. Human cells carry 22 autosome pairs, which are *homologous* in both males and females, and one set of allosomes (X and Y chromosomes), which determine sex and thus differ in males and females.

Deoxyribonucleic acid (DNA) forms the molecular basis of what is commonly referred to as "genetic material". The molecular structure of DNA was first described by Watson and Crick (1953) on basis of X-ray diffraction data by Rosalind Franklin. A chromosome

is a single DNA molecule composed of two strands that form a double helical structure. Each strand is a chain of *nucleotide* subunits containing one of four *nucleobases*; adenine (A), guanine (G), cytosine (C), and thymine (T), which constitute the alphabet of the genetic code. The DNA double helix is held together through hydrogen bonds between complementary nucleobases on opposite strands. The human genome contains more than 3 billion such *basepairs*. The chemical structure of the DNA double helix is illustrated in Figure 1.1 (this page).

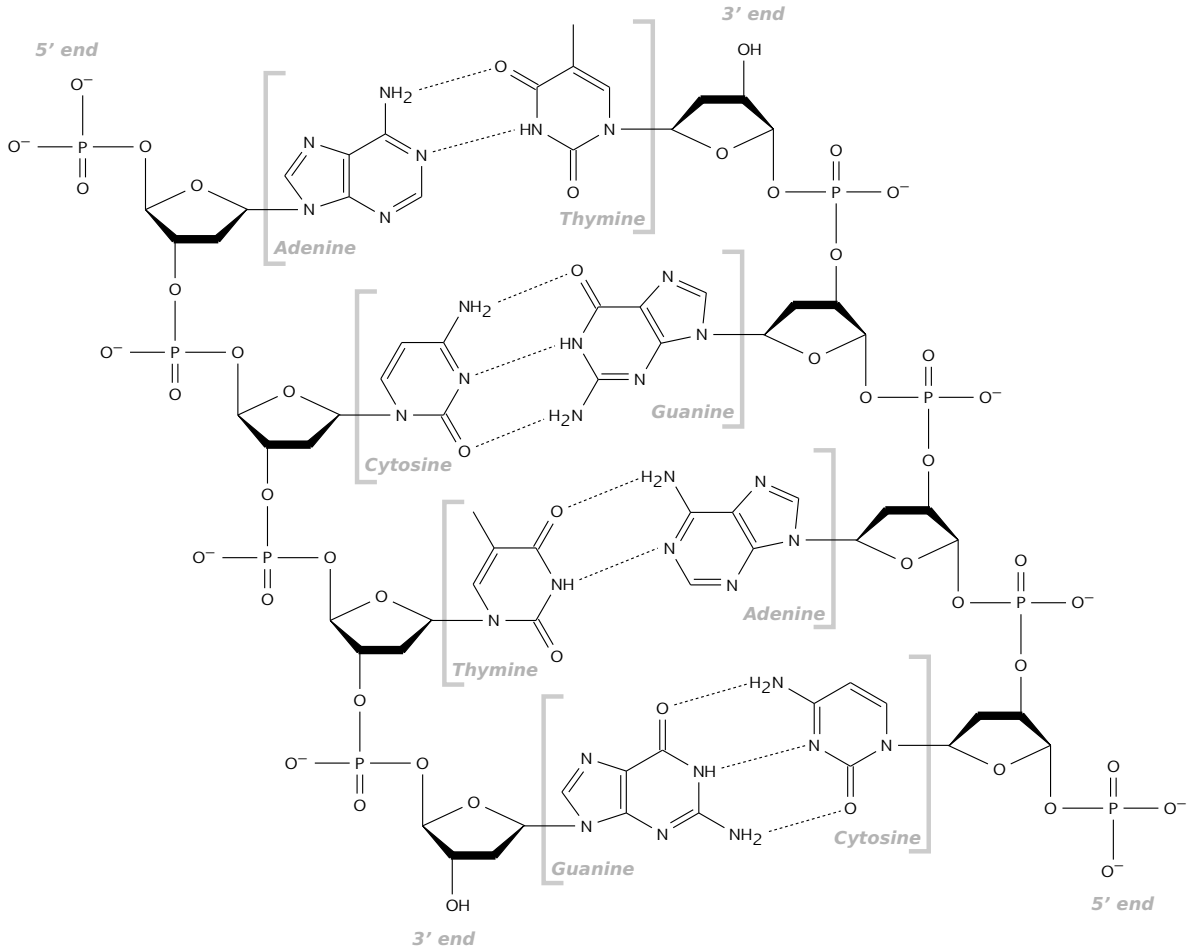


Figure 1.1: The chemical structure of DNA. The DNA molecule is a long chain polymer of individual nucleotide building blocks. Each nucleotide is composed of a phosphate residue, a deoxyribose sugar (pentose), and a nucleobase (adenine, guanine, cytosine, or thymine). The phospho-deoxyribose subunits are identical at each nucleotide and form the backbone of a DNA strand along which the sequence of nucleobases may vary. The DNA in living cells is typically composed of two complementary strands, which are connected through hydrogen bonds between complementary nucleobases. The figure shows a hypothetical sequence of four base pairs, where hydrogen bonds (*dotted* lines) can only be formed between nucleobases as indicated.

It is the sequence of base pairs along a chromosome which stores and thereby constitutes "genetic information". The expression of information typically occurs at a *gene* coding region of a chromosome. A gene is an organised structure of DNA elements, which can be divided into regulatory sequence regions and protein-coding regions (*exons*) that can be separated by non-coding DNA segments (*introns*). The sequence of basepairs instructs the *transcription* from double-stranded DNA into single-stranded ribonucleic acid (RNA) and the *translation* into proteins. Regulation of gene expression directs cell growth and maintenance, as well as the development of an organism and its ability to interact with and react to the environment.

The sum of observable characteristics is referred to as the *phenotype* of an individual. The expression of phenotypic traits varies among the members of a population due to genetic variation as well as environmental influences. For example, traits such as blood type or eye colour are determined genetically, whereas most of the phenotypic variability seen in a population arises from interactions between genetic and environmental factors. Typical examples are the effects of diet or stress on complex traits such as body weight or health.

Note that the meaning of the word *gene* has changed over time (*e.g.* see Slack, 2014). Historically, before the molecular basis of DNA was discovered, a gene was informally defined as the smallest unit of heredity, referring to the determinant of a characteristic that is transmitted from parent to offspring. A gene may be observed in different variant forms in the population, each distinguished as an *allele*. Further, a *locus* (plural *loci*) refers to the physical location of a gene on a chromosome, but may also be used in reference to the position of a single nucleotide (or *site*) in the genome. When a set of sites on a single chromosome is considered, *i.e.* the alleles observed at one or more loci, the term *haplotype* is used. While one *maternal* and one *paternal* haplotype can be distinguished in a diploid individual, its *genotype* refers to the sum of the inherited genetic information at one or more loci in the two chromosomes. An individual can be *homozygous* for a particular allele at a given site if the allele is identical in both parents, or *heterozygous* if the inherited alleles differ. This terminology is further clarified in Figure 1.2 (next page).

The following sections describe the main processes which generate genetic variation and, thereby, phenotypic variation in a population; namely mutation (Section 1.2.1, next page) and recombination (Section 1.2.2, page 9).

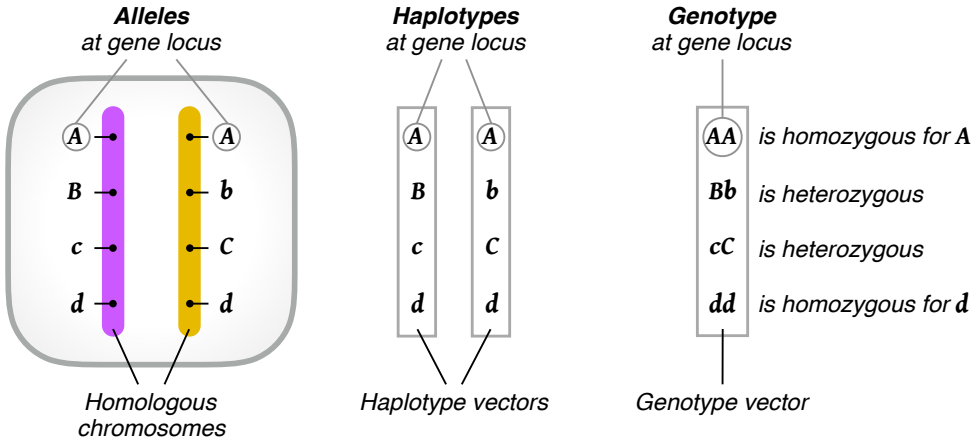


Figure 1.2: Alleles, haplotypes, and genotypes. A pair of homologous chromosomes is shown (*left*) on which four gene loci are highlighted; labelled as A, B, C, and D. Maternal and paternal chromosomes are shown in *purple* and *yellow* (arbitrarily coloured). Each gene may have two allelic states (in this example), distinguished by capitalisation of the label. Each chromosome has a corresponding haplotype at each locus (*middle*). Genotypes do not distinguish chromosomes and are represented as the sum of allelic information inherited from both parents (*right*). Note that the term *haplotype* may refer to the allelic state observed at a single nucleotide or a set of alleles observed along a chromosome. Likewise, the term *genotype* may refer to the allelic dosage at a single site or a vector of observed genotypic information.

1.2.1 Mutation

A mutation constitutes a lasting change in the genetic sequence, *e.g.* caused by imperfect DNA replication during cell division or due to errors in the DNA repair process. The change may initially be only present in one cell, but it is passed on to daughter cells in the course of successive cell divisions (*mitosis*). If mutations occur in the germline, *i.e.* germ cells which give rise to haploid *gametes* (sperm and egg cells) during *meiosis*, the nucleotide sequence is permanently altered in all cells of the progeny. If a mutation has no effect on the reproductive success of an individual, it is said to be selectively *neutral*; otherwise, a mutation may lead to a selective advantage or disadvantage, *e.g.* due to a *beneficial* or *deleterious* effect on the phenotype, respectively. In humans, the average rate of mutation per site and per generation, denoted by μ, is typically as low as one mutation event every 100 million base pairs. More specifically, recent studies suggest a mutation rate of $\mu \approx 1.1 \times 10^{-8}$ (Roach *et al.*, 2010) or $\mu \approx 1.2 \times 10^{-8}$ (Scally and Durbin, 2012).

Mutations generate the genetic variation that is observable in a population; several classes of genetic *variants* can be distinguished (*e.g.* see Frazer *et al.*, 2009). A change at a single position on the chromosome results from a *substitution* of one base for another, which in sample data is observed as a single-nucleotide polymorphism (SNP). Nucleotides may be added to or removed from the sequence, due to *insertions* or *deletions* respectively,

commonly referred to as *indels*. Larger changes to the chromosomal structure may also be distinguished. This thesis is mainly concerned with genetic variation observed at individual positions in the genome. In the following, the term "mutation" is used in reference to substitutions at single loci that result in observable SNPs in sample data. It is further assumed that SNP loci are *biallelic*, *i.e.* there are two alleles that segregate in a population (sample) at a given locus; this is the case for the vast majority of SNPs.

1.2.2 Recombination

Recombination refers to the reorganisation of alleles during meiosis in sexually reproducing organisms, which is facilitated through the physical exchange of genetic material between maternal and paternal chromosomes, such that new combinations of alleles are generated and transmitted to the offspring. Two main mechanisms of recombination can be distinguished.

Chromosomal crossover refers to the overlap of two chromatids (replicated maternal and paternal chromosomes) with subsequent, mutual exchange of homologous DNA segments.

Gene conversion is a non-reciprocal exchange of genetic material. The DNA sequence at a section in one of the chromatids is replaced by a copy of the sequence on the other chromatid, resulting in the loss of its original sequence.

Here, chromosomal crossover is implied as the acting mechanism of recombination, whereas gene conversion is not considered in this thesis. In the following, the term *recombination* therefore refers to crossover events between two homologous chromosomes.

Consider the haplotypes at two loci in an individual which is heterozygous for both the alleles at these loci. Given gene locus $\mathcal{A}$ with alleles A and a, and locus $\mathcal{B}$ with alleles B and b, the observed allelic configurations are (A, B) on one of the chromosomes and (a, b) on the other. If no recombination occurs between the two loci during meiosis, the resulting gametes retain the configuration as present in the parental chromosomes; *i.e.* the offspring may either receive (A, B) or (a, b). In presence of recombination, in particular if the number of recombination events between loci is odd, the association between the two loci is broken such that either (A, b) or (a, B) are transmitted to the offspring. An even number of recombination events between the two loci reverts the configuration of alleles. Both cases (odd and even numbers of recombination events) are illustrated in Figure 1.3 (next page).

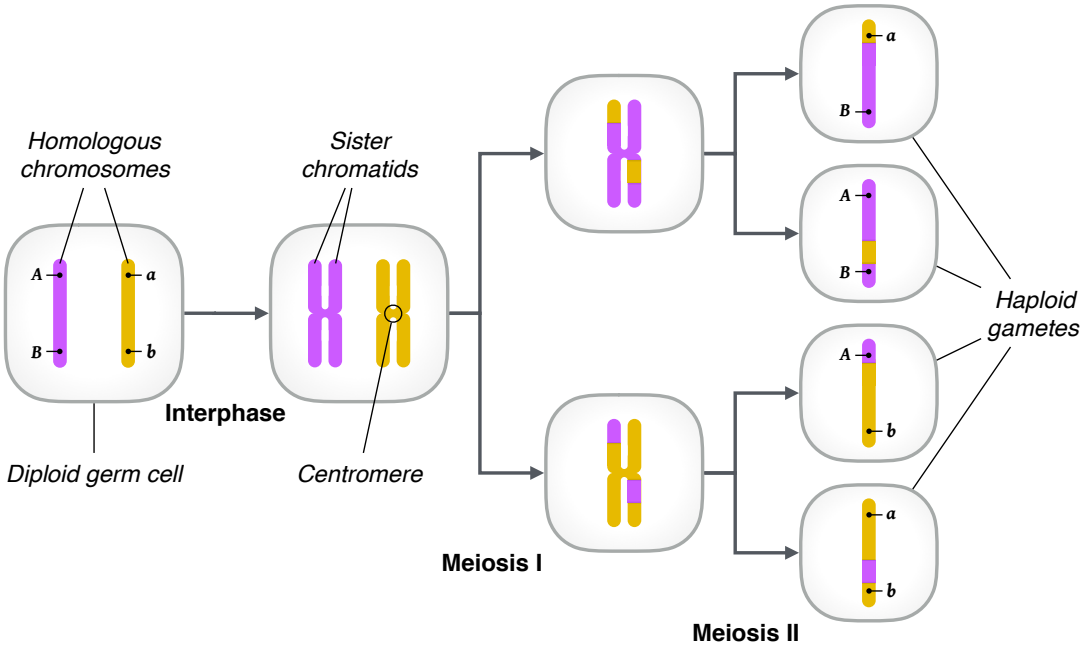


Figure 1.3: Illustration of recombination during meiosis. One pair of homologous chromosomes is shown at the beginning of the meiotic cell cycle (*left*). Maternal and paternal chromosomes are shown in *purple* and *yellow* (arbitrarily coloured). The allelic configuration at two sites is indicated on both chromosomes; (*A*, *B*) and (*a*, *b*). DNA sequences are replicated during the *Interphase* of meiosis, where each chromosome forms two identical *sister chromatids* which are held together at the *centromere*. Homologous chromosomes are paired at the beginning of the first cell division (*Meiosis I*), during which sequence segments are exchanged between chromatids through crossover. In the second cell division (*Meiosis II*), the four chromatids are then separated into haploid gametes (*right*).

1.2.2.1 Genetic linkage

A direct consequence of meiotic recombination is the phenomenon of genetic linkage, which was discovered by Morgan (1911) in experiments on *Drosophila*. Linkage describes the concept that genetic markers located in close proximity to each other are less likely to be separated by recombination during meiosis. This concept was further developed by Sturtevant (1913), who proposed that the frequency of recombination between a set of markers can be used to determine the linear order of genes on a chromosome. It was this idea that paved the way for the development of molecular and statistical methods for the purpose of *linkage analysis*, through which it became possible, for example, to detect the chromosomal location of genetic variants implicated in human disease.

The earliest models of recombination go back to Haldane (1919), who defined *genetic distance* as the expected number of recombination events per meiosis between two loci. The unit of genetic distance is called a *Morgan*. However, it is more common to express genetic distance in units of centiMorgan (cM), where 1 Morgan is equal to 100 cM. For example,

if two loci sit 1 cM apart on a chromosome, the expected number of recombination events between them is 0.01 per generation, meaning that the two loci are separated once every 100 meioses on average. In humans, a distance of 1 cM corresponds to about 1 million base pairs; *i.e.* 1 Megabase (Mb). The genetic distance translates into the rate of recombination, here denoted by ρ. The human genome exhibits an average rate of $\rho \approx 1 \times 10^{-8}$ per site per generation. However, the recombination rate varies among chromosomes and more so along the length of each chromosome.

1.3 Models in population genetics

Over the last century, the field of population genetics has evolved from a mainly theoretical area of study into a more applied area of research. More recently, the field has adapted to the exponential growth of available molecular data and continues to fill a niche in the computational sciences so as to be able to analyse the increasing amounts of data and to answer questions of biological as well as medical meaning. This section outlines the statistical concepts on which many of the current analytical approaches are based. Coalescent theory is of particular importance for the understanding of the statistical methods developed in this thesis, for which the Wright-Fisher model may serve as an introduction.

1.3.1 Wright-Fisher model

One of the most influential models in population genetics is the Wright-Fisher model of reproduction (Fisher, 1930; Wright, 1931), which describes how gene frequencies evolve over time in a finite population. Because the Wright-Fisher model is often implied in other statistical applications in population genetics, it is pertinent to explore its properties in greater detail. In particular, the following describes the effects of "random genetic drift" in an idealised population.

In its simplest form, the Wright-Fisher model considers a gene locus at which two alleles, A and a, are observed; *i.e.* the locus is *biallelic*. A population of N haploid individuals is assumed, where N remains constant in each generation. All individuals die at the same time at which all individuals in the next generation are born; *i.e.* time is measured in discrete, non-overlapping generations. The effects of mutation or selection are ignored, such that alleles are *neutral* and the probability of producing offspring is equal for each individual. It follows that reproduction is considered as a random

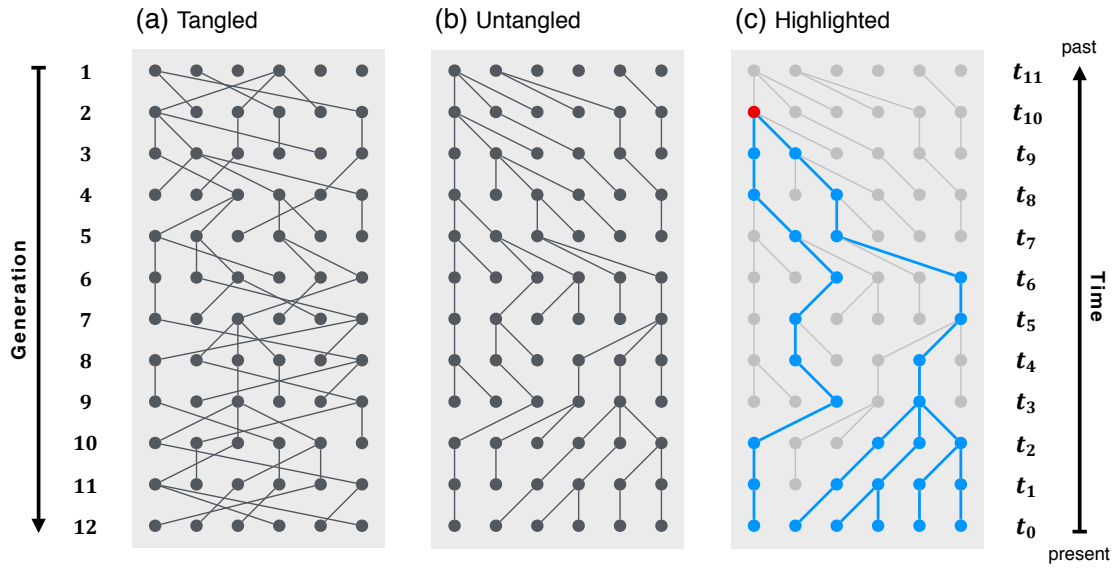


Figure 1.4: Example genealogy in a Wright-Fisher model. A population of size $N = 6$ is shown in Panel **(a)**, which is observed over 12 generations. In the neutral Wright–Fisher model, one individual is chosen at random (with replacement) in each generation to produce offspring for the next generation, repeated N times. The genealogy of the population is more clearly seen after individuals have been sorted such that their lineages do not cross; see Panel **(b)**. Note that not every individual produces offspring, such that some lineages go extinct. If this process is repeated over many generations (forward in time), it can be seen that all individuals in the present generation derive from a single individual in the past, which is indicated in Panel **(c)**. The ancestry of the present population (*blue*) is traced back to a single ancestor (*red*) at time $t = 10$ generations ago.

sampling process, in which the alleles that are transmitted into the next generation are drawn (with replacement) from the gene pool of the current population. An example is illustrated in Figure 1.4 (this page).

Since each draw has only two possible outcomes, A or a, each generation is produced by a series of independent Bernoulli trials such that allele frequencies are binomially distributed. Let X_t denote the number of A alleles in generation t. Given $X_t = i$ allele copies (or individuals which carry the allele), the probability of drawing the A allele is equal to its frequency in the current generation, denoted by $\pi_i = {}^{i}/_{N}$. The probability of observing $X_{t+1} = j$ copies in the next generation is

$$P(j \mid i) \;=\; \binom{N}{j} \pi_i^j \, (1 - \pi_i)^{N-j} \tag{1.1}$$

for $0 \leq i, j \leq N$, and where $\sum_{j=0}^{N} P(j \mid i) j = i$. From the binomial distribution follows that the expected number of alleles in the next generation can be expressed as

$$\mathbb{E}[X_{t+1} \mid X_t] \;=\; N\pi_i \;=\; N\frac{X_t}{N} \;=\; X_t \tag{1.2}$$

and the variance is given by

$$\mathrm{Var}[X_{t+1} \mid X_t] \;=\; N\pi_i(1-\pi_i) \;=\; X_t\left(1-\frac{X_t}{N}\right). \tag{1.3}$$

Equation (1.2) implies that $\mathbb{E}[X_t] = \mathbb{E}[X_{t-1}]$ and thereby $\mathbb{E}[X_t] = \mathbb{E}[X_0]$; *i.e.* the expected number of alleles in each generation is (on average) equal to the initial allele count. This result is reminiscent of the Hardy-Weinberg principle (Hardy, 1908; Weinberg, 1908), which states that the relative allele frequency remains constant in each generation if mating is random, but in which the population size is assumed to be infinite. However, due to the behaviour of a stochastic process in a finite population, the number of allele copies may eventually *drift* to 0 or N, even in a single generation. Several examples of how the allele frequency may change in populations of different sizes are shown in Figure 1.5 (this page).

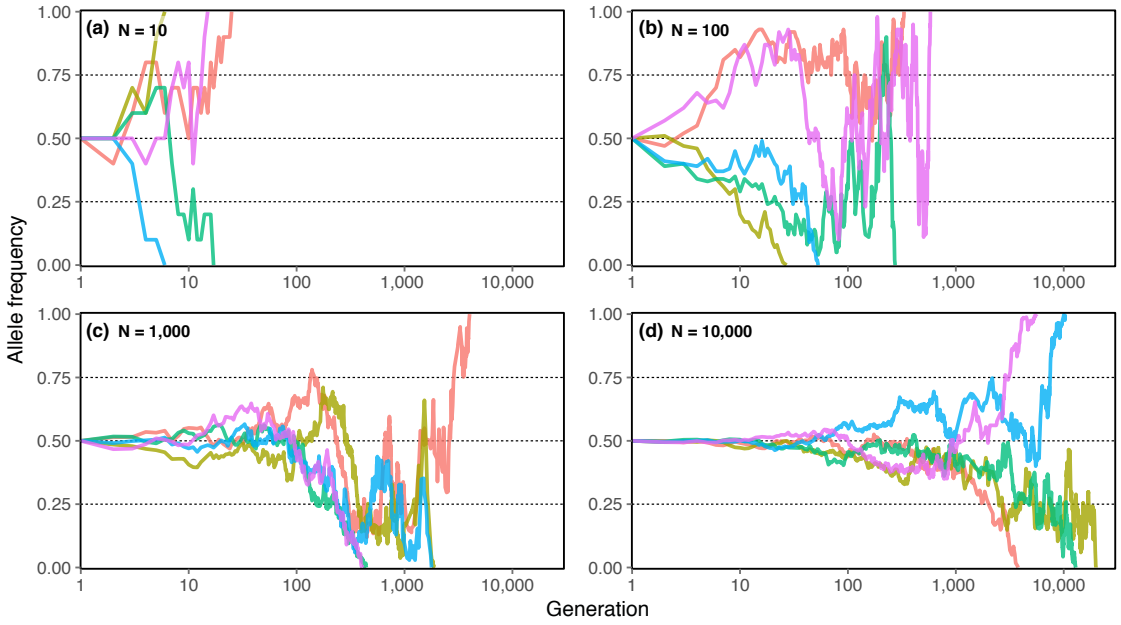


Figure 1.5: Allele frequency changes over time simulated under the Wright-Fisher model. A haploid population was simulated under four different constant values of population size, N, as indicated in each panel. The change in allele frequency is shown by generation. For each value of N, five replicate simulations were conducted (distinguished by colour). Note that the allele frequency does not change after it has reached 0 or 1; *i.e.* the allele is said to have become *fixed* in the population.

Because the frequency of an allele in a particular generation only depends on the frequency distribution in the previous generation, it follows from this property that the reproductive process is itself a Markov chain, with transition probabilities as described by Equation (1.1) and a state space in $\{0, \ldots, N\}$. The states 0 and N are absorbing,

which means that if the population consists of $X_t = 0$ or $X_t = N$ alleles, it remains so in all future generations. A consequence of this Markov process is that an allele will either go extinct or reach *fixation* (*e.g.*, see Ewens, 2012). Let the time until either of the two alleles has reached fixation be denoted by T. From Equation (1.2) follows that the probability that an allele reaches fixation is

$$P(X_T \in \{0, N\}) = \frac{X_0}{N} \tag{1.4}$$

which means that the probability of a given allele reaching fixation is equal to its initial frequency.

Without the introduction of new alleles through mutation, the Wright-Fisher model predicts that genetic variation is inevitably lost over time, due to random drift resulting from sampling error in a finite population. Hence, an important extension of the Wright-Fisher model is the incorporation of mutations. Suppose that allele A mutates into allele a with rate μ_A, and a into A with rate μ_a. The transition probability given in Equation (1.1) still holds, but allele frequency can be expressed such that π_i is dependent on mutation rate, namely

$$\pi_i \;=\; \frac{i}{N}\,(1-\mu_A) \;+\; \left(1-\frac{i}{N}\right)\mu_a\,. \tag{1.5}$$

If $\mu_A, \mu_a > 0$, then transitions from any state into any other state remain possible in each generation and permanent fixation is avoided. Note that in a population in which the effects of mutation and genetic drift are in statistical equilibrium allele frequencies are expected to follow the Hardy-Weinberg principle; *i.e.* the population is in Hardy-Weinberg equilibrium (HWE).

1.3.2 Coalescent theory

The coalescent is arguably the most frequently employed genealogical method in population genetics. The concept and the statistical properties of the coalescent were first described by Kingman (1982a,b,c) and it is therefore often referred to as "Kingman's coalescent". The term "n-coalescent" is also frequently used to emphasise the importance of the sample size, n, in the genealogical process within a much larger population. The coalescent, at its core, is a collection of stochastic models which provide the means to generate predictions about population dynamics under a variety of models of genetic variation and demography (Wakeley, 2008). Note that the term "prediction" may sound odd given that the coalescent looks backward in time to reconstruct a possible

genealogy given a set of population parameters. The coalescent is often used to simulate the ancestry of a sample, from which particular model parameters can be inferred, for example, on basis of biological observations. The first computational algorithm for simulations under the coalescent (named "`ms`") was devised by Hudson (1990). Over the past decades, coalescent theory has grown extensively. Hence, this section provides only a summary of the basic properties of the coalescent as relevant for this thesis. For a more thorough presentation of the subject see, for example, Fu and Li (1999), Neuhauser (2001), Nordborg (2001), Hein *et al.* (2004), and Wakeley (2008).

In contrast to the Wright-Fisher model, as well as other approaches which model the genealogical history of a population forward in time, the coalescent process reconstructs the genealogy of a sample by tracing the ancestry of individuals (or genes) backward in time. Ancestral relationships between individuals are represented as lineages in a genealogical tree. In each generation, each individual independently chooses one ancestor at random. If two individuals choose the same ancestor by chance, their lineages are joined; *i.e.* they *coalesce*. The time at which two lineages join is referred to as a *coalescent event*. This process is repeated until only one lineage is left, which belongs to the most recent common ancestor (MRCA) of the sample.

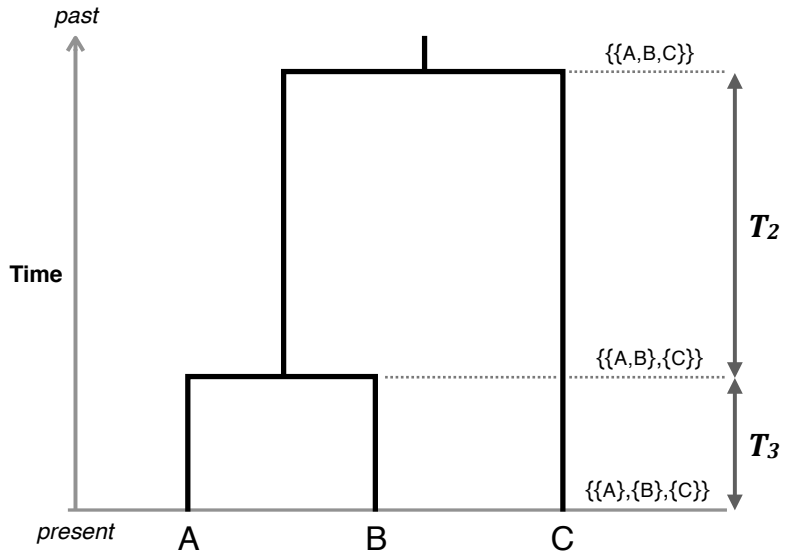


Figure 1.6: Topology of a genealogical tree in the coalescent. The genealogical relationship of three haploid individuals is shown, A, B, and C, which represent separate lineages at present, but where A and B are the first to coalesce (back in time). The waiting time between successive coalescent events is denoted by T_i, where i is the number of ancestral lineages at a given time interval, which changes from i to $i-1$ at coalescence. Figure modified from Nordborg (2001).

The history of a sample is reflected in its genealogy and can be described in terms of the topology of the tree and the lengths of the connecting branches. The branch length corresponds to the time interval between two successive coalescent events, which is of central interest in describing the coalescent process. Let this waiting time be denoted by T_i, where i corresponds to the number of distinct lineages during the time interval, which changes from i to $i-1$ at coalescence. An example of a simple genealogical tree is shown in Figure 1.6 (page 15), in which the waiting times between coalescent events are indicated. In the following, the concept of the standard coalescent is described by assuming a haploid population of constant size, N, in which the effects of mutation, selection, recombination, or other biological processes are not involved.

For now, consider a sample of $n = 2$ individuals taken at the present time, which are followed back in time until the first coalescent event. Since there are N possible ancestors, the probability that a particular ancestor is chosen by one of the individuals is equal to N^{-1}. The probability that two individuals choose the same ancestor independently is N^{-2}. Hence, the probability that any of the possible ancestors is chosen by two individuals is equal to $N \times N^{-2} = N^{-1}$, and the probability that none is chosen is $1 - N^{-1}$. To arrive at the probability that two lineages coalesce $t > 0$ generations back in time, it is implied that they do not choose the same ancestor in previous generations. Because generations are independent, the probability that the two lineages are distinct over $t-1$ generations is

$$P(T_2 \geq t \mid N) \;=\; \left(1 - \frac{1}{N}\right)^{t-1} . \tag{1.6}$$

Therefore, the probability that two lineages coalesce t generations back in time is geometrically distributed with rate N^{-1}, such that

$$P(T_2 = t \mid N) \;=\; \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N} \tag{1.7}$$

which arises from the number of independent Bernoulli trials needed until the same ancestor is chosen by two lineages. It follows from the geometric distribution that the expected number of generations up to and including the coalescent event is

$$\mathbb{E}[T_2 \mid N] \;=\; \frac{1}{N^{-1}} \;=\; N \tag{1.8}$$

and the variance is

$$\mathrm{Var}[T_2 \mid N] \;=\; \frac{1 - N^{-1}}{N^{-2}} \;=\; N^2 \left(1 - \frac{1}{N}\right). \tag{1.9}$$

A notable result is that the expected time to the first coalescent event is equal to the size of the population; see Equation (1.8). It is therefore convenient to scale time in units of N generations, namely

$$\tau \;=\; \frac{t}{N} \tag{1.10}$$

where the time, τ, is continuous (as opposed to time measured in distinct generations) and referred to as the *population-scaled* time. The probability that a pair of lineages remains distinct during a given time interval can now be approximated using the exponential distribution if the population size is sufficiently large, *i.e.* as N tends to infinity; namely

$$P(T_2 > \tau \mid N) \;=\; \left(1 - \frac{1}{N}\right)^{\lfloor N\tau \rfloor} \quad \xrightarrow{\quad N \to \infty \quad} \quad \mathrm{e}^{-\tau} \tag{1.11}$$

where $\lfloor N\tau \rfloor$ is the largest integer that does not exceed $N\tau$ (*e.g.*, see Nordborg, 2001).

The above can now be extended to consider a sample of $n \geq 2$ individuals. Let i denote the number of distinct lineages in the current generation. In the immediately previous generation, there are i ancestral lineages if no coalescent event has occurred, or $i - 1$ otherwise. The probability of no coalescence in the previous generation can be derived by letting each lineage choose a different ancestor. Let the first lineage choose among N ancestors with probability $N/N = 1$, the second lineage then chooses among the remaining $N - 1$ ancestors with probability $(N-1)/N$, the third chooses among $N - 2$ ancestors with probability $(N-2)/N$, and so on. Given i lineages in the current generation, the probability that they also have i ancestors in the immediately previous generation therefore is

$$\begin{aligned} P_{i,i}(N) \;&=\; \left(\frac{N}{N}\right)\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N}\right)\cdots\left(\frac{N-(i-1)}{N}\right) \\ &=\; \left(\frac{N}{N}\right)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{i-1}{N}\right) \\ &=\; \prod_{k=1}^{i-1}\left(1-\frac{k}{N}\right) \\ &=\; 1 - \frac{\sum_{k=1}^{i-1} k}{N} + \mathcal{O}\left(\frac{1}{N}\right) \quad = \quad 1 - \binom{i}{2}\frac{1}{N} + \mathcal{O}\left(\frac{1}{N}\right) \end{aligned} \tag{1.12}$$

where the binomial coefficient, $\binom{i}{2} = \sum_{k=1}^{i-1} k$, corresponds to the number of possible pairs.

Similarly, to derive the probability of a coalescent event in the immediately previous generation, it is implied that i lineages have $i - 1$ ancestors. Let any two lineages choose the same ancestor with probability $\binom{i}{2}\frac{1}{N}$ while the remaining lineages choose a different

ancestor. It follows that

$$\begin{aligned} P_{i,i-1}(N) &= \binom{i}{2}\frac{1}{N} \times \left(\frac{N-1}{N}\right)\left(\frac{N-2}{N}\right)\cdots\left(\frac{N-(i-2)}{N}\right) \\ &= \binom{i}{2}\frac{1}{N} \times \prod_{k=1}^{i-2}\left(1-\frac{k}{N}\right) \\ &= \binom{i}{2}\frac{1}{N} + \mathcal{O}\left(\frac{1}{N}\right). \end{aligned} \tag{1.13}$$

Note that the term $\mathcal{O}(N^{-1})$ describes the limiting behaviour of Equations (1.12) and (1.13) and captures all terms that decrease more rapidly than $1/N$ as N tends to infinity. Mathematically, $\mathcal{O}(N^{-1})$ corresponds to the *diffusion* limit of the continuous process, which can be ignored if the population size is sufficiently large (*e.g.*, see Wakeley, 2008). By doing so, it is assumed that not more than two lineages coalesce at a given time and that the resulting tree has a binary topology. Hence, in the limit and if $i \ll N$, the probability of no coalescence (1.12) and the probability of coalescence (1.13) in the immediately previous generation, respectively, can be written as

$$P_{i,i}(N) \approx 1 - \binom{i}{2}\frac{1}{N} \qquad P_{i,i-1}(N) \approx \binom{i}{2}\frac{1}{N}.$$

Using the above result, it follows that the waiting time until a coalescent event can again be approximated in terms of the exponential distribution as given below.

$$P(T_i > \tau \mid N) \approx \left(1 - \binom{i}{2}\frac{1}{N}\right)^{\lfloor N\tau \rfloor} \xrightarrow{N \to \infty} \mathrm{e}^{-\binom{i}{2}\tau} \tag{1.14}$$

Thus, in the continuous-time coalescent, the approximate waiting time between successive coalescent events, T_i, is exponentially distributed with rate $\binom{i}{2}$, from which follows that the expected value is

$$\mathbb{E}[T_i] = \frac{1}{\binom{i}{2}} = \frac{2}{i(i-1)} \tag{1.15}$$

and the variance is

$$\mathrm{Var}[T_i] = \frac{1}{\binom{i}{2}^2} = \frac{4}{i^2(i-1)^2}. \tag{1.16}$$

An important result of the coalescent is that an expectation for the time to the most recent common ancestor ($\mathrm{T_{MRCA}}$) can be derived dependent on the sample size, n. Given the sum of branch lengths that need to be traced back to arrive at the MRCA,

$$T_{\mathrm{MRCA}} = T_n + T_{n-1} + \cdots + T_2$$

the expectation can be expressed as

$$\mathbb{E}[T_{\text{MRCA}} \mid n] \;=\; \sum_{i=2}^{n} \mathbb{E}[T_i] \;=\; \sum_{i=2}^{n} \frac{2}{i(i-1)} \;=\; 2\left(1 - \frac{1}{n}\right). \tag{1.17}$$

Therefore, if $n \ll N$, $\mathbb{E}[T_{\text{MRCA}}] \approx 2$ in units of population-scaled time, which implies that on average the number of generations until the entire sample has coalesced into a single ancestral lineage is equal to about twice the population size; *i.e.* $2N$.

1.3.2.1 Effective population size

Natural populations rarely adhere to the assumptions made by mathematical models. One such example is the rather unrealistic assumption that the population size remains constant over time. The rate at which coalescent events occur in the genealogy of a sample is conditional on the size of the population in each generation, which in reality is often highly variable. Statistical models in population genetics therefore resort to the concept of an effective population size, denoted by N_e, to substitute the census population size, N.

The effective population size is one of the central concepts of population genetics, which was introduced by Wright (1931) and further developed by many others; *e.g.* Crow and Kimura (1970). The value of N_e is commonly defined as the number of individuals in an "ideal" Wright-Fisher population (Fisher, 1930; Wright, 1931) which shows the same value of a given genetic property of interest as seen in the non-ideal, natural population (Ewens, 2012).

Although the effective size is of crucial interest in population genetics, the complexity of N_e is difficult to define and estimate, because it collapses a large number of variable stochastic factors into a single parameter. Properties such as heterozygosity and allele frequency in a population of finite size will fluctuate over time, due to the stochastic sampling process of a finite number of gametes, which is influenced by variable factors such as sex ratio, number of breeding individuals, and variance in reproductive success. Several definitions of N_e have been developed, for example dependent on the rate of increase in homozygosity (inbreeding effective size) or the variance in allele frequency change from one generation to the next (variance effective size), but which can only predict different aspects of the underlying population history; see review by Wang (2005).

Note that N_e may differ from the census size of a population by several magnitudes. For example, the human population currently counts several billion individuals globally, whereas the long-term, diploid effective size is commonly defined in the order of $N_e \approx 10{,}000$; *e.g.* based on estimates from DNA polymorphism data (*e.g.* Takahata, 1993; Yu *et al.*, 2001).

For consistency with the definitions provided so far, the following sections in this chapter keep N to denote the population size, but this is substituted by N_e in the remaining chapters.

1.3.2.2 The coalescent with mutation

Mutations are essential to generate genetic diversity and maintain genetic variation in a finite population. The standard coalescent relies on the assumption that variant alleles are selectively neutral; *i.e.* the effect of mutation is independent of the genealogical process. As such, mutation events can be superimposed on the coalescent tree by placing mutations on all branches proportional to their length. An example is illustrated in Figure 1.7 (this page), in which several mutation events are shown to give rise to the variation observed in the DNA sequence of a sample.

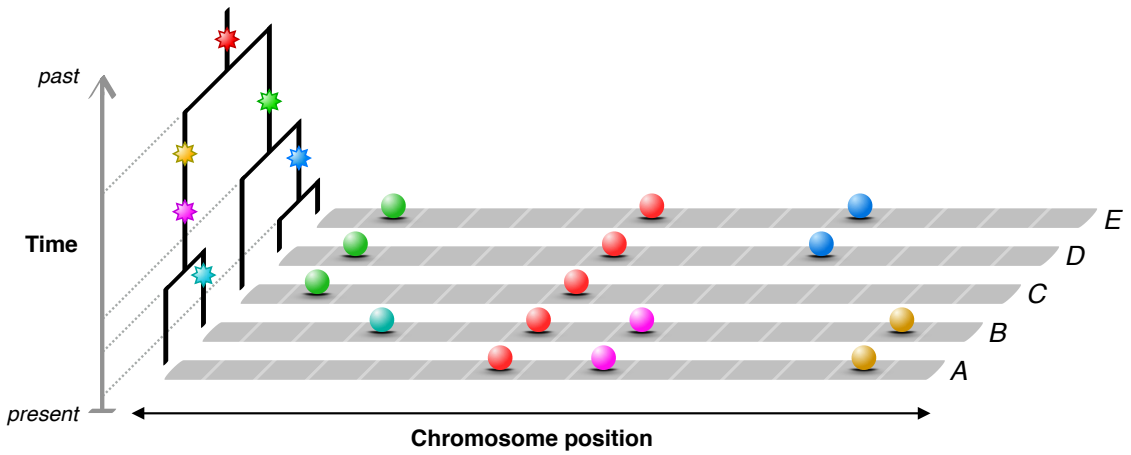


Figure 1.7: Mutation events on a genealogical tree in the coalescent. The genealogy of a sample of five haploid individuals ($A - E$) is shown on the left. The time of each coalescent event is indicated by a *dotted* line. Mutation events (*stars*) are placed along the branches of the tree. Each mutation event alters the allelic state at a random position on the chromosome, giving rise to a new allele, which is inherited by all descendants of the ancestral individual in which the mutation occurred. Horizontal lanes (*grey*) represent the chromosome sequence of the individuals, on which the derived alleles are depicted as *marbles*; colours correspond to the mutation event from which the alleles derive.

Given a constant rate of mutation per site per generation, μ, the expected number of mutations on a branch in the genealogical tree, *i.e.* a lineage that is t generations long, is $t\mu$. If time is scaled in units of N generations, see Equation (1.10) on page 17, the corresponding value is expressed by $\tau N\mu$, such that the rate of mutation per site per

unit of time is equal to $N\mu$. However, for historical reasons (*e.g.*, see Wakeley, 2008), the population-scaled mutation rate is given by the compound parameter

$$\theta = 2N\mu \tag{1.18}$$

where θ is assumed to be constant in the limit $N \to \infty$. Note that the factor of 2 relates to the formulation of the expected number of pairwise differences between two haploid sequences, which is equal to θ (Tajima, 1993). Thus, θ describes the amount of genetic diversity in a population.

Because mutations effectively count events that occur independently, the probability distribution of mutation is described by a Poisson process with rate parameter $\theta/2$ (Wakeley, 2008). It follows that the probability of observing K mutations on a branch of length L is itself Poisson distributed with parameter $\theta L/2$;

$$P(K = k \mid L) = \left(\frac{\theta L}{2}\right)^k \frac{1}{k!} \mathrm{e}^{-\frac{\theta L}{2}} \tag{1.19}$$

where $L = t$ if measured in discrete generations or $L = N\tau$ if measured on a continuous time scale. It follows from the Poisson distribution that $\mathbb{E}[K \mid L] = \mathrm{Var}[K \mid L] = \theta L/2$.

Suppose that each mutation event creates a new allele and that each site can only mutate once in the history of the sample; such a setting is generally referred to as the infinite sites model (Kimura, 1969; Watterson, 1975). Under this assumption, the number of segregating sites (or *variant* sites) observed in sequence data in a sample of size n, is equal to the sum of mutation events that occurred in the history of the sample. The total branch length of the tree thereby determines the expected value of the number of segregating sites, denoted by S_n. From the sum of all branch lengths, *i.e.*

$$T_{\text{total}} = iT_i + (i-1)T_{i-1} + (i-2)T_{i-2} + \ldots + 2T_2$$

where i is the number of distinct lineages during a given time interval, the expected value of the total branch length can be computed as

$$\mathbb{E}[T_{\text{total}} \mid n] = \sum_{i=2}^{n} i\, \mathbb{E}[T_i] = \sum_{i=2}^{n} i\, \frac{2}{i(i-1)} \tag{1.20}$$

where $\mathbb{E}[T_i]$ is given by Equation (1.15) on page 18. From the above, the expected value of S_n can be derived as follows.

$$\mathbb{E}[S_n] = \frac{\theta}{2}\, \mathbb{E}[T_{\text{total}}] = \frac{\theta}{2} \sum_{i=2}^{n} i\, \frac{2}{i(i-1)} = \theta \sum_{i=1}^{n-1} \frac{1}{i} \tag{1.21}$$

By rearrangement, the following equation can be obtained;

$$\hat{\theta}_W = \frac{S_n}{\sum_{i=1}^{n-1} \frac{1}{i}} \tag{1.22}$$

which is an unbiased estimator of the genetic diversity in a sample of sequence data; also known as Watterson's θ (Watterson, 1975). With regard to the calculation of the effective population size as described in the previous section (page 19), it can be seen that an estimate for the value of N_e can be obtained, for example, from Equations (1.18) and (1.22) given an estimate of the mutation rate.

1.3.2.3 The coalescent with recombination

Recombination is ubiquitous in nature and crucially involved in the spread of genetic variability in populations of sexually reproducing organisms. Hudson (1983) showed that the genealogical process in the coalescent can be extended to model recombination along the sequence of a sample. In contrast to neutral mutation events, which do not affect the topology of a tree under the standard coalescent, recombination events have a considerable effect on the structure of the genealogy.

Consider the sequence of one of the chromosomes present in a diploid individual. Due to recombination, different sections of the chromosome can be traced back to the ancestral material in two parents in the immediately previous generation, and further to four grandparents in the second previous generation, and so on. It becomes clear that the ancestral origin of the chromosomal sequence is distributed over many parallel lineages back in time. For example, a useful (but limited) representation of this process is seen in family trees (*pedigrees*) in which ancestral lineages *branch* back in time such that the number of ancestors appears to double in each generation. Obviously, this progression cannot go on indefinitely because in a finite population any individual will be to some degree related to any other individual (their pedigrees may partially share the same ancestors). As shown by Wiuf and Hein (1997), all chromosomal lineages will eventually coalesce back onto a single lineage which is the *ultimate* MRCA of the chromosomal sequence.

The coalescent with recombination includes coalescent events as well as branching events, but where the genealogy of a sample of sequences cannot be represented by a single tree. This is because recombination alters the genealogical relation between different segments of the ancestral material such that two chromosomes may be closely related at a particular segment, but distantly related at another segment. The chromosomal

sequence is superimposed by a sequence of *marginal* trees of different topology. This tree sequence can be represented in a graph structure. The most common way to represent the genealogy of a sample of sequences is the ancestral recombination graph (ARG) which was first described by Griffiths (1991) in a two-locus model, but which was later generalised by Griffiths and Marjoram (1996, 1997b) in regards to the infinite sites model. Figure 1.8 (this page) illustrates a minimal example of an ARG for a sample of four chromosomes, in which mutation events are included to emphasise the pattern of allelic variation resulting from recombination between two loci. In the following, the basic properties of the generalised ARG are presented.

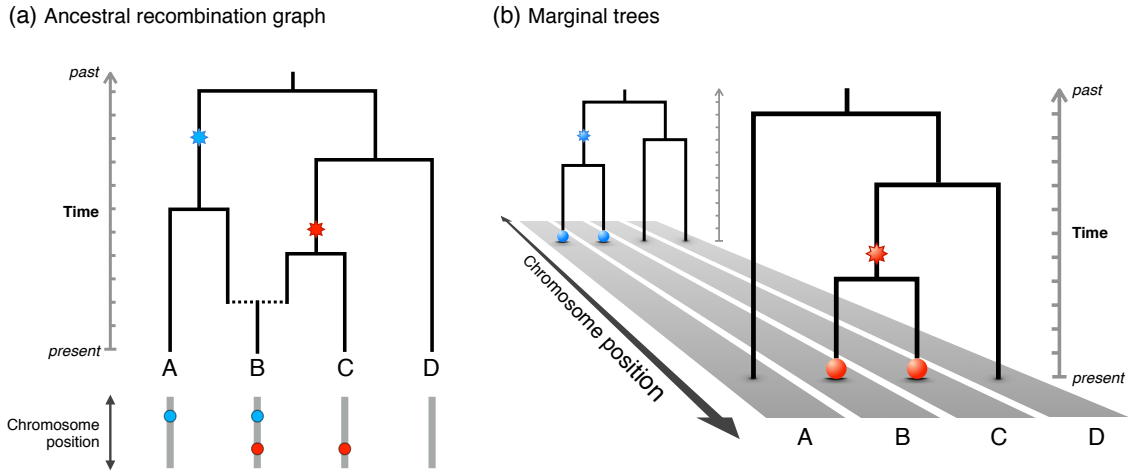


Figure 1.8: Illustration of the ancestral recombination graph. Panel **(a)** shows the ARG for a sample of four chromosomes, labelled by A, B, C, and D. The *dotted* horizontal line denotes the time of a recombination event between chromosomal lineages. Mutation events are shown as *stars*. The chromosomal positions of derived alleles are indicated below the ARG. The corresponding marginal trees are shown in Panel **(b)**, where each lane (*grey*) represents the chromosomal sequence on which the derived alleles sit (shown as *marbles*).

Given the rate of recombination per site per generation, ρ, the population-scaled recombination rate is given by the compound parameter

$$\phi = 4N\rho \tag{1.23}$$

which is assumed to be constant in the limit $N \to \infty$.* The factor of 4 results from time being scaled in units of $2N$ generations, accounting for the fact that the population is diploid. Note that this adjustment permeates the coalescent and implies similar changes in other equations. For example, the scaled mutation rate given in Equation (1.18) on page 21 needs to be written as $\theta = 4N\mu$ if considered in a diploid population.

* Note that in the literature r is often used to denote the per-generation recombination rate and ρ to denote the population-scaled recombination rate.

Given a sample of n chromosomes, the number of chromosomal lineages, i, may increase (due to recombination) or decrease (due to coalescence) back in time. First, consider the event of no recombination and no coalescence; *i.e.* the value of i remains the same in the previous generation (*e.g.*, see Tavaré, 2004). The probability of this event is

$$(1-\rho)^i \times \left(\frac{N-1}{N}\right)\left(\frac{N-2}{N}\right)\cdots\left(\frac{N-(i-1)}{N}\right) \tag{1.24}$$

where $(1-\rho)^i$ corresponds to the probability that none of the lineages recombine; the other terms refer to the probability of no coalescence, which was already defined in Equation (1.12) on page 17. Now, because the rate at which one lineage branches into two lineages back in time is equal to $\phi/2$, Equation (1.24) can be written as

$$1 - \frac{i\phi}{2N} \; - \; 1 - \binom{i}{2}\frac{1}{N} \; + \; \mathcal{O}\left(\frac{1}{N}\right). \tag{1.25}$$

For the event $i \to i+1$, which can only be facilitated through recombination, it follows that the probability of a recombination event in the previous generation is given by

$$\frac{i\phi}{2N} \; + \; \mathcal{O}\left(\frac{1}{N}\right). \tag{1.26}$$

The term $\mathcal{O}(N^{-1})$ is the diffusion limit of the function and corresponds to the probability that more than one recombination event occurs at a given unit of time, which can be ignored for larger population sizes; *i.e.* as N tends to infinity. Similarly, a coalescent event in the previous generation means that $i \to i-1$, for which the probability has already been described in Equation (1.13) on page 18. Also, as shown in Equation (1.14) on page 18, the probability of coalescent events, in the limit $N \to \infty$, is exponentially distributed with rate

$$\binom{i}{2} \; = \; \frac{i(i-1)}{2}\,. \tag{1.27}$$

Likewise, in the limit, recombination follows the same distribution in the coalescent at rate

$$\frac{i\phi}{2}\,. \tag{1.28}$$

It follows that the coalescent with recombination can be described as a continuous-time Markov chain with a *birth-death* process. Lineages are "born" through recombination or "die" due to coalescence backward in time (*e.g.*, see Tavaré, 2004; Wakeley, 2008). The state space is delimited by $i = n$ at present and $i = 1$ at an MRCA. The transition rates can be summarised as follows.

$$i \; \to \; \begin{cases} i-1 & \text{at rate} \quad \frac{i(i-1)}{2} & \text{if lineages coalesce} \\ i+1 & \text{at rate} \quad \frac{i\phi}{2} & \text{if lineages recombine} \end{cases} \tag{1.29}$$

Importantly, because the rate of coalescence is quadratic in the number of lineages and the rate of recombination is at most linear, the number of lineages cannot increase indefinitely (Wiuf and Hein, 1997). As a result, the ancestry of all chromosomal segments are eventually traced back to a single ancestral chromosome in the ultimate MRCA.

1.4 Advances in high-throughput genomic technologies

In this section, I provide a brief review of the developments in high-throughput genomic technologies that have been achieved over the past 40 years. I further highlight some of the milestone projects that have contributed substantially to our understanding of the human genome, namely the Human Genome Project (HGP), the International HapMap Project (HapMap), and the 1000 Genomes Project (1000G). Data from HapMap and 1000G have been used extensively in this thesis. Note that a detailed presentation of the history and biochemistry of available technologies, as well as a comprehensive list of human sequencing projects, is beyond the scope of this chapter (for review, *e.g.* see Metzker, 2009; Naidoo *et al.*, 2011; Liu *et al.*, 2012; Mardis, 2017).

1.4.1 Next-generation sequencing

The first DNA-based organism to have its genome fully sequenced was the bacteriophage ΦX174 (5,386 basepairs), which was undertaken by Sanger *et al.* (1977) based on the previously developed chain-termination sequencing method (Sanger and Coulson, 1975). This technology formed the backbone of the coming era of whole-genome sequencing (WGS), which has dominated the field since 1977 and was the main method employed to sequence the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001).

Following the initialisation of the Human Genome Project (HGP) in 1990, and the publication of the draft sequence of the human genome in 2001, it was proclaimed in 2004 that the sequence of the human genome was "essentially complete" (International Human Genome Sequencing Consortium, 2004). However, it became clear that available technologies could not realistically be applied to generate sequence data for larger samples due to the significant requirements in labour, cost, and time. The National Human Genome Research Institute (NHGRI), United States, therefore announced an initiative with the aim of developing novel DNA sequencing methods (awarding more than $38 million in grants).* Ultimately, it was hoped to decrease the cost of sequencing to $1,000 or

* https://www.genome.gov/12513210/2004-release-nhgri-seeks-next-generation-of-sequencing-technologies/ [Date accessed: 2017-03-15]

less per genome (Mardis, 2006). As a result, major advances have been made in the development of commercially available sequencing and genotyping technologies, which fostered a groundbreaking synergistic relationship between research and industry, and several large-scale international projects have been initiated; see Figure 1.9 (this page).

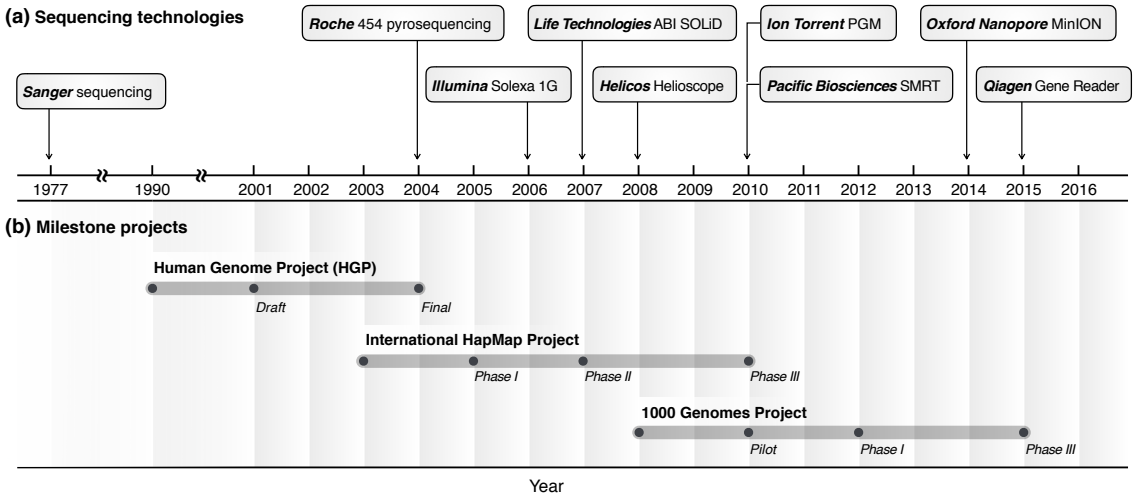


Figure 1.9: Timeline of sequencing technologies and milestone projects. Panel **(a)** shows the year of commercial introduction of successfully established next-generation sequencing (NGS) platforms until 2016, following the introduction of the Sanger *et al.* (1977) sequencing method. Panel **(b)** illustrates the timeline of three major projects that were undertaken to sequence (or genotype) the human genome. Figure modified from Mardis (2017, Figure 1) and Naidoo *et al.* (2011, Table 1).

Sanger sequencing is now regarded as the "first-generation" of sequencing technologies, while more recently developed techniques are commonly referred to as "next-generation" sequencing (NGS), which allow higher volumes of samples to be processed in shorter time and reduced cost (Metzker, 2009). The first next-generation sequencer was the *Roche GS 20 System* by *Roche 454*, so called *pyrosequencing*, which became commercially available in 2004. Novel and diverse NGS instruments rapidly became available over the past decade; notable examples include companies such as *Illumina*, *Pacific Biosciences*, and recently *Oxford Nanopore*, to name a few. The NGS platforms shown in Figure 1.9 follow Mardis (2017).

The arrival and commodification of NGS technologies have made it feasible to sequence a whole human genome within days or weeks, rather that months or years. There is an ongoing reduction in labour and cost, while speed and accuracy of data generation is improving. For example, the cost of the Human Genome Project (HGP) sequencing the first human genome has been estimated at more than $3 billion. The first human diploid genome (James Watson) was sequenced for less than $1 million (Wheeler *et al.*, 2008). Currently, the goal of the $1,000 genome is surprisingly close; see Figure 1.10 (next page).

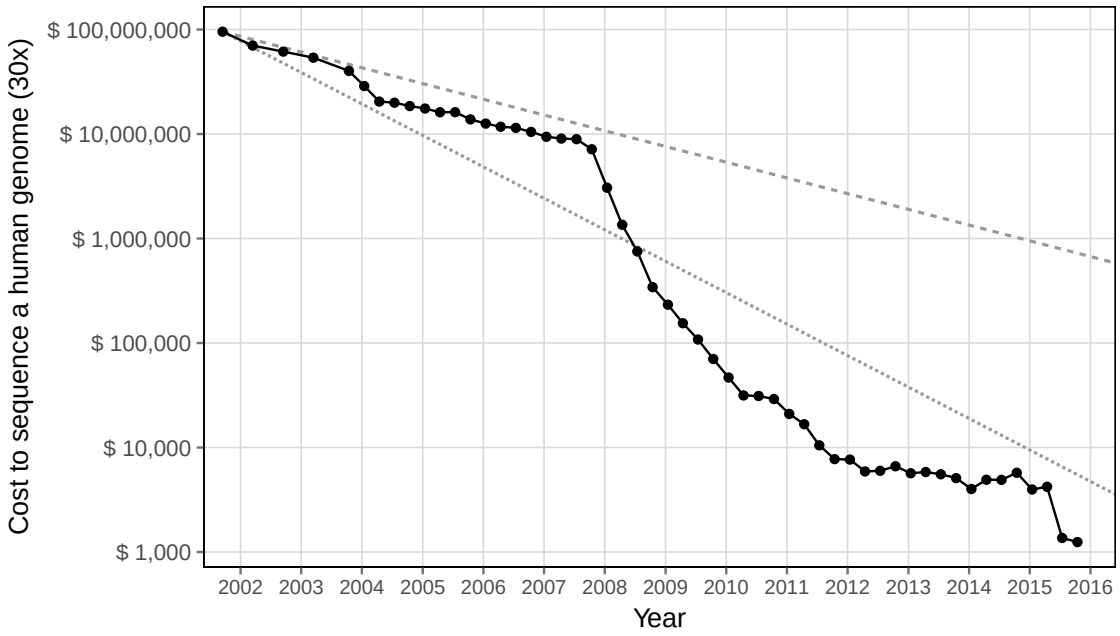


Figure 1.10: Timeline of cost reduction in DNA sequencing. Technological improvements in whole-genome sequencing have led to drastic reductions in cost while simultaneously improving accuracy and speed of data generation. The plot shows the development of price per human-sized genome sequenced at 30x depth (price given in US dollars) since the publication of the first draft sequence of the human genome in 2001. The costs shown between 2001 and 2007 are based on the Sanger sequencing method (*first-generation* methods); since 2008, costs are based on *next-generation* technologies. The hypothetically expected rate of cost reduction per genome is indicated according to Moore's law (Moore, 1965); the price halves every two years (*dashed*) or every year (*dotted*). Data provided by the National Human Genome Research Institute (NHGRI): `https://www.genome.gov/sequencingcostsdata/` [Date accessed: 2017-03-15].

1.4.2 Exploration of the human genome

Our understanding of genetic information and the forces that shape variation in a population has grown substantially since the early breeding experiments on pea plants conducted by Mendel (1866), who formulated the fundamental laws of genetic inheritance, rediscovered more than 30 years later (Correns, 1899; De Vries, 1900; Tschermak, 1900). Yet, our patience to wait for such important insights has been decreasing exponentially.

Before the HGP was planned, an initial human genetic linkage map had been established using restriction fragment length polymorphisms (RFLPs) in 1980 (Botstein *et al.*, 1980). A second-generation linkage map of the human genome had been constructed by 1993, using microsatellite markers (Weissenbach, 1993). In 2001, linkage disequilibrium (LD) patterns had been documented for parts of the genome, using a combination of early sequencing methods and genotyping (Daly *et al.*, 2001; Reich *et al.*, 2001).

The release of the draft sequence of the human genome in 2001 led to numerous large-scale projects. For example, GWA analyses of complex diseases required the identification of genetic markers prior to interrogation; to this end, the International HapMap Project (HapMap) was initiated to validate several million SNP markers and to examine LD patterns within different populations, eventually providing haplotype information for a representative global sample. In addition, a central aspect of the HapMap effort was to develop methods enabling GWA analysis.

The HapMap Project consisted of several phases of data acquisition and release. Phase I involved the genotyping of 1.3 million SNPs in 270 individuals from four global populations (International HapMap Consortium, 2003). Subsequently, Phase II aimed to increase the genotyping density in these same individuals to further improve the ability to map associations, supplementing the Phase I release with another 2.1 million SNPs (International HapMap Consortium *et al.*, 2007). In conjunction with the Human Genome Project and the SNP Consortium (McCarroll *et al.*, 2008), approximately 11 million common SNPs had now been identified. Finally, Phase III focussed on the coverage of additional populations, culminating in a total of 1,397 samples from 11 populations (Release 3), of which 692 individuals had been additionally sequenced at selected regions (International HapMap 3 Consortium *et al.*, 2010).

With the advantage of new NGS technologies, the 1000 Genomes Project was launched in 2008, with the aim of sequencing the genomes of at least 1,000 individuals across different populations, in order to provide a comprehensive resource of observed human genetic variation that could be leveraged by GWA studies and research in population genetics. The pilot phase described approximately 15 million SNPs, most of which had not been identified previously (Altshuler *et al.*, 2010). Several pilot projects were undertaken, including low-coverage WGS of 179 individuals from four populations, high-depth sequencing of two trios (parents and child), and targeted exome-sequencing of 697 individuals from seven different populations.

The variants discovered in the pilot stage were common (> 5% minor allele frequency); that is, low-frequency variants were underrepresented. Although the prevalent hypothesis at that time proposed that the variants underlying common diseases will also be common in the population (Lander, 1996; Chakravarti, 1999), it was argued that low-frequency and rare variants may also contribute to disease susceptibility and therefore could further our understanding of complex phenotypes (Pritchard, 2001). But to capture variants that occur at lower frequencies per population sample, it was necessary to sequence hundreds or thousands of genomes (Kaiser, 2008).

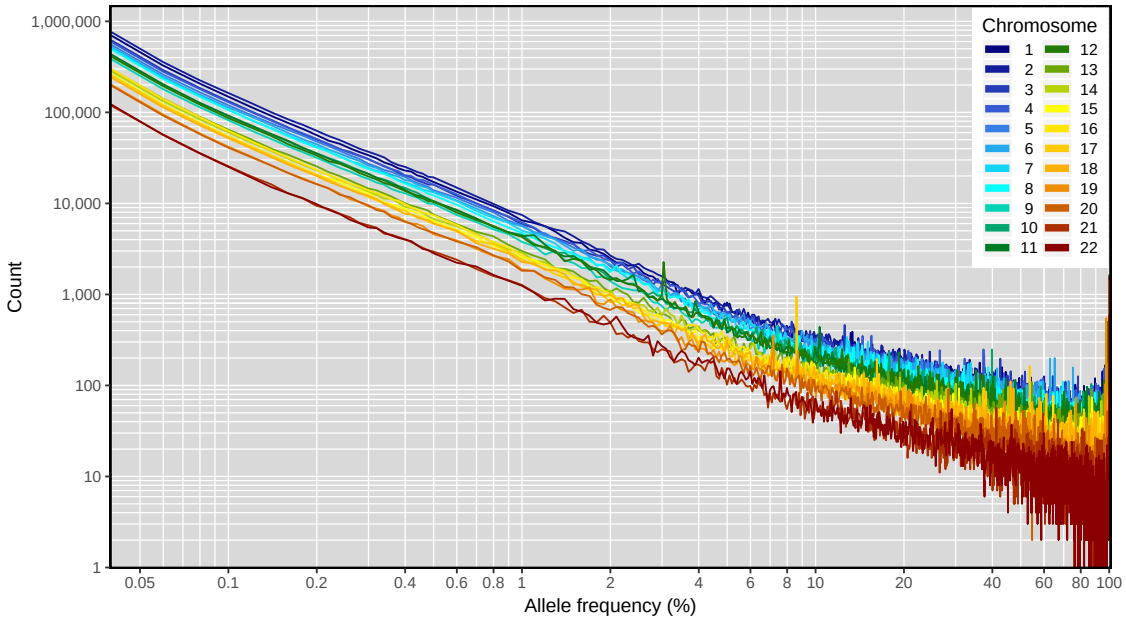


Figure 1.11: Allele frequency spectrum in the 1000 Genomes Project. The allele frequency distribution is shown per chromosome (1–22) for all variants contained in the final release dataset of 1000G Phase III. Singletons (private mutations observed only once in the sample) were excluded. Note that data are shown on log-log scale.

This led to Phase I of the 1000 Genomes Project, carried out on 1,029 individuals from 14 populations, and comprising a combination of low-coverage WGS, targeted exome sequencing, and genotyping by microarray. This resulted in the profiling of 38 million SNPs in total, with the majority being rare (1000 Genomes Project Consortium *et al.*, 2012). It must be noted that low-coverage sequencing is unlikely to capture rare variants with high accuracy, as they may not be called correctly due to inherent sequencing errors. However, Phase I represented a crucial step towards achieving complete characterisation of the genetic variation present in the human genome. Phase II of the project focussed on methods development, while increasing the sample size to 1,700 indivudals; these methods were applied to a total of 2,504 samples from 26 populations in Phase III, leading to a final release dataset of 84.7 million SNPs and the completion of the project (1000 Genomes Project Consortium *et al.*, 2015). Notably, although Phase III conducted low-coverage whole-genome (> 4×) and high-coverage exome (> 50×) sequencing, variants were called using improved methods (*e.g.* haplotype-aware variant callers and methods based on *de novo* assembly) to produce the final dataset. Figure 1.11 (this page) illustrates the allele frequency spectrum of all variants identified through the 1000 Genomes Project (final release, Phase III); shown per chromosome after removal of private mutations (singletons).

1.5 Genome-wide association studies

The International HapMap Project was instrumental to the design of GWA studies by validating millions of SNPs in the human genome and revealing the structure of genetic variation through patterns of LD in different populations. Due to the non-independence of markers, association analyses may only interrogate a modest subset of variants to detect common risk alleles. It was shown that the efficiency of GWA studies could be maximised by scanning only a fraction ($\approx 1\%$) of the 11 million SNPs that were known at that time (de Bakker *et al.*, 2005; Pe'er *et al.*, 2006). The availability of HapMap data was used to guide the development of genotyping arrays, to tag SNPs markers that are informative to capture most of the variation between individuals.

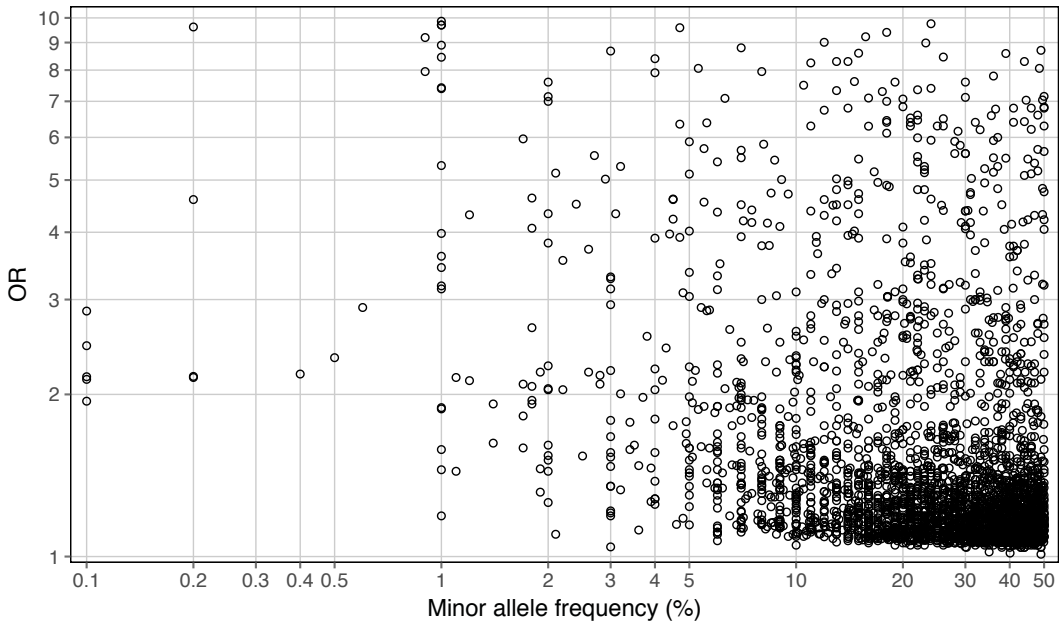


Figure 1.12: Significant risk-associated variants listed in the NHGRI-EBI Catalogue. Results are shown for 3,186 unique variants which were reported as being significant at p-value $\leq 5 \times 10^{-8}$ and for which odds ratio (OR) values were available in the database. Note that different studies may report different minor allele frequency (MAF) and OR. Duplicate entries (variants reported in more than one study) were removed, after calculating the median value of MAF and OR across duplicates; frequencies were then rounded to three decimal places. Data were taken from `http://www.ebi.ac.uk/gwas/` [Date accessed: 2017-01-20].

The first proper GWA study was undertaken by Klein *et al.* (2005), who successfully identified a common variant of large effect size to be significantly associated with age-related macular degeneration. The number of subsequent GWA studies rapidly increased; by 2007, more than 100 studies had been published, which was considered

as the "breakthrough of the year" by *Science* (Pennisi, 2007). Currently, the GWAS Catalogue maintained by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EBI) lists 2,324 publications and reports more than 30,000 unique SNP associations of which more than 8,000 are significant at p-value $\leq 5 \times 10^{-8}$ for approximately 1,000 traits (Burdett *et al.*, 2016).* The bulk of these results is summarised in Figure 1.12 (page 30), in which I show the relation between risk effect size and allele frequency for identified risk-associated variants at p-value $\leq 5 \times 10^{-8}$.

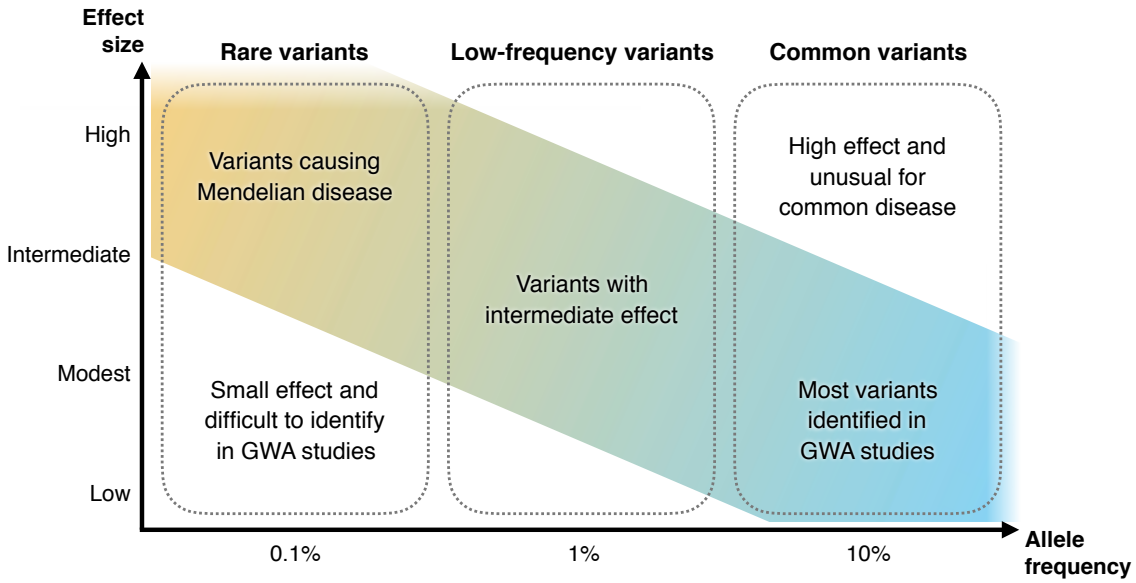


Figure 1.13: Risk-related variants by allele frequency and effect size. Rare, low-frequency, and common variants are distinguished by (minor) allele frequency. Note that frequency values are only indicated as approximate guides. Figure adapted from McCarthy *et al.* (2008, Box 7) and Manolio *et al.* (2009, Figure 1).

In contrast to traditional linkage approaches, which have high power to locate low-frequency variants of large effect size (*e.g.* Mendelian diseases), genome-wide association was designed and has proven to be powerful for interrogating common variants with modest effects. This disparity is illustrated in Figure 1.13 (this page), which outlines a seemingly categorical distinction between rare, low-frequency, and common variants based on expected penetrance and the ability to detect effects resulting from such genetic factors. The limitations of both approaches lie at the extremes (outside the band indicated in Figure 1.13).

Notably, rare variants with modest or low penetrance are difficult to detect by either linkage or GWA analysis. Since it became apparent that the human genome harbours an abundance of rare and low-frequency variants, it has been suggested that a large

* NHGRI-EBI GWAS Catalogue: http://www.ebi.ac.uk/gwas/ [Date accessed: 2017-01-20]

proportion of rare variants may also have functional implications with low to modest effects (Coventry *et al.*, 2010; Keinan and Clark, 2012; Tennessen *et al.*, 2012). Using GWA methods, the interrogation of alleles observed at very low (rare) frequencies may represent a conceptual limitation, however, it is hoped that the detection of low-frequency variants with intermediate effect can be improved.

1.6 Identity by descent

Relatedness among individuals is a natural property of genetic inheritance. Although this observation may seem trivial as we all inherit our DNA from somebody,* knowledge about the genetic relationship between individuals is crucial to many applications in genetic research. The validation of individual relationships is of particular interest in family-based methods such as linkage analysis (Purcell *et al.*, 2007; Albrechtsen *et al.*, 2009), or to exclude pedigree errors that would influence statistical power in linkage studies (Boehnke and Cox, 1997), but also in population-based (case-control) association studies of purportedly unrelated individuals, where unreported relatedness may lead to spurious results due to population stratification, *i.e.* systematic differences in the ancestry of individuals (Freedman *et al.*, 2004; Voight and Pritchard, 2005).

The relationship between individuals is indicated by the alleles they have in common, where two alleles are said to be *identical by descent* if they have been co-inherited from a common ancestor (Thompson, 1974, 1975). The concept of identity by descent (IBD) was introduced by Cotterman (1940) and extended by Malécot (1948) who provided probability formulations of IBD in related individuals; the term "identity by descent" was coined by Crow (1954). Notably, Malécot (1948) defined IBD as the probability that no mutation occurred since the common ancestor; see also Slatkin (2008a). In contrast, identity by state (IBS) refers to alleles that are observed to be the "same", but which may not be shared by descent.

1.6.1 Single-locus concept

Traditional measures of relatedness define IBD as the gametic relationship at a single locus, for which in particular the inbreeding coefficient and the kinship coefficient introduced by Wright (1921, 1922) have been relevant. For example, the probability that two homologous alleles are identical by descent in the same diploid individual is

* Until CRISPR/Cas9 genome editing has been established (*e.g.* see Cai *et al.*, 2016); in reference to the term *identity by descent* (IBD) I propose the term *identity by modification*, or IBM. [*Castigat ridendo mores*]

given by the inbreeding coefficient. However, such traditional approaches often assume that the relationship status of the individuals is known or can be derived from possible pedigree relationships, where ancestors are defined with respect to the founders of a pedigree. It has been argued that ancestry defined in reference to a founder sample is "something arbitrary" (Maynard Smith, 1989, p 141); see Rousset (2002). Moreover, this definition of IBD (in particular the distinction between IBD and IBS) seems to be in conflict with coalescent theory, which postulates that every allele is technically identical by descent in the individuals which carry them, because all shared mutations in the genome can be traced back to a common ancestor at different times in the past (Powell *et al.*, 2010); that is, given the assumptions of the infinite sites model (Kimura, 1969; Watterson, 1975).

1.6.2 Genealogical concept

Given the recent advances in genomic technologies, single-locus concepts of IBD have become less common and are supplanted by genealogically defined concepts of *haplotype sharing by decent* in large samples of unrelated individuals (Thompson, 2013; Wakeley and Wilton, 2016). For example, the inference of IBD sharing has been useful to provide information about historical migration events and to reconstruct the demographic history of a population (Palamara *et al.*, 2012; Palamara and Pe'er, 2013; Harris and Nielsen, 2013).

If an allele at a given locus has been co-inherited (recently) by two or more individuals, it is likely that alleles at the surrounding loci on the same chromosome were also derived from the same ancestral lineage in those individuals. The definition of IBD is therefore extended to refer to homologous chromosomal *segments* that are identical by descent if they have been co-inherited without intervening recombination from a common ancestor (Hayes *et al.*, 2003; Powell *et al.*, 2010), such that the genealogical relationship between two haplotypes is the same along the shared region. Consequently, meiotic recombination is seen as the driving force that shapes the patterns of relatedness among individuals. The length of a shared IBD segment is delimited by recombination events that occurred independently in each lineage; IBD therefore results from the unique pairwise relationship between two gametes. To illustrate the genealogical concept of IBD, consider the example shown in Figure 1.14 (next page).

Note that recombination events may not always result in the termination of an IBD segment. This is because a coalescent event may join the two lineages broken up by recombination back together (back in time), forming a 'closed loop' in the ARG (see Griffiths and Marjoram, 1997a, Theorem 2.4). Further, haplotype segments that are

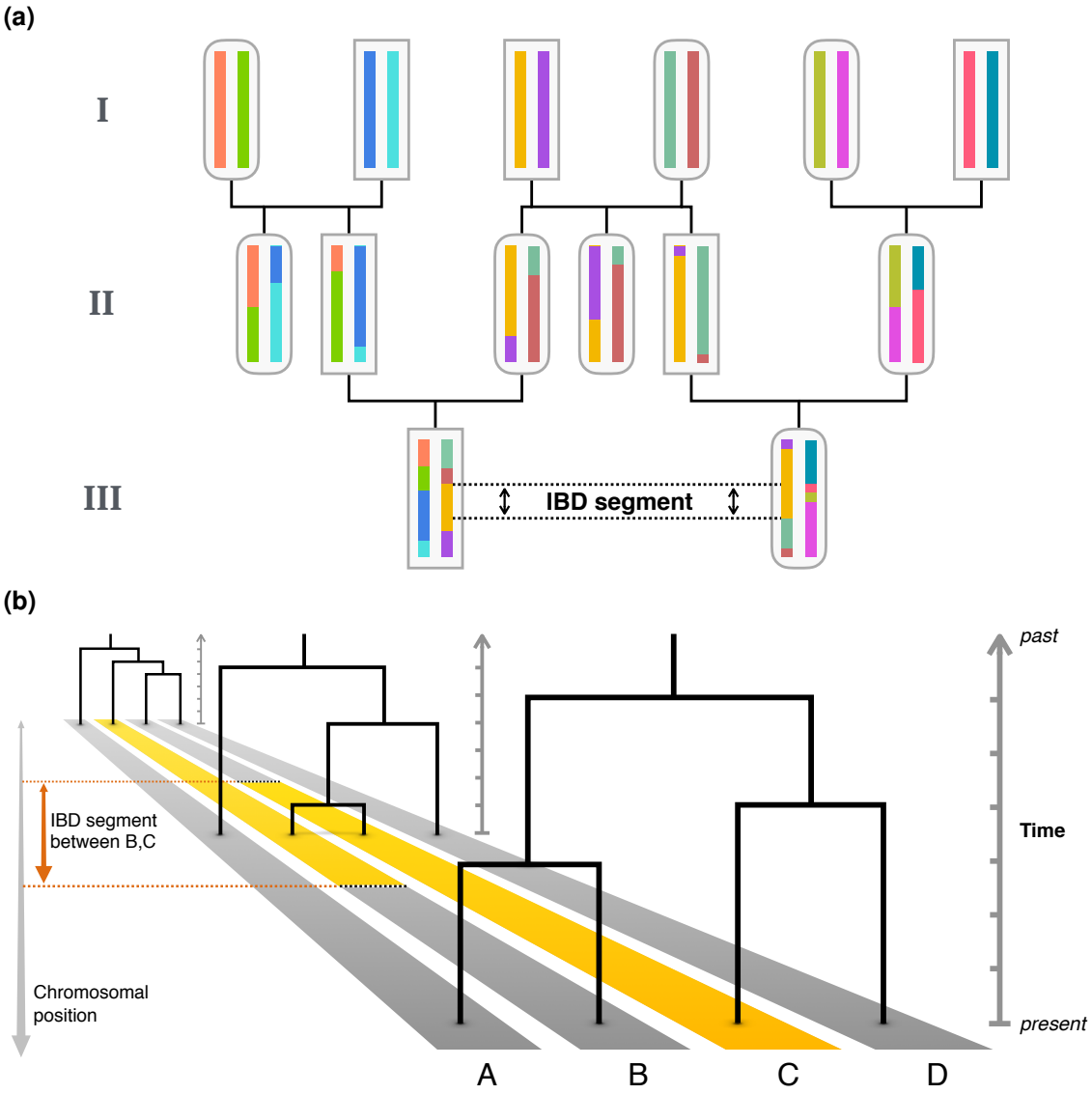


Figure 1.14: Illustration of haplotype sharing by descent. Panel **(a)** shows a three-generation pedigree; generation I consists of the founders of the pedigree. The two individuals shown in generation III are first-degree cousins. Male and female individuals are distinguished by square and round shapes, respectively. Each individual carries a diploid genome, shown as two large homologous chromosomes. The colour of each chromosome indicates the "identity" of the shared ancestral haplotype, which is shuffled with the other haplotype present in the same individual due to meiotic recombination in each generation, such that the offspring receives a unique arrangement of haplotype segments per chromosome from each parent. The "shared" haplotype refers to the overlapping region of haplotypes that are identical by descent; *i.e.* the IBD segment shared by the two individuals in generation III, indicated by the *orange* ancestral haplotype. For simplicity, colours indicate ancestry relative to the founders of the pedigree shown. Panel **(b)** illustrates the different genealogies along the length of the sequence of four chromosomes (A, B, C, and D), indicated by three marginal trees. The IBD segment co-inherited by chromosomes B and C is found at the overlapping region of the shared ancestral haplotype of the MRCA (*orange*). Note that the four chromosomes given in Panel (b) show a simpler arrangement of haplotypes than shown in Panel (a).

identical by descent may not actually be "identical", because the alleles observed along the shared sequence may differ. This is because mutations accumulate along each lineage independently, such that IBD segments separated by many meioses carry an increasing number of pairwise mutational differences. Likewise, it is expected that the length of the shared segment is decreasing over time due to recombination. As such, the "signal" of IBD might be lost for relatively old relationships, which can be described as the genetic "event horizon". In practice, the detection of IBD segments is therefore often limited to recently inherited shared haplotypes (*e.g.* < 100 generations); see Browning (2008).

1.7 Allele age estimation

There has been growing interest in being able to estimate the age of alleles that segregate in contemporary human populations; that is, the time since an allele was introduced into a population through a mutation event. The age of an allele, in conjunction with patterns of allele sharing, would allow us to better understand human evolutionary history and past demographic events and processes. It has been suggested that by knowing the age of alleles, geneticists will be able to build a "time machine" to explore our past (Slatkin and Rannala, 2000).

A number of mechanisms can affect the frequency at which an allele that emerged at some unknown point in the past is observed in a population. For example, an allele might be under purifying selection and hence on its way to becoming extinct. Conversely, it might endow a selective advantage and is therefore increasing in frequency. If the allele is neutral it could be subject to random genetic drift or simply be present due to a founder effect. Finally, the heterozygous state might have a selective advantage, meaning that the allele is held at a steady frequency in the population despite being "old" (Colombo, 2007).

1.7.1 Theoretical results

The field of population genetics has been fascinated with the possibility of estimating the time of mutation events. Early and often purely theoretical approaches had been conceived prior to the discovery of the coalescent. For example, Kimura and Ota (1973) found that the frequency of an allele can be used as an estimator for its age, which they derived in a diffusion process. The expected age of a neutral allele in a constant population is given by

$$\mathbb{E}[t_m] = \frac{-2x}{1-x} \log(x) \tag{1.30}$$

where x denotes the frequency of an allele observed in a sample; the age, here denoted by t_m, is scaled in units of $2N$. Notably, this and other contributions to the field by Kimura were deserving of a dedicated review (Watterson, 1996).

Related results were provided by Maruyama (1974) and Li (1975), who considered allele age as a random variable for which the probability of reaching fixation or extinction is regarded in presence of selection (*i.e.* assuming that the allele is beneficial or deleterious, respectively). Using diffusion methods, they have shown that (purifying) selection reduces the average age of an allele, whereas mutations that increase fitness also increase the average age. Watterson (1976) further developed the theory to provide the probability distribution of allele age conditional on its frequency; see review by Slatkin (2000) and Slatkin and Rannala (2000).

An alternate approach was proposed by Thompson (1976), who considered the age of an allele as a fixed parameter to derive the likelihood function for the age using a discrete branching process model, given the number of allele copies found in a sample. Notably, Thompson (1976) has shown that it is unrealistic to arrive at an exact point estimate for the age of a given variant in a sample, due to the stochastic nature of genetic evolution in natural populations. However, it is possible to derive a confidence interval to delimit the period during which a mutation event is likely to have occurred.

Later, Griffiths and Tavaré (1998) extended these earlier results in context of the coalescent. For example, the following formulation describes the expected age of an allele under a constant population size and the assumption of the infinite sites model (Kimura, 1969; Watterson, 1975);

$$\mathbb{E}[t_m] \;=\; 2\binom{n-1}{b}^{-1} \sum_{j=2}^{n} \binom{n-j}{b-1} \frac{n-j+1}{n(j-1)} \tag{1.31}$$

where b denotes the number of allele copies observed in a sample of size n. The above is equivalent to Equation (1.30) and provides conform estimates based on allele frequency alone. Nonetheless, a general conclusion reached by the field was that the distribution of allele age based on its frequency alone is too broad to provide reliable age estimates, which meant that there was only little practical utility (see Slatkin, 2000).

However, due to the growing interest in exploring the genetic and genealogical basis of human disease, several other methods have been developed, most of which based on *intra-allelic variability*, which is defined as the extent of variability observed at closely linked markers (Slatkin and Rannala, 2000; Slatkin and Bertorelle, 2001). Note that this idea can be seen as a progenitor to the genealogical IBD concept presented in the previous

section (page 33); that is, before recombination had been first mentioned in the definition of IBD (Hayes *et al.*, 2003).* These methods have been applied to numerous cases, some of which are summarised in the following section.

1.7.2 Application in human disease research

I provide three examples of studies in which the age of an allele has been estimated. The first two studies below represent early examples that have been conducted in context of a specific disease on limited data; *i.e.* prior to the high-throughput sequencing era. The third and more recent study was conducted "blindly", in a hypothesis-generating approach on more than a million protein-coding variants using exome-sequencing data, without targeting specific loci of known disease association.

Serre *et al.* (1990) analysed the $\Delta F508$ mutation of the *CFTR* gene, which had been identified as causing cystic fibrosis, and is higher in frequency in European populations compared to other populations. They used restriction fragment length polymorphism (RFLP) data from 240 French families, estimating the age from the variation observed at two linked loci. As a result, they estimated this mutation to have occurred 3,000 to 6,000 years ago, which was consistent with an estimate of approximately 3,000 years found by Slatkin and Rannala (2000), who replicated the study on intronic microsatellite data provided by Morral *et al.* (1994).

Risch *et al.* (1995) examined six closely linked microsatellite markers in data from 59 Ashkenazi Jewish families with idiopathic torsion dystonia (ITD), a rare disorder involving involuntary and sustained muscle contractions. They showed that cases with early-onset ITD (Oppenheim's dystonia) are due to a single founder-mutation. Based on linkage analysis and observations of strong LD around the ITD locus, they estimated this mutation to have emerged around 350 years ago (assuming 25-year generations). However, Labuda *et al.* (1996) provided a correction to account for founder effects in the model, which suggested that the mutation originated several centuries earlier than reported by Risch *et al.* (1995), during a period when the Jewish population was founded in eastern Europe. This corrected result was further confirmed through re-analysis by Slatkin and Rannala (2000).

As noted by Slatkin and Rannala (2000), the recombination and mutation rates (as well as other demographic parameters such as the exponential growth rate) used to estimate allele age represent a source of uncertainty. For example, the age range reported by Serre *et al.* (1990) was estimated based on several values that were consistent with data available at the time.

* Note that the connection between identity by descent, linkage, and recombination had been anticipated long before (*e.g.* see Donnelly, 1983).

In a more recent study, Fu *et al.* (2012) used exome data from 6,515 individuals and estimated the age of more than 1 million protein-coding SNPs, using a simulation-based approach under several established demographic models. In addition, they predicted whether variants were deleterious using a range of different methods, Interestingly, they found that the probability that a variant was predicted to be deleterious was strongly related to estimated allele age. Fu *et al.* (2012) found that some of the genes surveyed, among those which had been associated with human diseases, showed a significant excess of putative deleterious variants which were estimated to have a relatively recent origin through mutation. For example, several of those genes had been implicated in coronary artery atherosclerosis (*CPE*), hereditary spastic paraplegia (*KIAA0196*), premature ovarian failure (*LAMC1*), and Alzheimer's disease (*LRP1*). In fact, the majority of identified deleterious variants within gene-coding regions were rare in frequency, enriched for mutations of large effect size, and indicated to have emerged relatively recently, within the last 5,000–10,000 years.

In general, it has been argued that the large number of rare variants observed in the human genome is due to a recent, explosive population growth, following a bottleneck population size after the expansion out of Africa 50,000–100,000 years ago, and the advent of agriculture, approximately 10,000 years ago (Coventry *et al.*, 2010; Keinan and Clark, 2012; Tennessen *et al.*, 2012). For example, the effects of (weak) purifying selection can be considered as being too slow to purge young alleles with disadvantageous phenotypic consequences from the population, such that there might be an unrecognised large abundance of rare variants in the human genome which could influence disease risk in yet unaccounted ways. An argument to the contrary, however, suggests that recent demographic changes such as population growth may have had negligible impact on the mutational load carried by an individual on average (Simons *et al.*, 2014). As such, the amount of ascertained rare variants may not necessarily contribute to complex disease risk unless they exert strongly deleterious effects on fitness. Thus, it remains to be seen whether rare variants play an important or an inconsequential role with regard to complex disease susceptibility; to that end, knowledge about their age may lead to a better understanding of disease aetiology. Regardless, the estimation of allele age still remains a matter of curiosity.

We chose it because we deal with huge amounts of data.
Besides, it sounds really cool.

— Larry Page, co-founder of Google Inc.

2

Meta-imputation of multiple reference datasets for genome-wide association studies

Contents

2.1 Introduction

Genome-wide association (GWA) studies have identified thousands of genetic risk factors that influence disease susceptibility and complex disease phenotypes. A contributing factor to this success is the ability to statistically estimate, or *impute* genotypes that have not been observed in a study sample. Genotype imputation has become a standard technique in GWA studies where it is used to increase the number of variants to achieve higher power in association analysis as well as to facilitate meta-analysis of association results across different studies (Marchini *et al.*, 2007; Marchini and Howie, 2010). Methods for genotype imputation match patterns of genetic variation observed in a study sample

with a more densely typed set of haplotypes in a reference panel. The extent of shared variation is informative for estimating the most likely genotypic states at other, unobserved variant sites in the same individuals. Commonly employed imputation methods are, for example, `Beagle` (Browning and Browning, 2016), `MACH` (Li *et al.*, 2010), and `IMPUTE2` (Howie *et al.*, 2009, 2011).

Genotypes can be imputed with remarkably high accuracy, allowing researchers to assay only a modest number of markers in sampled individuals, which makes large-scale data collection feasible and cost-effective (Li *et al.*, 2009). The accuracy of imputation is dependent on several factors. These include the number of genotyped markers in the study sample, the size of the reference panel, and the genetic similarity between sampled and reference individuals (Howie *et al.*, 2009; Roshyara and Scholz, 2015). The coverage of the reference panel further influences the power to find significant associations. The availability and choice of reference data therefore becomes crucial in considerations of statistical power in study design.

One of the first larger sets of publicly available reference genomes was established by the International HapMap Project (HapMap), which identified 3.1 million variants through genotyping of 270 individuals from four continental populations (International HapMap Consortium *et al.*, 2007; International HapMap 3 Consortium *et al.*, 2010). More recently, the 1000 Genomes Project (1000G) released reference data in three phases at progressively increasing sample size, currently reaching over 88 million variants from low-coverage whole-genome sequencing (WGS) of 2,504 individuals from 26 populations (1000 Genomes Project Consortium *et al.*, 2012, 2015). Due to ongoing advances in next-generation sequencing (NGS) technologies and reductions in costs, large-scale WGS studies have become routine. However, genetic variation generally shows extensive stratification dependent on geography and ethnicity. Also, disease risk factors can be segregated on a much finer scale. Therefore, any study may only capture the variation present in the population or study cohort sampled, particularly among lower frequency and rare variants.

To increase the chance of detecting significant risk variants through GWA methods, it would be desirable to combine sequencing data from different studies to generate a single, large reference panel for imputation. However, the integration of independently produced datasets is not straightforward due to differences arising from different sequencing platforms, coverage, and strategies to filter and call variant genotypes. It is not directly feasible, for example, to compile an unbiased union of variant calls across studies,

because monomorphic sites cannot be distinguished from sites that were filtered or missed. Conversely, retaining the intersection of variants that are present in all panels would dispose of much information.

One solution would be to re-process raw sequence or genotype data from multiple studies together, where variants are jointly called and phased over a combined set of samples. For example, in a large-scale collaborative effort, the Haplotype Reference Consortium (HRC) has recently created a reference panel from study data of 20 participating cohorts, which included a total of 64,976 human haplotypes in its first release (McCarthy *et al.*, 2016). This dataset currently represents the largest single resource of human genetic variation, but currently only includes samples of European ancestry. Although data are not accessible publicly, an online service has been provided for imputation and phasing from the internally stored reference dataset.*

Here, I propose an alternate solution in which multiple reference panels are separately imputed into a given study sample after which the genotype datasets produced are merged. Because imputed data may only differ in variant coverage, while the sample set is identical, it is feasible to merge data and integrate genotype information at overlapping sites. The underlying intuition is that the accuracy of an imputed genotype is indicated by its posterior probability or other metrics that result from the imputation process; for example *allelic* R^2 in `Beagle`, $\hat{r}^2$ in `Mach`, and *info-score* in `IMPUTE2`. The presented method applies such information to select from or assign higher weights to candidate genotypes, thereby indirectly leveraging information across different reference panels.

The following section (2.2) describes the approach by which sets of imputed genotype data are combined to form an integrated, larger genotype dataset; the method is referred to as *meta-imputation*. I considered several strategies to combine data based on different summary metrics. To be able to efficiently evaluate this method, as well as for application to genomic datasets on a larger scale, I implemented the method as a computational tool written in `C++` called `meta-impute`.† For assessment of meta-imputation, I constructed multiple, smaller reference panels from a larger dataset, which enabled comparisons between meta-imputation and direct imputations from both single and whole reference data. An additional analysis was conducted using data from several independent studies. The composition of reference data is described in Section 2.3 (page 46). The performance of meta-imputation was evaluated in regards to genotype accuracy and power to detect significant association signals. An accuracy analysis was conducted in Section 2.4 (page 48). Statistical power was analysed in a series of association experiments using simulated case-control data, which is described in Section 2.5 (page 62). Results are jointly discussed in Section 2.6 (page 70).

* HRC: `http://www.haplotype-reference-consortium.org` [Date accessed: 2017-02-05]
† Meta-imputation software (`meta-impute`): `https://github.com/pkalbers/meta-impute`

2.2 Approach

There are several ways by which genotypes imputed from independent sources can be combined at overlapping sites. To provide the means to explore a range of possibilities, the presented solution is implemented as a two-step process. First, a *score metric* is obtained for each genotype which, second, informs a *merge operation*. The general approach of meta-imputation is based on the assumption that a given metric is informative for distinguishing candidate genotypes that are more or less likely to reflect the underlying, true genotypic state. Here, several score metrics (Section 2.2.2, page 44) and two merge operations (Section 2.2.3, page 46) were considered, which are described after introducing principal notation and the general algorithm below.

2.2.1 Description of the method

It is convenient to think of genotype data as being arranged in a matrix, G, of size $M \times N$ where M is the number of observed variant sites and N is the diploid sample size. Let g_{ij} denote the genotype observed at marker i in individual j, such that g_i refers to the vector of genotypes of size N at the ith site, and g_j the vector of genotypes of size M belonging to individual j. Meta-imputation combines the information contained across several such genotype matrices. Let L denote the number of available genotype datasets imputed from different reference panels, such that $G_1, \ldots, G_L$ are available, and $k \in \{1, \ldots, L\}$ is used to identify a particular matrix; note that $L \geq 2$ is assumed. Because reference data were imputed into the same study sample, the number of individuals, N, is constant in each matrix but M_k may vary due to differences in coverage per reference panel.

Meta-imputation combines available genotype matrices in an aggregated matrix, A, of size $M_A \times N \times L$ where M_A is the number of variants in the combined set of sites across imputed panels. The algorithm merges genotype information at overlapping variants by gathering those that correspond to the same genomic position per chromosome. Here, the word *analogue* is used to refer to the set of available data vectors that correspond to the same variant. Let a_i denote an analogue variant, *i.e.* the set of overlapping genotype vectors at the ith site in the aggregated matrix, and a_{ij} an analogue genotype, *i.e.* the set of overlapping genotypes at this site in individual j. Note that the number of genotypes referred to by a_{ij} may vary dependent on presence in the reference panel. Let l denote the number of overlapping variants in an analogue, where $1 \leq l \leq L$, such that l_i refers to the size of a_i.

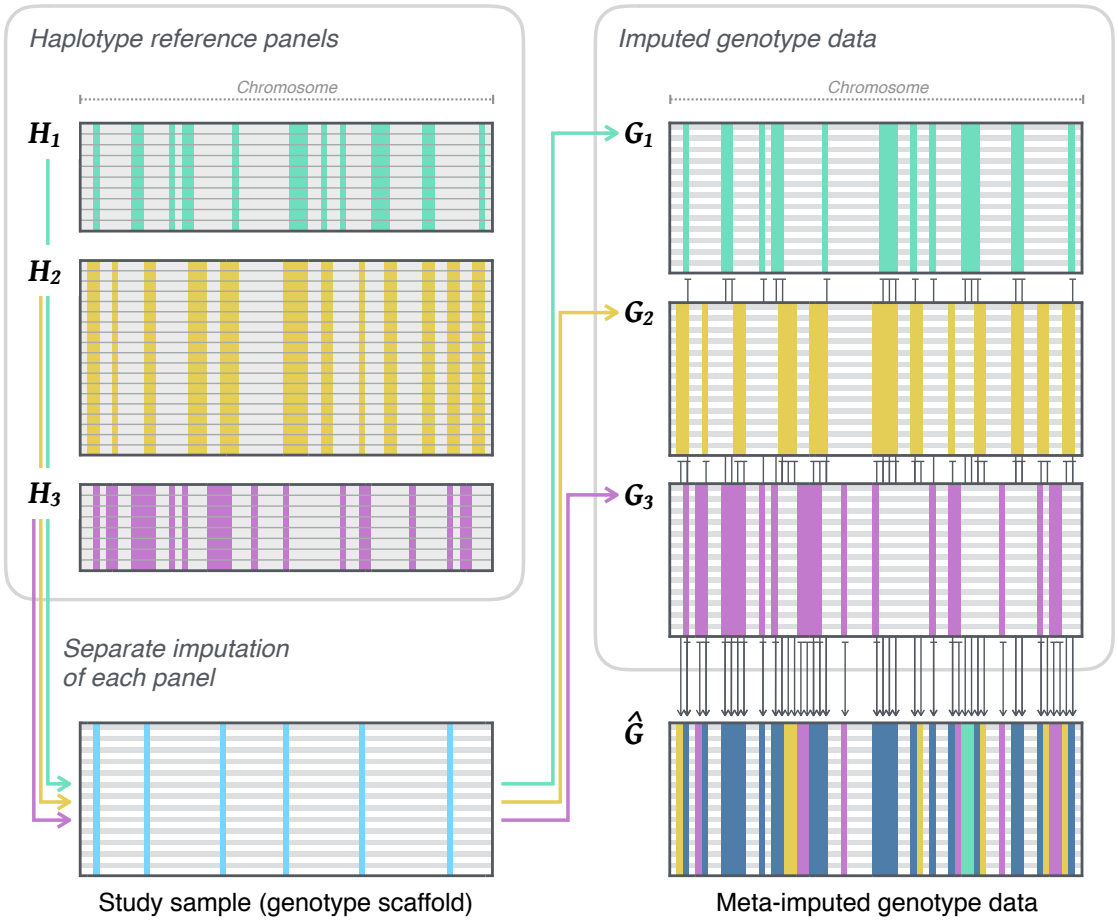


Figure 2.1: Illustration of the meta-imputation concept. An example of three haplotype reference panels is shown; denoted by H_1, H_2, and H_3, where haplotypes are indicated by row (*grey*) and observed variant sites are indicated by column. Each panel may vary in sample size and coverage. Reference data are separately imputed into the same study sample, which is a "scaffold" of typed genotype markers, where individual genotypes are indicated by row (alternating *grey-white*) and observed markers by column (*light-blue*). Each imputation returns an imputed genotype dataset, denoted by G_1, G_2, and G_3, containing marker genotypes as present in the corresponding reference panel, but where the number of individuals, N, is the same as in the study sample in each imputed dataset. Imputed data are combined through meta-imputation, such that the resulting genotype dataset, $\hat{G}$, contains the union of variant sites across panels. Variants merged across multiple datasets are indicated (*dark-blue*); the markers specific to a given panel are indicated by their corresponding colour.

Genotype formats may differ according to the type of data available. Note that the following considers single-nucleotide polymorphisms (SNP) specifically. In generic terms, a genotype can be observed in one of three possible states; homozygous for the reference allele, heterozygous, or homozygous for the alternate allele, which can be encoded by the alternate allele count (*allele dosage*); that is 0, 1, or 2, respectively. Imputed genotypes are typically expressed by the uncertainty associated with the imputation process. Here,

an imputed genotype is considered as a tuple $(p_0,\ p_1,\ p_2)$ of sum 1, representing the inferred posterior probability per genotypic state. Hence, a_{ijk} refers to a genotype tuple at the ith site in individual j taken from G_k.

The meta-imputation algorithm assigns a score value, s_{ijk}, to each a_{ijk}; *i.e.* each candidate genotype per analogue variant. A *meta-imputed* genotype is formed, denoted by $\hat{g}_{ij}$, by merging candidate genotype data conditional on the score assigned. At sites where $l_i = 1$, that is a given variant was imputed from only one reference panel, genotype data are retained as is, to capture as much variation as available from each separate imputation. The resulting genotype matrix, $\hat{G}$, contains the union of variants across input datasets. A simplified illustration of the meta-imputation concept is given in Figure 2.1 (page 43).

2.2.2 Score metrics

The score metrics considered in this work are described below; asserted 2-letter codes are used for the remainder of this chapter.

Maximum probability (`MP`). The mode of the probability distribution of a candidate genotype is taken as the value of the genotype's score; that is the maximum value in the tuple of posterior probabilities, which takes values in $[0, 1]$. The score is separately obtained for each candidate genotype, a_{ijk}, such that

$$s_{ijk} \ = \ \max\Big[(p_0, p_1, p_2)_{ijk}\Big] \ . \tag{2.1}$$

`IMPUTE2` information score (`IS`). The information score (or *info-score*) is used, which is a quality metric of the difference between observed and expected information, dependent on the imputed genotype distribution and estimated allele frequency; see definition below (Marchini and Howie, 2010, S3, eq. 16; modified here to correspond to present notation).

$$\mathrm{I}_{ik} \ = \ \begin{cases} 1 - \frac{\sum_{j=1}^{N} f_{ijk}\ e_{ijk}^2}{2N\hat{\theta}_{ik}(1-\hat{\theta}_{ik})} & \text{if } \hat{\theta}_{ik} \in (0, 1) \\ 1 & \text{if } \hat{\theta}_{ik} = 0, \hat{\theta}_{ik} = 1 \end{cases} \tag{2.2}$$

where $e_{ijk} = p_{1\,ijk} + 2p_{2\,ijk}$ is the expected allele dosage, similarly $f_{ijk} = p_{1\,ijk} + 4p_{2\,ijk}$, and $\hat{\theta}_{ik}$ is an estimate of the unknown population allele frequency, calculated as

$$\hat{\theta}_{ik} \ = \ \frac{\sum_{j=1}^{N} e_{ijk}}{2N} \ . \tag{2.3}$$

The `IMPUTE2` info-score takes values in $[0, 1]$ where values close to 0 or 1 indicate low or high certainty, respectively. This and other information measures (*e.g.* `Beagle` R^2 or `Mach` $\hat{r}^2$) are commonly used as a filter criterion in quality control (QC) of imputed GWA data.

Because meta-imputation was evaluated using IMPUTE2 for imputations (see Section 2.4.1.2, page 49), it is justifiable to use this information measure as a score metric. Since the info-score is calculated per imputed variant, the same score value is assigned to each candidate genotype imputed from a given reference panel at each site. Its value is assigned to each candidate genotype at a given imputed variant; that is

$$s_{ijk} = \mathrm{I}_{ik} \; \forall j \; . \tag{2.4}$$

Sample certainty (SC). A simple measure of imputation certainty is calculated per individual, such that a score value is assigned to genotypes across variants. This metric is calculated as the proportion of an individual's genotypes which have a maximum probability that satisfies a threshold rule, defined as

$$s_{ijk} = \frac{\sum_{i=1}^{M} \mathrm{I}_{ijk}}{M} \tag{2.5}$$

where

$$\mathrm{I}_{ijk} = \begin{cases} 1 & \text{if } \max\left[(p_0, p_1, p_2)_{ijk}\right] \geq r \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

where r is an arbitrarily defined value. In the present implementation, this threshold was set to $r = 0.9$. The intention of the SC metric is to prioritise imputations from reference haplotypes which show a closer fit to the genetic variation observed per individual in the study sample, which is assumed to be captured by the posterior probability at imputed genotypes. It must be noted that more sophisticated approaches for the estimation of genetic similarity exist, which provide summary statistics that could be used in place of the present score metric. Possible examples range from multi-locus statistics to fine-scale measures of population structure and demographic history (*e.g.* McVean *et al.*, 2004; Lawson *et al.*, 2012).

Random score (RS). In addition, the option to assign random score values to candidate genotypes was included, to be considered as a control against which the above metrics were compared. The score was explicitly calculated as

$$s_{ijk} = \frac{\mathrm{rand}(R)}{100} \, , \qquad R \in \{1, 2, \ldots, 99\} \tag{2.7}$$

where $\mathrm{rand}(\cdot)$ is a function which uniformly selects one value from R at random, such that $0 < s_{ijk} < 1$.

2.2.3 Merge operations

Any operation to merge the information available per analogue genotype can be divided into one of two conceptually distinct approaches; either one candidate genotype is selected and others are discarded, or a new genotype tuple is mathematically derived from available data. Accordingly, I considered the following two operations; note that the specified 3-letter codes are used henceforth.

Maximum score selection (`MSS`). A candidate genotype is selected by using score metrics as a ranking criterion, where the genotype tuple with the highest assigned score is selected from an analogue genotype in a_{ij} and retained as is in $\hat{g}_{ij}$; see below.

$$\hat{k} \;=\; \underset{k\in\{1,\ldots,l_i\}}{\arg\max}\left[s_{ij}\right] \qquad \textbf{s.t.} \qquad \hat{g}_{ij} \;=\; a_{ij\hat{k}} \tag{2.8}$$

If the highest score value is equal in more than one candidate genotypes, one is selected at random from those with the highest score.

Weighted linear combination (`WLS`). Tuple values of the meta-imputed genotype are derived from candidate genotypes as a linear combination of their posterior probability per genotypic state. This is calculated as the weighted average over analogue genotype probabilities, using corresponding score values as weights. Each candidate genotype thereby contributes to the resulting probability distribution in $\hat{g}_{ij}$. Probability values in each tuple a_{ijk} are multiplied by their assigned s_{ijk} after normalising scores such that values in s_{ij} sum to 1. The tuple of the meta-imputed genotype is then constructed by calculating the sum over the weighted probabilities at each genotypic state; see below (the mathematical definition follows Stone (1961)).

$$\hat{g}_{ij} \;=\; (\hat{p}_0, \hat{p}_1, \hat{p}_2)_{ij} \;=\; \sum_{k=1}^{l_i} (p_0, p_1, p_2)_{ijk}\; s_{ijk} \tag{2.9}$$

Implicitly, the resulting probability distribution in $\hat{g}_{ij}$ sums to 1. In contrast to `MSS` above, the weighted linear combination of genotype data does not discard available information. But note that tuple values may not be regarded as posterior probabilities when candidate genotypes were combined using `WLS`, but rather as "pseudo-probabilities".

2.3 Generation of reference datasets

Multiple reference panels were derived from the 1000 Genomes Project (1000G) Phase I dataset, which comprises both low-coverage whole-genome sequencing and whole-exome sequencing data of 1,092 individuals from 14 populations of European, East-Asian,

African, and admixed American ancestries.* This original dataset was split into non-overlapping subsets in two scenarios, A and B, reflecting situations when reference data of similar or distinct ethnic backgrounds would be available for imputation into a given study sample; see details below.

Scenario A included four panels composed of individuals belonging to European sub-populations (CEU, FIN, GBR, and TSI) as an example use case when different reference data of similar ethnic background are available.

Scenario B included four panels from different continental populations (AFR, AMR, ASN, and EUR) as an example use case when panels of distinct-ancestry samples are available.

Because sample sizes of the population groups considered in Scenario B differed in 1000G (more than in Scenario A), extracted individuals were randomly drawn from each group to create panels of equal size. Note that this was done to be able to better compare imputation accuracy among the generated panels, but which is not a requirement when using the meta-imputation method. Monomorphic sites and singletons were removed in each generated panel to more closely resemble data from independently conducted studies, where singleton or monomorphic variant calls are likely to be removed in the final dataset. In the following, the term *split panel* is used to denote subset reference data from 1000G. Throughout, analyses were limited to data from one chromosome, namely chromosome 20. This was done to allow for a larger number of replicate analyses, as will be described in Section 2.5 (page 62).

Generated split panels were used for separate imputations and subsequent integration of estimated genotype data through meta-imputation. To compare meta-imputed genotypes to those that were directly imputed from a unified panel, two additional reference datasets were generated from 1000G per scenario, which combined samples across respective split panels; referred to as the *intersection* panel and the *union* panel. The union reference contained variation as present in the original dataset, but for the individuals contained across split panels in a given scenario, and with monomorphic and singleton variants removed. The intersection reference contained the same set of individuals as the union panel, but where variant sites not shared across all split panels were removed. Unlike the split panels, from which imputed data were combined in meta-imputation, the genotype datasets obtained in imputations from the intersection and union panels were used in direct comparisons to meta-imputed data. The process of reference data generation is illustrated in Figure 2.2 (next page). A summary of the final reference datasets in each scenario is given in Table 2.1 (next page).

* Note that I completed work on the *meta-imputation* project prior to the release of 1000G Phase III (1000 Genomes Project Consortium *et al.*, 2015).

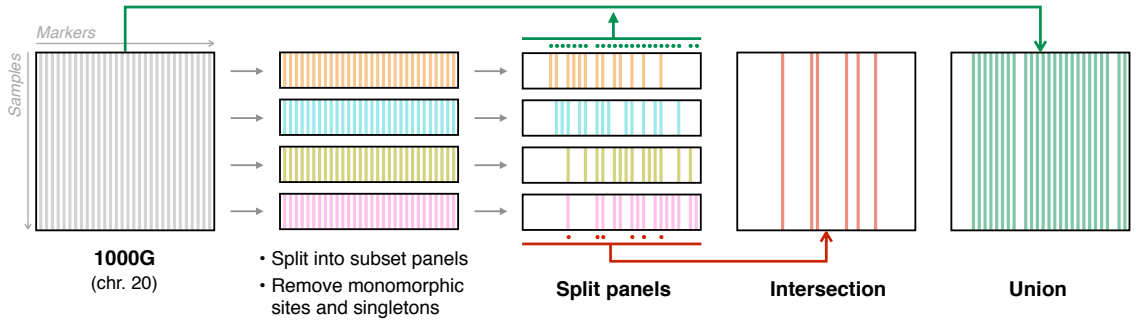


Figure 2.2: Generation of reference panels in each scenario. The original 1000 Genomes dataset (Phase I, chromosome 20) was used to generate multiple, smaller panels for imputation. This was done in two scenarios to create data of similar or distinct ethnic backgrounds. In each scenario, data were split into four *split* panels of approximately equal size. Monomorphic sites and singletons were removed in each split panel. Two additional panels were generated from the obtained split panels per scenario; one *intersection* panel and one *union* panel, both of which contained the union of individuals across split panels, but where the intersection panel only included sites if captured in all split panels, and the union panel included all sites as observed in the original dataset (except monomorphic or singleton sites as per the individuals included).

Table 2.1: Dimensions of generated reference data used for imputations. Panels included in Scenarios A and B were generated from the 1000G Phase I dataset. These "split" panels are named after their respective population codes in 1000G. Only data from chromosome 20 were considered. Note that split panels in Scenario B were reduced to match the size of the smallest panel in that scenario. Both the *intersection* and the *union* panels were created from the combined set of individuals across panels in each scenario.

Scenario A			Scenario B		
Panel	Samples	Variants	Panel	Samples	Variants
CEU	85	197,252	*AFR*	181	429,088
FIN	93	205,093	*AMR*	181	307,454
GBR	89	202,707	*ASN*	181	209,209
TSI	98	207,583	*EUR*	181	233,527
Intersection	365	168,744	Intersection	724	144,259
Union	365	253,852	Union	724	559,172

2.4 Accuracy of estimated genotypes

Evaluation of genotype accuracy was done in two parts. First, each combination of score metric and merge operation was tested and compared to select the best performing setting for downstream analyses. Second, meta-imputed genotypes generated under the selected setting were examined in comparison to genotype data imputed from each split reference panel, as well as the intersection and union imputations. Details about the methods used are given in the section below. Results are presented in Section 2.4.2 (page 51).

2.4.1 Methods

Calculation of genotype accuracy requires that the true genotypic states at untyped variants in a study sample are known. This was done by using a larger dataset from which a subset of variants was drawn to form an imputation scaffold. Missing variants were then re-imputed from available reference panels. The generation of the genotype scaffold is described below, followed by details about imputation, quality control, and the calculation of genotype accuracy.

2.4.1.1 Generation of genotype scaffold data (study sample)

The study sample used for imputations was extracted as a scaffold from data of the Genetics of Type 2 Diabetes Project (GoT2D), consisting of 2,657 individuals of Central and Northern European descent (Fuchsberger *et al.*, 2016).* The dataset is composed of data obtained on several platforms; low-coverage whole-genome sequencing (~ 5×), high-coverage whole-exome sequencing (~ 82×), and genotyping data using the *Illumina Omni2.5 Array*. To maintain a congruent set of markers in the genotype scaffold, variants typed on the latter were extracted from the larger GoT2D dataset, yielding 40,255 variants of in total 387,499 SNPs on chromosome 20 in GoT2D, after removing monomorphic sites and singletons. Remaining sites were masked for comparison after imputation, where imputed variants were matched to their corresponding sites in the masked dataset to calculate genotype accuracy.

Figure 2.3 (next page) shows the site frequency spectrum (SFS) of variants captured in the GoT2D dataset, highlighting the discrepancy between the frequency distribution observed at all captured variants and those obtained through genotyping only. Rare variants (*e.g.* at allele frequency ≤ 1%) are underrepresented in the genotyped set of SNPs and, thus, in the extracted genotype scaffold. Imputation from a reference panel into the scaffold attempts to fill these gaps, including sites with alleles occurring at lower frequencies.

2.4.1.2 Imputation and quality control

Imputations were performed using `IMPUTE2` version 2.3.0 (Howie *et al.*, 2009), and executed in consecutive chunks of 5 Megabases (Mb). The GoT2D dataset comprises already phased haplotypes, so imputations were carried out on pre-phased genotypes (command line argument `-use_prephased_g` in `IMPUTE2`). Because meta-imputation is indirectly based on

* GoT2D Consortium: http://www.type2diabetesgenetics.org/projects/got2d [Date accessed: 2016-12-02]

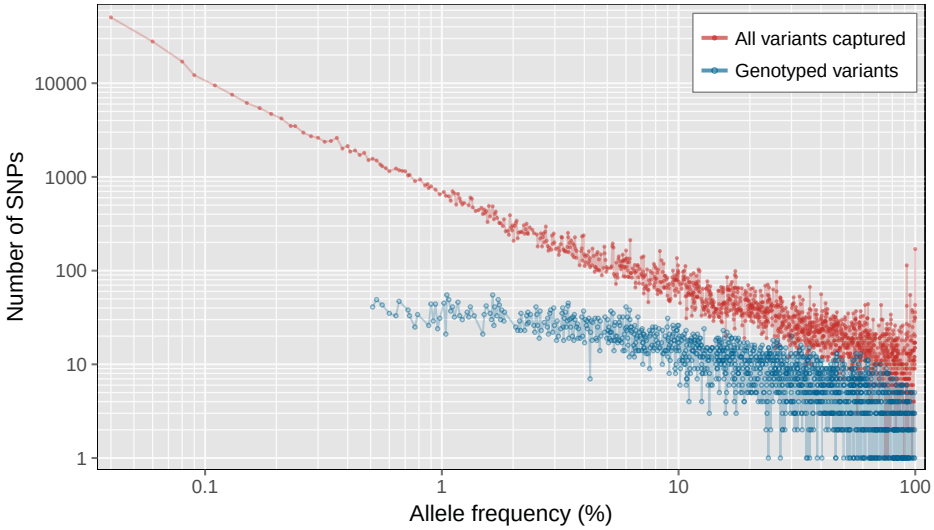


Figure 2.3: Site frequency spectrum of variants captured in GoT2D, chr. 20. The site frequency spectrum (SFS) is shown for SNPs as captured in the full GoT2D dataset (*red*) and SNPs genotyped using *Illumina Omni2.5 Array* (*blue*), given the allele frequencies observed in the full GoT2D sample for chromosome 20, after removing singletons and monomorphic sites.

information from more reference haplotypes than available in each separate imputation, the number of haplotypes that inform the imputation process was set to the maximum number present in a given reference panel (command line argument `-k_hap` in `IMPUTE2`). This was done to minimise potential biases in comparisons between meta-imputed and imputed genotypes, but is not a requirement for general applications of this approach.

Imputed and meta-imputed genotype data were filtered in QC, removing variants at `IMPUTE2` info-score < 0.4 and at deviations from Hardy-Weinberg equilibrium (HWE) at p-value $< 1 \times 10^{-4}$. These metrics were computed using `QCTOOL`,* which was performed on both directly imputed and meta-imputed datasets. Imputed data were filtered before the assessment of imputation accuracy, but not before integration through meta-imputation. The proportion of variants retained after QC was used as an indicator for data quality in comparisons between imputed and meta-imputed data, as well as to characterise the quality achieved though using different meta-imputation settings. Hence, QC results were separately reported for each part of the analysis. A summary of the described analysis is illustrated in Figure 2.4 (next page).

2.4.1.3 Calculation of genotype accuracy

Genotype accuracy was calculated as the squared Pearson correlation coefficient, r^2, as a measure for the strength of the linear relationship between imputed and masked genotype vectors, such that r^2 was computed per site. This was done after conversion

* `QCTOOL`: `http://www.well.ox.ac.uk/~gav/qctool` [`Date accessed: 2016-12-02`]

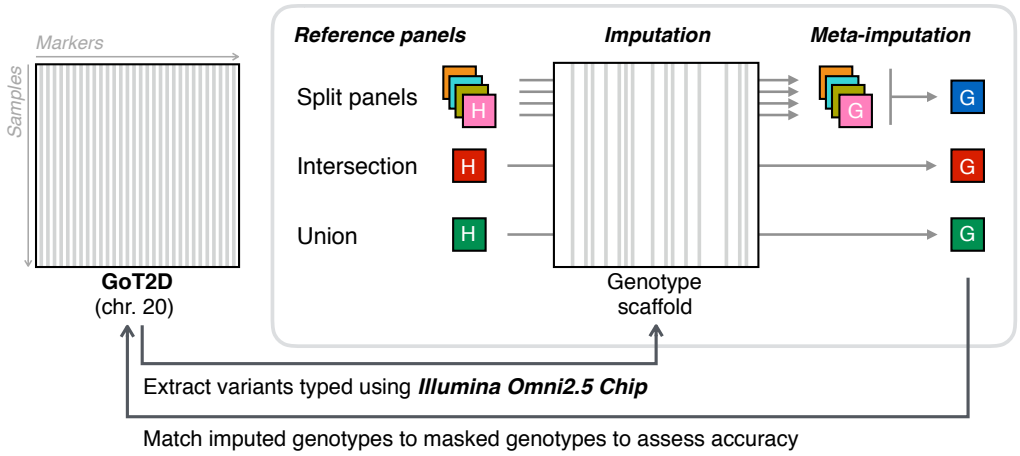


Figure 2.4: Illustration of the accuracy assessment process. Imputations were performed on the same genotype scaffold, which consisted of genetic markers obtained through genotyping using *Illumina Omni2.5 Chip*, which was part of the GoT2D dataset. This scaffold was extracted from GoT2D data, where remaining markers were masked for subsequent calculation of accuracy (squared Pearson correlation coefficient, r^2) at corresponding sites after imputation. Several reference panels were available, which were imputed into the same scaffold. Meta-imputation was applied to the imputed datasets obtained from split panels, which were generated as distinct subsets from the 1000G dataset. The intersection and union panels were separately imputed into the scaffold and subsequently compared to meta-imputed data on corresponding variant sets.

of genotypes to allelic dosage, calculated as $d = 0p_0 + 1p_1 + 2p_2$ where $d \in \{0, 1, 2\}$ for masked genotypes or $0 \leq d \leq 2$ when calculated from imputed genotype probabilities. Note that the Pearson correlation coefficient is defined as the covariance divided by the product of the standard deviation (SD) of two random variables. This is problematic if SD = 0, which is the case when variant genotypes are imputed as being monomorphic. To compensate for this loss in precision towards lower frequencies, the coefficient was set to $r^2 = 0$ for monomorphic variants. Imputed and masked genotype data were sorted into minor allele frequency (MAF) bins, based on their population frequency (MAF in the GoT2D dataset). In the following, accuracy is reported as mean r^2 calculated at corresponding variants per MAF bin.

2.4.2 Results

Accuracy of meta-imputed genotypes was explored for each combination of score metric and merge operation. The best performing setting was then chosen for comparison to direct imputations, as well as further analysis in Section 2.5 (page 62).

2.4.2.1 Comparison of meta-imputation settings

Each combination of score metric and merge operation produced an identical set of variants; that is, the combined set of variants across imputed panels. In total, 253,852 variants were returned from each meta-imputation in Scenario A (European subpopulations) and 559,172 in Scenario B (continental populations); *i.e.* the same number as captured by the union panel. Meta-imputed datasets were further reduced to the set of variants that matched to masked variants in the original GoT2D dataset. Variants contained in the genotype scaffold were removed, as these were not imputed. This retained 181,561 and 196,300 variants in Scenarios A and B, respectively.

First, I report the impact of quality control (QC) to characterise the different meta-imputation settings by the number of retained sites. Briefly, recall that QC was carried out to remove variant sites at `IMPUTE2` info-score < 0.4 and at deviations from HWE at p-value $< 1 \times 10^{-4}$. I then report genotype accuracy measured on sites retained after QC for each setting.

Table 2.2: Variants retained after quality control per meta-imputation setting. The number of variants retained after quality control (QC), n, per meta-imputation setting (combination of score metric and merge operation) in Scenario A and B. The percentage is given relative to the set of sites matched to masked variants in the GoT2D dataset and after removing sites contained in the imputation scaffold; 181,561 and 196,300 in A and B, respectively. Variants were removed at `IMPUTE2` info-score < 0.4 and at deviations from HWE at p-value $< 1 \times 10^{-4}$.

Merge	Score	Scenario A		Scenario B	
		n retained	(%)	n retained	(%)
`MSS`	`MP`	168,595	(92.9)	178,034	(90.7)
	`IS`	169,455	(93.3)	179,677	(91.5)
	`SC`	168,686	(92.9)	179,449	(91.4)
	`RS`	165,877	(91.4)	171,517	(87.4)
`WLC`	`MP`	161,079	(88.7)	166,458	(84.8)
	`IS`	162,511	(89.5)	169,860	(86.5)
	`SC`	160,464	(88.4)	165,907	(84.5)
	`RS`	160,369	(88.3)	165,787	(84.5)

The number of variants retained after QC differed among meta-imputation settings; see Table 2.2 (this page). Merge operations had a higher impact on the quality of meta-imputed genotypes than score metrics. In Scenario A, on average 92.6 % (±0.431 % SE) of variants were retained when `MSS` (maximum score selection) was used as the merge operation, with fewer retained using `WLC` (weighted linear combination), where 88.7 % (±0.272 % SE) were retained on average. This was similar in Scenario B, where 90.3 % (±0.977 % SE) and

85.1 % (±0.491 % SE) were retained on average under MSS and WLC, respectively. Most of the variants removed in either setting were low in frequency. For instance at MAF ≤ 1%, 74.6 % (±0.705 % SE) and 68.3 % (±0.495 % SE) passed QC in Scenario A when using MSS and WLC, respectively, as well as 72.2 % (±1.73 % SE) and 61.7 % (±0.962 % SE) in Scenario B, respectively. Among score metrics, the number of variants that passed QC was lowest for RS (random scores) in each comparison; for example, 67.5 % and 60.5 % at MAF ≤ 1% in A and B, respectively.

Although MSS overall preserved a relatively large proportion of markers after QC, the accuracy of retained genotypes was overall lower compared to data produced under WLC. Imputation accuracy improved after QC as illustrated in Figure 2.5 (next page), which shows mean r^2 calculated in MAF bins of equal size on log-scale. The differences among settings were small, in particular among score metrics when WLC was used, but where differences in accuracy become more pronounced after QC, which highlighted a clear distinction between merge operations. Throughout, mean r^2 was higher for genotype data produced under WLC. In Scenario A, for example, mean r^2 at MAF ≤ 1% before QC was 0.472 ($\pm 0.886\times10^{-3}$ SE) in WLC and 0.464 ($\pm 0.894\times10^{-3}$ SE) in MSS, but showed a larger difference after QC, namely 0.605 ($\pm 1.01\times10^{-3}$ SE) and 0.472 ($\pm 0.886\times10^{-3}$ SE) in WLC and MSS, respectively. This was also seen in Scenario B, where mean r^2 at MAF ≤ 1% was 0.428 ($\pm 0.796\times10^{-3}$ SE) and 0.418 ($\pm 0.811\times10^{-3}$ SE) before QC in WLC and MSS, respectively, as well as 0.600 ($\pm 0.951\times10^{-3}$ SE) in WLC and 0.548 ($\pm 0.914\times10^{-3}$ SE) in MSS after QC. Accuracy differences between merge operations were more pronounced at higher MAF; as seen in Figure 2.5. For example, at MAF ≤ 5% after QC, mean r^2 was 0.873 ($\pm 0.442\times10^{-3}$ SE) and 0.811 ($\pm 0.580\times10^{-3}$ SE) in Scenario A for WLC and MSS, respectively, as well as 0.862 ($\pm 0.472\times10^{-3}$ SE) and 0.793 ($\pm 0.649\times10^{-3}$ SE) in Scenario B, respectively.

Accuracy as measured for each setting after QC is given in Table 2.3 (page 55), which shows mean r^2 computed in three broader MAF bins to summarise accuracy levels at rare variants (here defined at MAF $\in [0.00, 0.01]$), low-frequency (MAF $\in (0.01, 0.05]$), and common variants (MAF $\in (0.05, 0.50]$). The RS score metric overall resulted in less accurate genotype data compared to other metrics, in particular in Scenario B where RS was least accurate in all comparisons. This was not the case in Scenario A, where it showed a higher accuracy than IS and SC at rare and low-frequency variants. However, note that accuracy differences among score metrics were low overall in Scenario A (see

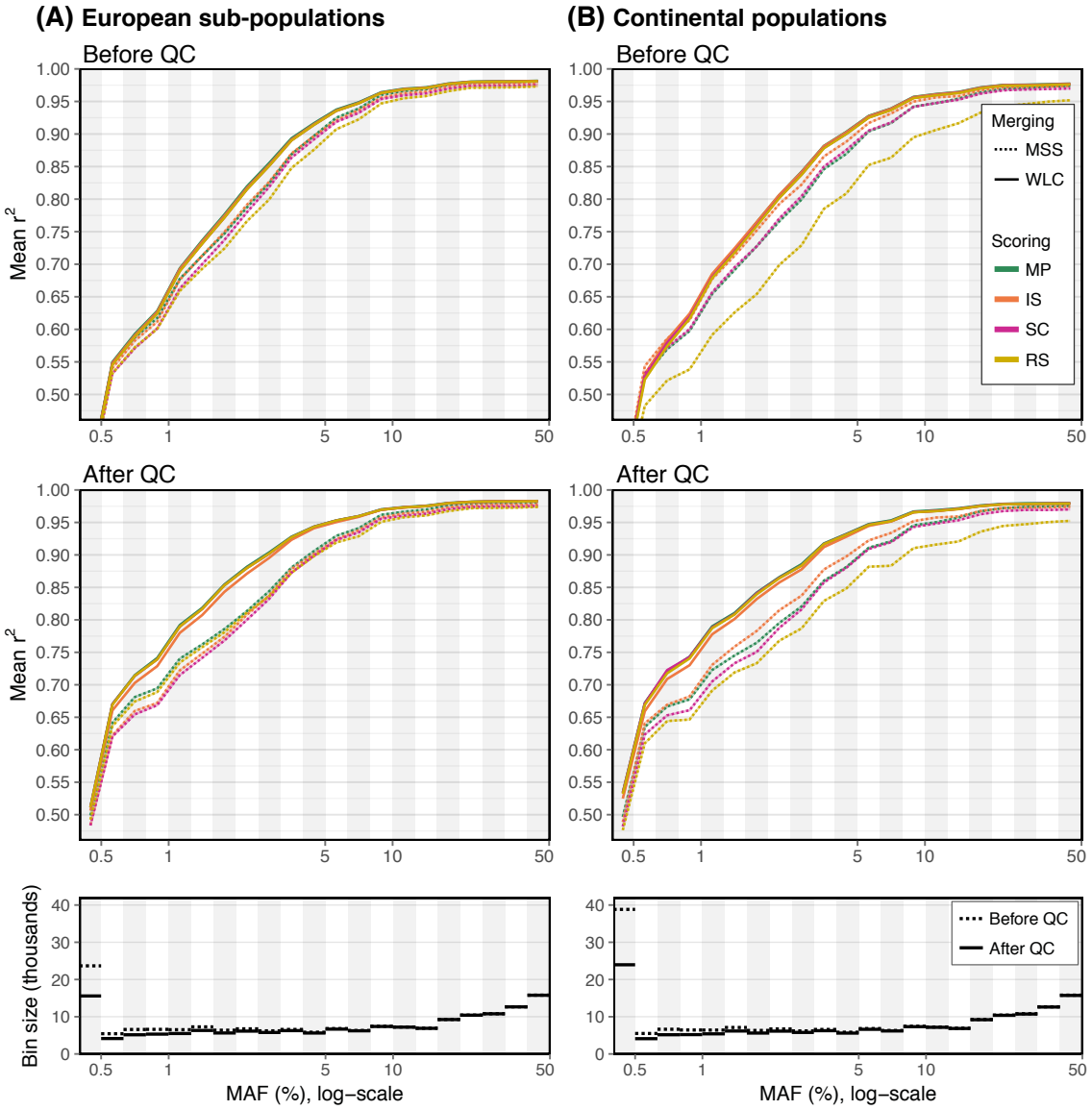


Figure 2.5: Accuracy comparison of score metrics and merge operations in meta-imputation. Each combination of merge operation (`MSS` and `WLC`) and score metric (`MP`, `IS`, `SC`, and `RS`) was examined in Scenarios A and B. Accuracy was measured as mean r^2 calculated between meta-imputed variants and variants masked in the GoT2D dataset. Results are shown both before and after QC. Bin sizes were defined on log-scale where *grey-white* bars indicate boundaries. The panels at the bottom indicate the number of variants per bin before QC (*dotted*) and the average number of variants per bin after QC (*solid*).

Table 2.3), due to the presumed higher genetic similarity between sample individuals and reference haplotypes (recall that the GoT2D sample is composed of individuals of Central and Northern European descent).

Regardless, `MP` (maximum probability) outperformed other score metrics in most comparisons; except in Scenario B, for low-frequency variants under `MSS`, where it was outperformed by `IS`. The `MP` score metric was found to further improve accuracy under

Table 2.3: Accuracy measured for each meta-imputation setting. Accuracy was measured as mean r^2 (±SE) per MAF bin; defined to reflect average levels of accuracy measured at rare, low-frequency, and common variants. Reported values were measured after QC for each meta-imputation setting (combination of merge operation and score metric), in Scenarios A and B. Retained variants were intersected across datasets per MAF bin. The number of SNPs at which mean r^2 was measured, n, is reported per scenario below the MAF range. The setting with the highest accuracy per MAF bin and per scenario is highlighted (**bold**).

MAF bin	Merge	Score	Scenario A	Scenario B
			Mean r^2(± SE*)	Mean r^2(± SE*)
[0.00, 0.01]	MSS	MP	**0.611** (2.057)	0.602 (1.997)
n_A = 28,341		IS	0.602 (2.068)	0.599 (2.005)
n_B = 33,443		SC	0.600 (2.041)	0.595 (1.961)
		RS	0.601 (2.052)	0.578 (1.977)
	WLC	MP	0.609 (2.041)	**0.603** (1.957)
		IS	0.608 (2.043)	0.603 (1.959)
		SC	0.608 (2.039)	0.601 (1.951)
		RS	0.608 (2.039)	0.601 (1.952)
(0.01, 0.05]	MSS	MP	0.871 (0.953)	0.858 (1.076)
n_A = 38,669		IS	0.859 (1.008)	0.860 (1.072)
n_B = 37,364		SC	0.855 (0.969)	0.846 (1.062)
		RS	0.847 (1.035)	0.807 (1.267)
	WLC	MP	**0.879** (0.867)	**0.867** (0.957)
		IS	0.876 (0.878)	0.866 (0.962)
		SC	0.876 (0.877)	0.864 (0.962)
		RS	0.876 (0.877)	0.863 (0.963)
(0.05, 0.50]	MSS	MP	0.975 (0.213)	0.969 (0.261)
n_A = 92,797		IS	0.971 (0.233)	0.970 (0.262)
n_B = 91,375		SC	0.969 (0.234)	0.965 (0.254)
		RS	0.965 (0.263)	0.940 (0.418)
	WLC	MP	**0.977** (0.175)	**0.974** (0.184)
		IS	0.976 (0.178)	0.973 (0.185)
		SC	0.976 (0.179)	0.973 (0.187)
		RS	0.976 (0.179)	0.972 (0.188)

* Standard error (SE) $\times 10^{-3}$

WLC, such that the combination of MP and WLC was seen to yield the highest accuracy in each MAF bin and in both scenarios (as highlighted in Table 2.3). Therefore, in the following, WLC was chosen as merge operation and MP as score metric; hence, the combination of MP and WLC is implied when referring to meta-imputation below.

2.4.2.2 Improvements of accuracy in comparison to direct imputations

Available split panels were imputed into the generated study sample and imputed genotype data were then combined through meta-imputation. The union and intersection panels were separately imputed for subsequent comparison to meta-imputed genotypes.

Before accuracy was measured, all data were subjected to QC and variants were removed when not matched to masked variants or when contained in the imputation scaffold. For simplicity, imputed datasets are referred to by the panel from which they were estimated.

Comparisons were based on mean r^2 calculated at corresponding (meta-)imputed and masked variants pooled by MAF bin. In addition, significant differences in the MAF distribution of imputed and corresponding meta-imputed variants were determined using the two-sample Kolmogorov—Smirnov (KS) test. However, significance was determined from the median of the KS test statistic, here denoted by $\widetilde{D}$, calculated at $n = 500$ randomly selected sites over 1,000 repeated draws. This was done to account for varying subset sizes retained in each comparison, and to avoid potential biases due to correlations of linkage disequilibrium (LD) at nearby markers. MAF distributions were significantly different if

$$\widetilde{D}_n \; > \; c(\alpha)\sqrt{\frac{2n}{n^2}} \tag{2.10}$$

for significance levels $c(0.05) = 1.36$ and $c(0.01) = 1.63$. A similar approach was applied by Pasaniuc *et al.* (2014) to compare signatures of functional enrichment in imputed data.

Table 2.4: Effect of quality control on imputed genotype data. The number (percent) of variants retained after QC for direct imputations (*i.e.* four split panels, intersection panel, and union panel) and meta-imputation. Numbers refer to variants retained after removing unmatched sites and those contained in the imputation scaffold.

Panel	Scenario A			Scenario B		
	Split	*n* retained	(%)	Split	*n* retained	(%)
Split panel (1)	CEU	135,218	(95.4)	AFR	123,662	(91.8)
Split panel (2)	FIN	141,017	(96.6)	AMR	155,266	(93.6)
Split panel (3)	GBR	137,277	(95.0)	ASN	99,531	(94.3)
Split panel (4)	TSI	138,613	(94.0)	EUR	161,364	(95.3)
Meta-imputed (1–4)	–	161,079	(88.7)	–	166,458	(84.8)
Intersection panel	–	116,980	(99.8)	–	92,312	(99.8)
Union panel	–	174,229	(96.0)	–	184,158	(93.8)

The numbers of retained variants for each panel are given in Table 2.4 (this page). Meta-imputed data showed the highest proportion of variants removed through QC. In Scenario A, 11.3 % of meta-imputed variants were removed, whereas only 0.197 % of variants in the intersection and 4.04 % in the union panel were removed, compared to an average of 4.74 % (±0.536 % SE) among split panels. Note that only 3.40 % of markers did not pass QC after imputation from the FIN sub-population. The proportion of meta-imputed genotypes removed after QC was also highest in Scenario B (15.2 %) which is

compared to only 0.163 % in the intersection and 6.19 % in the union panel, as well as 6.23 % (±0.735 % SE) on average in split panels, where the lowest proportion of removed variants was seen for the EUR panel (4.66 %). However, the number of retained variants in meta-imputed data (161,076 and 166,458 in A and B, respectively) exceeded those retained in any split panel or the intersection panel; see Table 2.4.

Each of the imputed datasets was compared separately to meta-imputation, on the same set of variants retained after QC. The distribution of accuracy (mean r^2) measured by MAF is shown in Figure 2.6 (next page); average accuracy measured for each imputation strategy in comparison to meta-imputation is given in Table 2.5 (page 59), where accuracy was measured by MAF to distinguish rare variants (MAF $\in [0.00, 0.01]$), low-frequency (MAF $\in (0.01, 0.05]$), and common variants (MAF $\in (0.05, 0.50]$).

In Scenario A, meta-imputation showed an improvement in accuracy over imputations from split panels. For example, for rare variants, the highest improvement among split panel comparisons was seen with the GBR sample, where mean r^2 was 0.637 ($\pm 3.37\times10^{-3}$ SE) for GBR and 0.659 ($\pm 3.30\times10^{-3}$ SE) for meta-imputed data. Differences were larger at low-frequency, where the highest improvement was seen in comparison with the TSI sample; 0.865 ($\pm 1.16\times10^{-3}$ SE) and 0.907 ($\pm 0.808\times10^{-3}$ SE) for TSI and meta-imputation, respectively. Only the union panel was higher in accuracy than meta-imputation; *e.g.* 0.697 ($\pm 1.87\times10^{-3}$ SE) and 0.627 ($\pm 2.00\times10^{-3}$ SE) for rare variants, respectively, and 0.893 ($\pm 0.801\times10^{-3}$ SE) and 0.877 ($\pm 0.859\times10^{-3}$ SE) at low-frequency variants, respectively. Meta-imputation showed approximately equal levels of accuracy as the union panel at common variants, where the difference in mean r^2 was 0.00318 ($\pm 0.811\times10^{-4}$ SE).

Genotype accuracy showed higher differences in Scenario B, where meta-imputation improved accuracy in most split panel comparisons. For rare variants, the highest difference was seen to genotype data imputed from the AFR split panel, where mean r^2 was 0.703 ($\pm 3.24\times10^{-3}$ SE), compared to 0.745 ($\pm 3.07\times10^{-3}$ SE) for meta-imputed genotypes. However, meta-imputation showed similar accuracy as the imputation from the EUR sample, where the difference in mean r^2 was 0.00208 ($\pm 0.716\times10^{-4}$ SE). Likewise, at low-frequency, mean r^2 was 0.879 ($\pm 1.63\times10^{-3}$ SE) for AFR and 0.948 ($\pm 0.767\times10^{-3}$ SE) for meta-imputation, and the difference in accuracy was 0.00249 ($\pm 0.367\times10^{-3}$ SE) with regard to the EUR split panel. As in Scenario A, differences were smaller for common variants, such that the difference in mean r^2 was below 0.001 in comparisons to imputations from the AFR, AMR, and EUR panels, but where the ASN sample

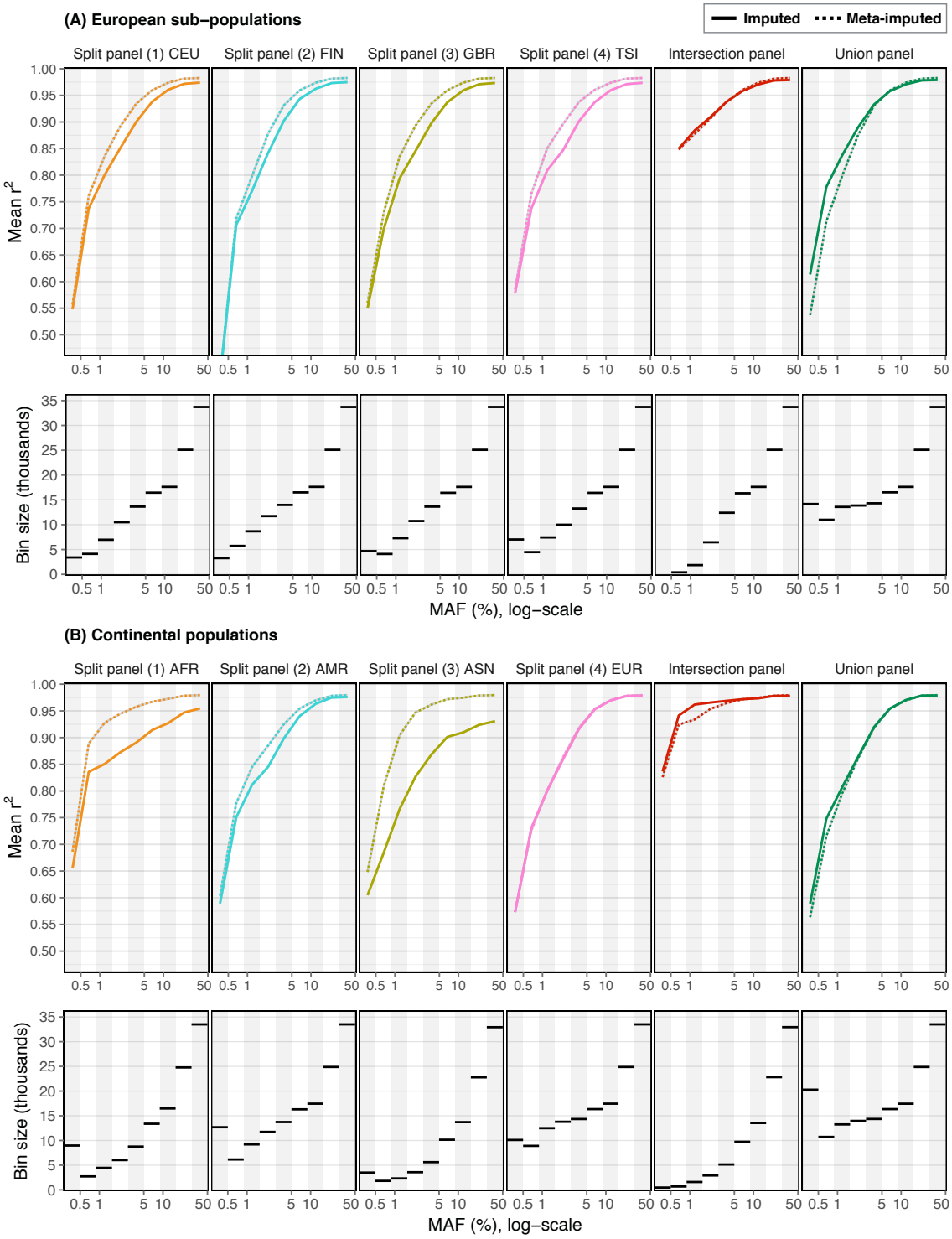


Figure 2.6: Accuracy comparison between meta-imputation and direct imputations. Accuracy was measured as mean r^2 per MAF bin, defined on log-scale where *grey-white* bars indicate boundaries. Each imputed panel (imputations from the four split panels, the intersection panel, and the union panel) was separately compared to meta-imputation on the same set of variants per bin (*i.e.* on the intersected set of SNPs per comparison); shown for variants retained after QC, in Scenarios A and B. MAF bins were defined on the actual allele frequencies as determined by the GoT2D dataset. Note that mean r^2 is not shown if the number of markers dropped below 50 per MAF bin. Panels at the bottom show the number of variants compared per bin.

Table 2.5: Accuracy of imputation strategies at rare, low-frequency, and common variants. Accuracy was calculated as mean r^2 per MAF bin on the same set of variants retained after QC in each comparison between meta-imputation and direct imputation, where n denotes the number of variants compared. The imputation strategy with the highest accuracy is highlighted (**bold**). The median of KS test statistic, $\widetilde{D}_{500}$, determined whether imputed and meta-imputed MAF distributions were significantly different; see Equation (2.10) on page 56.

MAF bin	Panel	n	Imputation		Meta-imputation		KS test†
			Mean r^2	(± SE‡)	Mean r^2	(± SE‡)	$\widetilde{D}_{500}$
Scenario A	(European sub-populations)						
[0.00, 0.01]	Split panel, CEU	8,636	0.665	(3.491)	**0.683**	(3.402)	0.046
	Split panel, FIN	10,416	0.619	(3.292)	**0.630**	(3.255)	0.024
	Split panel, GBR	10,023	0.637	(3.371)	**0.659**	(3.296)	0.034
	Split panel, TSI	12,763	0.654	(2.974)	**0.670**	(2.909)	0.098*
	Intersection panel	546	0.823	(9.901)	**0.824**	(9.832)	0.028
	Union panel	27,712	**0.697**	(1.873)	0.627	(2.001)	0.264**
(0.01, 0.05]	Split panel, CEU	30,012	0.866	(1.092)	**0.902**	(0.813)	0.040
	Split panel, FIN	32,969	0.853	(1.076)	**0.885**	(0.887)	0.032
	Split panel, GBR	30,442	0.860	(1.119)	**0.901**	(0.813)	0.036
	Split panel, TSI	29,445	0.865	(1.164)	**0.907**	(0.808)	0.036
	Intersection panel	20,604	**0.925**	(0.850)	0.923	(0.803)	0.034
	Union panel	39,213	**0.893**	(0.801)	0.877	(0.859)	0.088*
(0.05, 0.50]	Split panel, CEU	92,885	0.964	(0.269)	**0.977**	(0.180)	0.014
	Split panel, FIN	92,936	0.966	(0.250)	**0.977**	(0.181)	0.012
	Split panel, GBR	92,845	0.964	(0.273)	**0.977**	(0.181)	0.012
	Split panel, TSI	92,840	0.964	(0.283)	**0.977**	(0.180)	0.012
	Intersection panel	92,751	0.973	(0.212)	**0.977**	(0.180)	0.012
	Union panel	92,938	0.973	(0.212)	**0.977**	(0.181)	0.012
Scenario B	(Continental populations)						
[0.00, 0.01]	Split panel, AFR	12,495	0.703	(3.238)	**0.745**	(3.070)	0.030
	Split panel, AMR	20,416	0.653	(2.480)	**0.672**	(2.388)	0.040
	Split panel, ASN	5,661	0.640	(4.972)	**0.714**	(4.586)	0.082
	Split panel, EUR	21,223	**0.658**	(2.227)	0.656	(2.207)	0.048
	Intersection panel	1,364	**0.907**	(4.420)	0.892	(4.430)	0.172**
	Union panel	33,430	**0.653**	(1.871)	0.626	(1.894)	0.148**
(0.01, 0.05]	Split panel, AFR	18,468	0.879	(1.631)	**0.948**	(0.767)	0.048
	Split panel, AMR	33,093	0.861	(1.158)	**0.894**	(0.855)	0.036
	Split panel, ASN	11,181	0.837	(2.414)	**0.948**	(0.954)	0.114**
	Split panel, EUR	38,417	**0.867**	(0.969)	0.870	(0.910)	0.032
	Intersection panel	9,443	**0.967**	(0.811)	0.957	(0.816)	0.062
	Union panel	39,140	**0.871**	(0.949)	0.866	(0.924)	0.058
(0.05, 0.50]	Split panel, AFR	88,152	0.941	(0.426)	**0.976**	(0.172)	0.022
	Split panel, AMR	92,144	0.967	(0.263)	**0.973**	(0.191)	0.016
	Split panel, ASN	79,581	0.921	(0.519)	**0.978**	(0.165)	0.024
	Split panel, EUR	92,188	0.972	(0.218)	**0.973**	(0.193)	0.016
	Intersection panel	79,051	0.976	(0.201)	**0.977**	(0.168)	0.016
	Union panel	92,187	**0.973**	(0.214)	0.973	(0.193)	0.016

† Median of the Kolmogorov–Smirnov (KS) test statistic, $\widetilde{D}$; empirical CDF of imputed and meta-imputed MAF tested at $\alpha = 0.05$ (*) and $\alpha = 0.01$ (**).

‡ Standard error (SE) $\times 10^{-3}$.

showed the highest difference; mean r^2 was 0.921 ($\pm 0.519\times10^{-3}$ SE) for ASN and 0.978 ($\pm 0.165\times10^{-3}$ SE) for meta-imputation. The union panel was similar in accuracy as meta-imputation, where the overall difference in mean r^2 was 0.0106 ($\pm 8.47\times10^{-3}$ SE).

Imputations from the intersection panel in Scenario A and B showed approximately equal levels of accuracy to meta-imputation. The difference in mean r^2 averaged to 0.000811 ($\pm 1.43\times10^{-4}$ SE) across MAF in Scenario A, and 0.00824 ($\pm 4.82\times10^{-3}$ SE) in Scenario B. However, note that the number of variants in the intersection panel was the lowest among available panel data in both scenarios (Table 2.4), and was further reduced as accuracy was measured on the same sets of variants retained in both the intersection and the meta-imputed datasets. For example, the comparison between the intersection panel and meta-imputation included only 546 variants at MAF $\leq$ 1% in Scenario A and 1,364 variants in Scenario B, whereas each split panel and the union panel were compared on several thousands of variants at this frequency range. The high accuracy of genotypes imputed from the intersection panel may result from retaining only those variants that are "cosmopolitan" within the scope of the present evaluation.

Further, the empirical cumulative distribution function (CDF) of MAF at imputed and meta-imputed variants was compared per MAF bin. Differences are illustrated in Figure 2.7 (next page), which shows the CDF of compared variants in relation to the known population frequencies at masked variants in the GoT2D dataset; calculated by subtracting (meta-)imputed frequencies from masked frequencies (ΔMAF) at the same set of markers. Notably, meta-imputed frequencies showed high consistency with imputed frequencies at rare variants (MAF $\in [0.00, 0.01]$) across split panel imputations, but were skewed in comparison to imputations from the union panel in both scenarios. Significant differences were found for rare variant imputations from the TSI sample ($\widetilde{D} = 0.098$) in Scenario A, as well as for the union panel at rare and low-frequency variants (0.264 and 0.088, respectively). In Scenario B, imputed and meta-imputed differences were significantly different for rare variant imputations from the intersection panel and the union panel (0.172 and 0.148, respectively) and for the union panel at low-frequency variants (0.114). These results suggested that meta-imputation was able to correctly reproduce realistic allele frequency distributions from the combination of imputed genotypes from different sources, while achieving higher or similar accuracy compared to direct imputations from split panels. Results of KS tests in each comparison are given in Table 2.5 (page 59).

In summary, split panel imputations were either outperformed or similar levels of accuracy were achieved in direct comparisons to meta-imputed data; see Table 2.5 for a complete summary of genotype accuracy measured in each comparison. Although

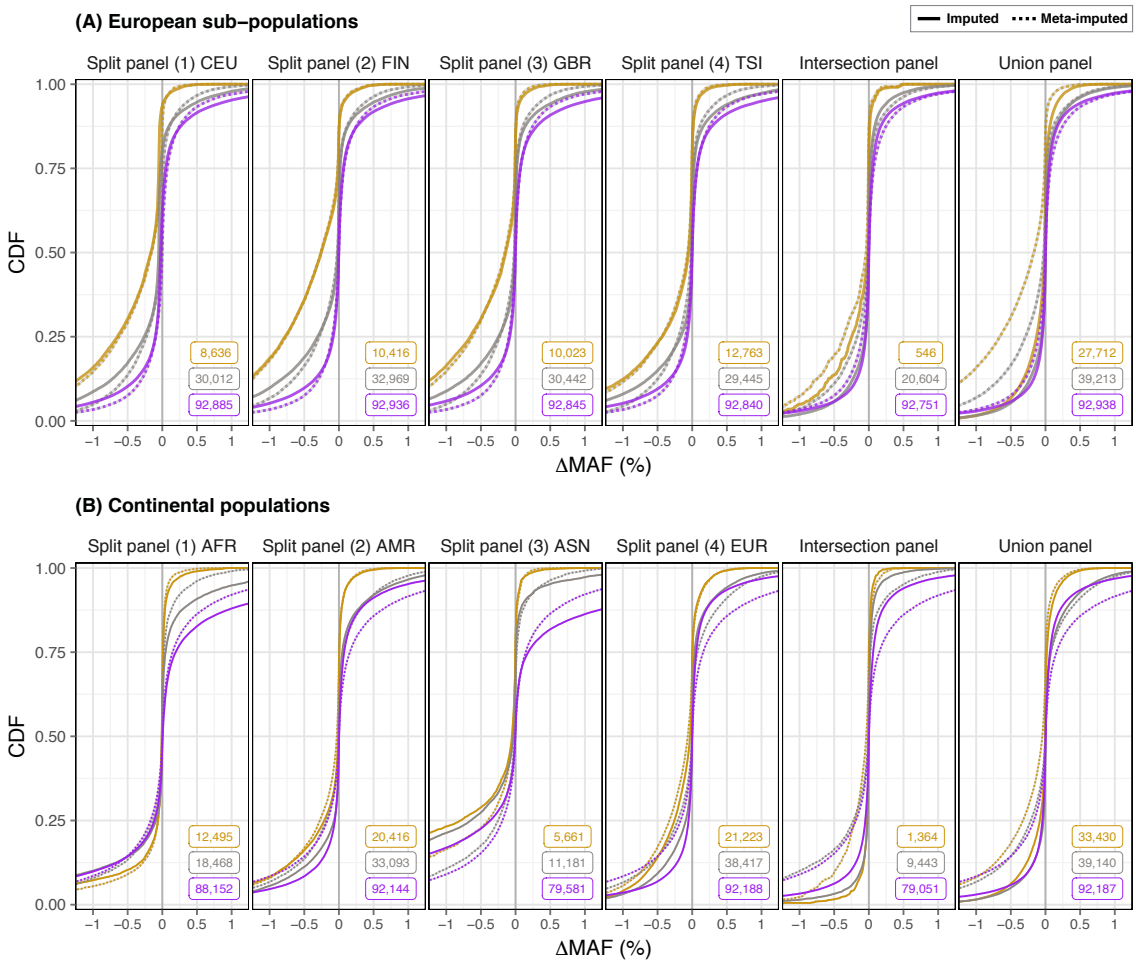


Figure 2.7: Difference between imputed and masked minor allele frequency. Comparison of imputed and meta-imputed MAF in relation to known population frequencies, compared on the same set as retained after QC in each comparison. Frequency difference, ΔMAF, was calculated as the MAF observed at a masked variant minus MAF at the corresponding (meta-)imputed variant, pooled in three MAF bins; rare variants (MAF $\in [0.00, 0.01]$; *yellow*), low-frequency (MAF $\in (0.01, 0.05]$; *grey*), and common variants (MAF $\in (0.05, 0.50]$; *purple*). Numbers per MAF bin per comparison are given in each panel (*colour-coded*).

imputations from the union panel outperformed meta-imputation, such differences may be expected given that the union panel contained all the information which meta-imputation had to leverage indirectly from several data sources. Nonetheless, the present evaluation of genotype accuracy was limited with regard to coverage; for instance, genotype data imputed from the intersection panel was found to be relatively high in accuracy and similar with regard to meta-imputed data, but the low number of variants present in the intersection panel may not yield similar improvements under realistic conditions in association analyses. Therefore, to provide a comprehensive assessment of the meta-imputation method and to account for a potential tradeoff between accuracy and coverage, I conducted a more extensive power analysis in the following section.

2.5 Power to detect significant risk signals

The power of meta-imputation to detect disease risk factors in association tests was evaluated using simulated sample data. This was done in consideration of expected power when causal risk factors vary in their allele frequency as well as risk effect size. In particular, a series of simulated case-control association experiments was conducted, from which the power to detect significant association signals was determined, at specified allele frequencies and effect size of simulated risk factors. The description of the methods used is provided below (Section 2.5.1, this page), followed by the presentation of results (Section 2.5.2, page 66).

2.5.1 Methods

The same regime to carry out imputation and QC was followed as described in Section 2.4.1 (page 49). An additional set of haplotype reference data was available from four independent sequencing studies, which were included here as Scenario C; see below.

Finns. A Finnish cohort composed of data from the Sequencing Initiative Suomi Project (*SISu*) and the *Finrisk* Project (Vartiainen *et al.*, 2010; Pajunen *et al.*, 2010; Lim *et al.*, 2014; Borodulin *et al.*, 2015); 4x depth; sample size and number of SNPs considered here were $N = 1{,}941$ and $M = 283{,}654$, respectively.

GoNL. The Genome of the Netherlands Project (Boomsma *et al.*, 2013; Deelen *et al.*, 2014; Genome of the Netherlands Consortium, 2014); 12x depth, consisting of a representative sample of 250 trio-families; $N = 748$, $M = 362{,}694$.

ORCADES. The Orkney Complex Disease Study of genetic epidemiology of an isolated population in northern Scotland (McQuillan *et al.*, 2008); 4x depth, family-based data; $N = 399$, $M = 236{,}755$.

UK10K. The *UK10K* Genome Sequencing Project (UK10K Consortium *et al.*, 2015); 6.5x depth; $N = 3{,}642$, $M = 527{,}199$.

Also, an intersection panel was prepared from these four datasets, but no union panel. As before, only data from chromosome 20 were considered. Note that the above datasets were part of the early stage HRC testing phase (McCarthy *et al.*, 2016).*

* Acknowledgement: Data provided by Professor Jonathan Marchini, Department of Statistics, University of Oxford; prior to the release of the HRC dataset.

2.5.1.1 Simulation of study sample data

Simulations were performed using HAPGEN version 2.2.0 (Su *et al.*, 2011), which requires a *template* dataset of haplotypes to reproduce realistic variant data in HWE, such that LD patterns in the simulated dataset are consistent with the haplotype sample. Individual sites can be simulated to independently act as causal disease variants with specified relative risk. The simulation generates two GWA samples of individuals that are affected (*cases*) or not affected (*controls*) by a disease phenotype. Data are identical in coverage as the template dataset.

Here, simulations were performed using GoT2D data (chromosome 20) to serve as the template dataset. The size of simulated case and control samples was fixed to 2,500 individuals each. Although a larger sample would have been beneficial in terms of signal detection through association testing, exceeding the size of the template dataset (N = 2,657) was expected to result in systematically biased allele frequency changes. For example, an iterative re-sampling strategy could be applied to introduce new low-frequency variants (*e.g.* following Moutsianas *et al.*, 2015). However, this was not done here because the effect size of risk variants (as defined during simulation) would likely be affected by such a sampling process.

A series of simulation experiments was conducted in which one variant was selected per simulation to act as a causal risk factor. Its relative risk (RR) was defined for heterozygous genotypes (RR_{het}) in a log-additive disease model (*i.e.* multiplicative on linear scale); the following three risk categories were defined.

$$\begin{aligned} \textit{Low risk}: &\quad RR_{het} = 1.2 \quad (RR_{hom} = 1.44) \\ \textit{Modest risk}: &\quad RR_{het} = 1.6 \quad (RR_{hom} = 2.56) \\ \textit{High risk}: &\quad RR_{het} = 2.0 \quad (RR_{hom} = 4.00) \end{aligned}$$

The analysis was performed by conducting 300 replicate simulations per risk category, where variants occurring at different frequencies were selected in three defined MAF intervals; very low frequency (MAF $\in [0.5, 1]\,\%$), low frequency (MAF $\in (1, 5]\,\%$), and high frequency (MAF $\in (5, 50]\,\%$), such that 100 variants were drawn from each interval and simulated as risk variants.

Note that variant selection was done at random, regardless of presence or absence of the selected variant in any of the available reference panels, so as to mirror conditions encountered under realistic GWA settings; *i.e.* when a causal variant itself is absent in an imputation reference, its risk effects may be detectable through LD at neighbouring sites.

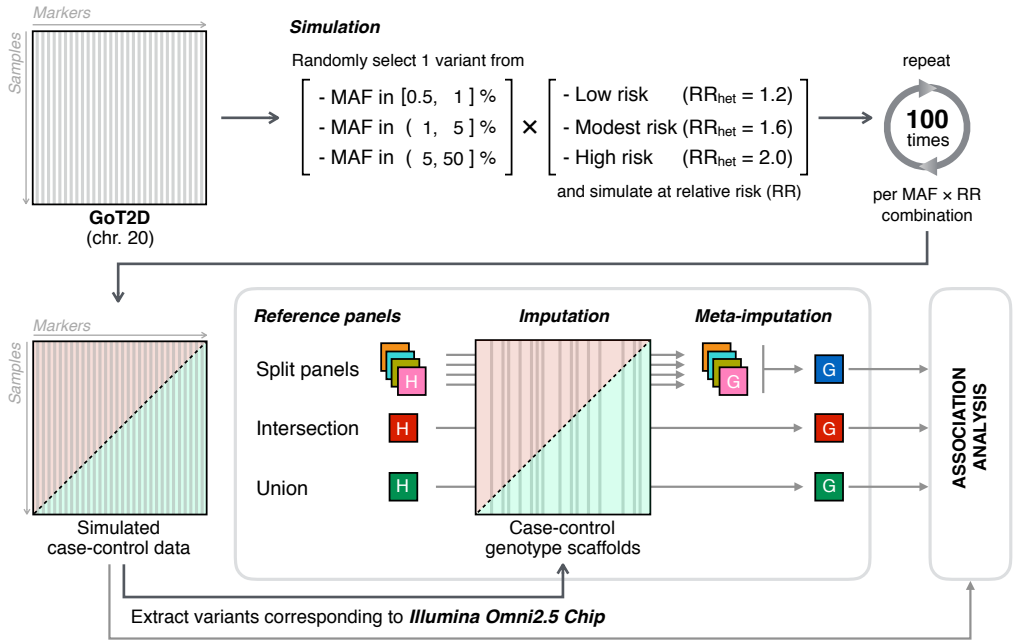


Figure 2.8: Illustration of the simulation process. Meta-imputation was assessed in terms of statistical power to detect significant risk association signals in a series of simulated case-control experiments. The GoT2D dataset was used as a template for simulations using HAPGEN (Su *et al.*, 2011), where one variant was randomly selected within one of three defined MAF intervals. The selected variant was then simulated to act as a causal disease variant in the simulated case-control dataset, where relative risk (RR_{het}) was defined according to one of three defined risk categories. In total, 100 replicate simulations were conducted per combination of MAF interval and risk category (per scenario). Simulated data were used to extract a genotype scaffold into which available reference panels were imputed, followed by meta-imputation of imputed datasets. Imputed and meta-imputed datasets were then subjected to association analysis, including the simulated (not imputed) datasets for comparison.

To generate a study sample for imputations, a variant scaffold was extracted from each simulation replicate. Because the set of simulated variants mirrored those in the GoT2D dataset, sites that matched with variants typed on *Illumina Omni2.5 Array* were identified and extracted. A scaffold thus contained 40,255 variants into which available reference panels were imputed. Note that simulations produced two datasets; one case and one corresponding control dataset. These were concatenated before imputation to ensure consistency in the imputation analysis. Imputed data were again separated into case and control samples prior to association analysis (described below). Because HAPGEN2 produces haplotype data, imputations were executed on pre-phased genotypes. A summary of the simulation process is illustrated in Figure 2.8 (this page).

2.5.1.2 Association analysis in imputed genotype data

Imputed case and control datasets were analysed using a frequentist score test under an additive model of association, implemented in SNPTEST version 2.5 (Marchini *et al.*, 2007). In contrast to the previous analysis (Section 2.4 on page 48), in which the variants not included in the extracted scaffold were masked to measure accuracy after imputation, here, the simulated case-control dataset was retained and separately examined in association analysis. This was done to enable comparisons of meta-imputed and imputed data to a non-imputed benchmark result for each simulation replicate.

The genomic control inflation factor, λ_{GC}, was calculated to investigate if systematic biases are present in association results, which is defined as the median of χ^2 test statistics resulting from case-control association tests divided by the expected median of the χ^2 distribution (Devlin *et al.*, 2001). Because the frequentist score test was used, λ_{GC} was calculated on basis of the resulting p-values from which the χ^2 statistic was calculated with one degree of freedom.

2.5.1.3 Calculation of power in replicate simulation experiments

Significant association signals were identified in each simulation and pooled by MAF interval and risk category, according to which variants were selected and simulated. The proportion of datasets in which significance was reached at the known risk variant was taken as a simple estimate for the statistical power to detect genetic risk effects. Note that the position of the simulated risk variant was known through simulation, but the variant itself may not be retained after imputation or QC. Therefore, signal detection was performed within a 1 Mb region around the position of the simulated risk variant, for any site reaching significance with this region.

Significance was defined at a nominal threshold of p-value $\leq 1 \times 10^{-6}$. Note that this threshold is higher (thus, less conservative) than commonly applied genome-wide thresholds, *e.g.* at 5×10^{-8} (*e.g.*, see Risch and Merikangas, 1996), because analyses were conducted on data from chromosome 20 only. However, to provide additional detail, power was estimated under a moving significance threshold; between p-value $\leq 1 \times 10^{-8}$ and p-value $\leq 1 \times 10^{-4}$. As a comparative measure between association results produced from the different imputation strategies, the difference in power between the non-imputed simulation dataset and a given (meta-)imputed dataset is reported, denoted by Δ_P, which is calculated as the average difference along the moving significance threshold.

2.5.2 Results

A number of 100 variants were selected per MAF interval such that there were 300 variants in total. Each was then simulated at the three defined risk categories such that 900 simulations were conducted from which a genotype scaffold was extracted for imputation. Given the four split panels, the intersection panel, and the union panel available per Scenario A and B, as well as the four independent reference datasets and the generated intersection panel in Scenario C, a total of 15,300 imputation analyses were performed. Imputed data were then combined in meta-imputation (except the intersection and union panels), resulting in 900 additional genotype datasets. Each dataset was then subjected to association analysis, including the non-imputed simulated case-control sample, which was used as a benchmark for comparisons. Hence, a total of 17,100 association analyses were conducted, where each was treated as an independent GWA study. All analyses were performed on whole-chromosome data (chromosome 20).

Association results were inspected with regard to inflation before and after QC; the difference is shown in Figure 2.9 (next page) where λ_{GC} is shown as the average per MAF interval. Inflation was slightly increased at higher risk variant frequencies. The difference of λ_{GC} before and after QC was small in Scenarios A and C, but noticeable in Scenario B, where association results of meta-imputed data were deflated ($\lambda_{GC} < 1$) before QC. After QC, λ_{GC} values of all imputed and meta-imputed datasets were approximately equal to inflation measured for the non-imputed simulation dataset in each scenario.

Association results were at $\lambda_{GC} \approx 1$ on average in each scenario when the simulated risk variant was very low in frequency (MAF $\in [0.5, 1]$ %), but increased to $\lambda_{GC} \approx 1.05$ for risk variants at higher frequencies (MAF $\in [5, 50]$ %). Although these results suggested no major inflation, higher values of λ_{GC} are generally expected in presence of population sub-structure, including cryptic relationships among individuals in the sample. One explanation why λ_{GC} increased at higher risk variant frequencies is that individuals in the case sample appeared to be more related to each other than the individuals in the control sample.

Association results for each imputation strategy (referring to results obtained on genotype data imputed from available reference panels and meta-imputation) were separately evaluated with regard to each combination of risk category and the MAF interval from which simulated risk variants were selected. The distribution of power measured under a moving significance threshold (between p-value $\leq 1 \times 10^{-8}$ and p-value $\leq 1 \times 10^{-4}$) is shown in Figure 2.10; for Scenario A (page 68), B (page 69),

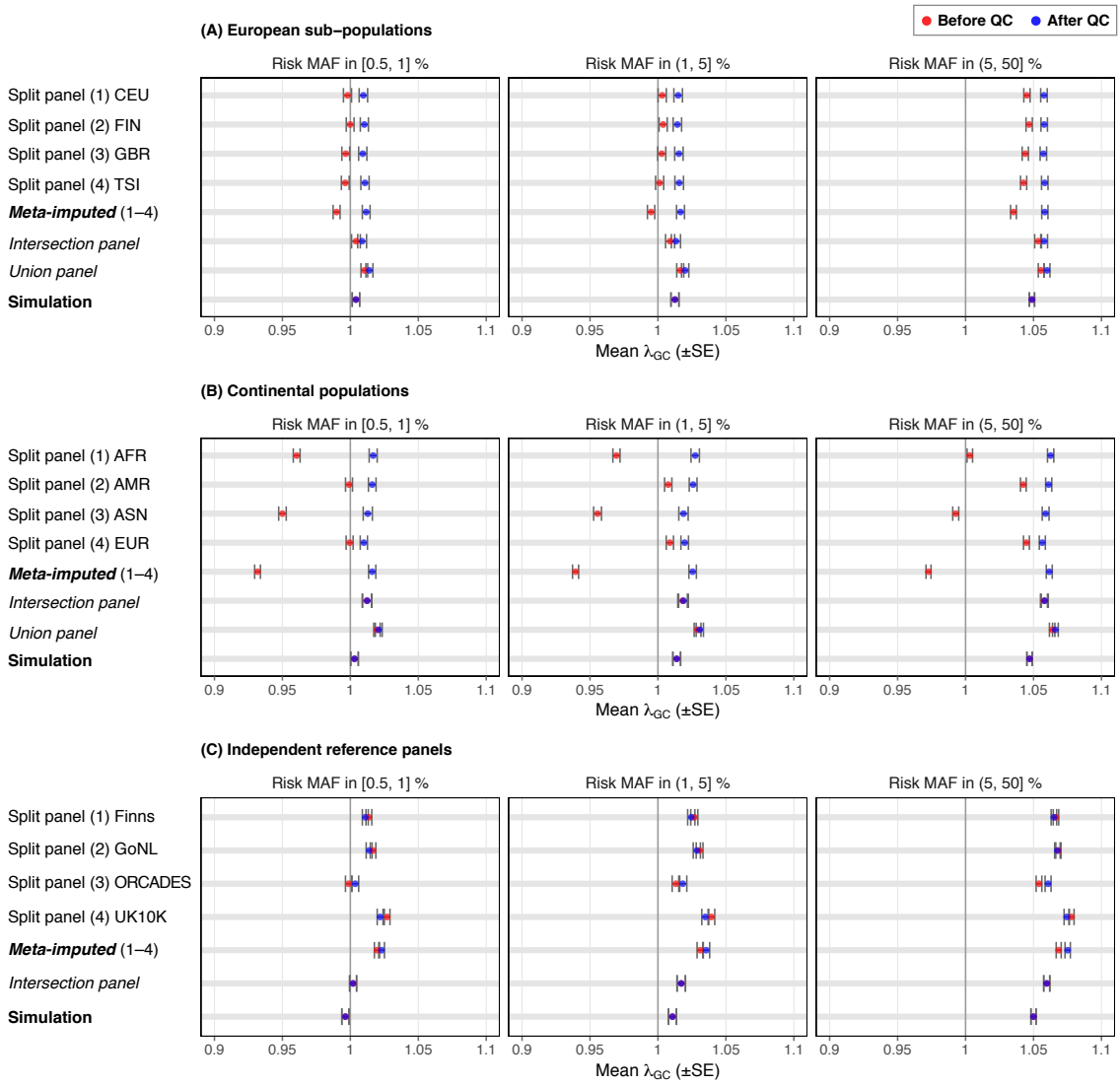


Figure 2.9: Inflation observed in simulated case-control experiments. Genomic control inflation factor calculated before (*red*) and after (*blue*) variants were filtered in QC, reported as mean λ_{GC} over replicate association results by MAF of the simulated risk variants.

and C (page 70). The results are summarised in Table 2.6, for power measured at the nominal significance threshold (p-value $\leq 1 \times 10^{-6}$) and the average difference to the non-imputed simulation benchmark (Δ_P) along the moving threshold, averaged per MAF interval of simulated risk variants; for Scenario A (page 71), B (page 72), and C (page 73).

The union panel was seen with the lowest average difference in power at very low frequencies of the simulated risk variant (MAF $\in [0.5, 1]$ %) in Scenario A, where Δ_P was 1.05 (±0.206 SE). Meta-imputation showed the lowest average difference at very low MAF in Scenario B, Δ_P = 1.23 % (±0.247 % SE), as well as Scenario C, Δ_P = 1.32 % (±0.248 % SE); but recall that Scenario C (independent reference panels) did not contain a union panel. However, even in the high risk category in each scenario,

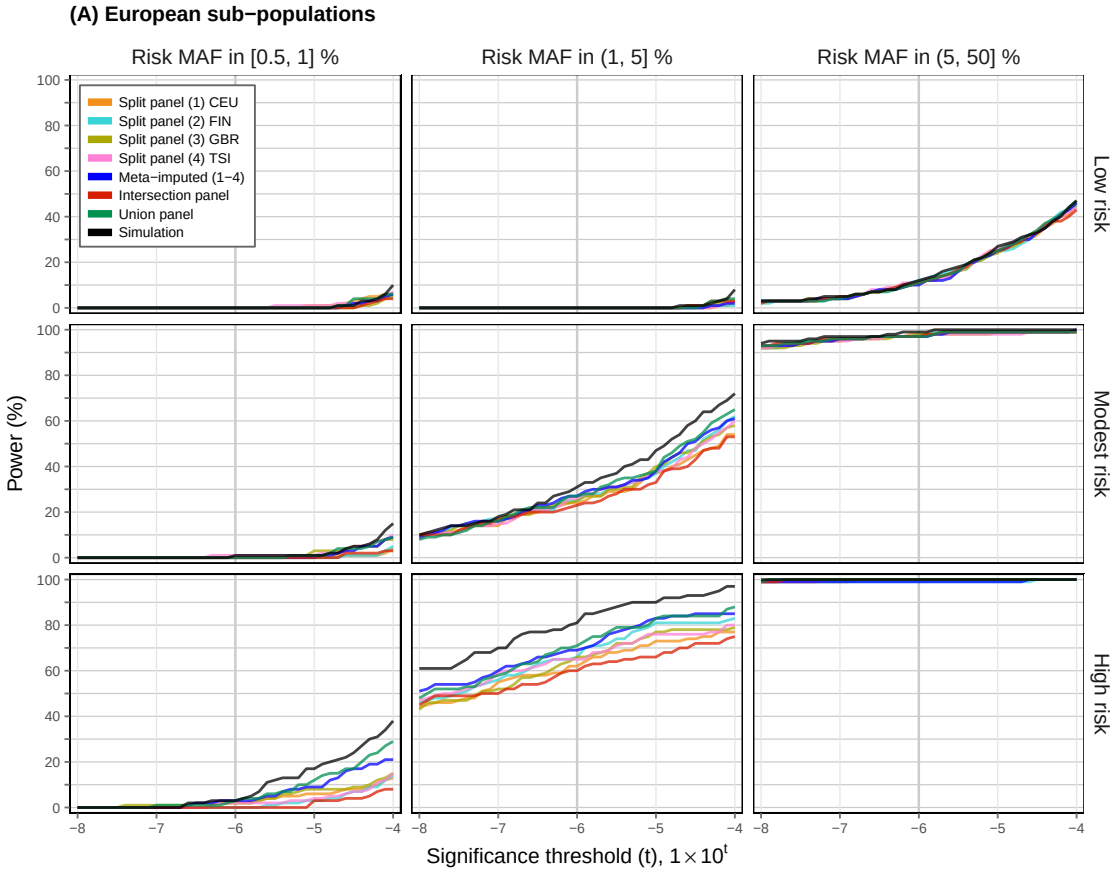


Figure 2.10: Power measured under a moving significance threshold. Power was calculated as the proportion of replicate association analyses ($n = 100$, per combination of risk category and MAF interval) in which any signal reached significance within 1 Mb around the position of a simulated risk variant. A moving significance threshold between p-value $\leq 1 \times 10^{-8}$ and p-value $\leq 1 \times 10^{-4}$ was applied to each association dataset.

estimated power did not exceed 3% for any imputation strategy when the simulated risk variant was very low in frequency, such that observed differences were negligible as these could be attributed to stochastic noise. Similarly, observed differences were small in each risk category when causal variants were selected from the high frequency interval (MAF $\in [5, 50]$ %), where the lowest average difference in power was recorded for the union panel in Scenario A, $\Delta_P = 0.537$ % (±0.067,6 % SE), the EUR split panel in B, 0.650 % (±0.072,1 % SE), and the intersection panel in C, 0.545 % (±0.091,8 % SE). However, note that $\Delta_P < 1\%$ in each strategy at high risk MAF in Scenarios A and C, but where some of the strategies showed larger differences in Scenario B, *e.g.* the ASN split panel and the intersection panel; 2.83 % (±0.177 % SE) and 2.56 % (±0.173 % SE), respectively.

Noticeable differences were seen among imputation strategies for simulated risk variants selected at low frequency (MAF $\in [1, 5]$ %). The union panel was recorded with the lowest difference in power relative to the non-imputed simulation benchmark

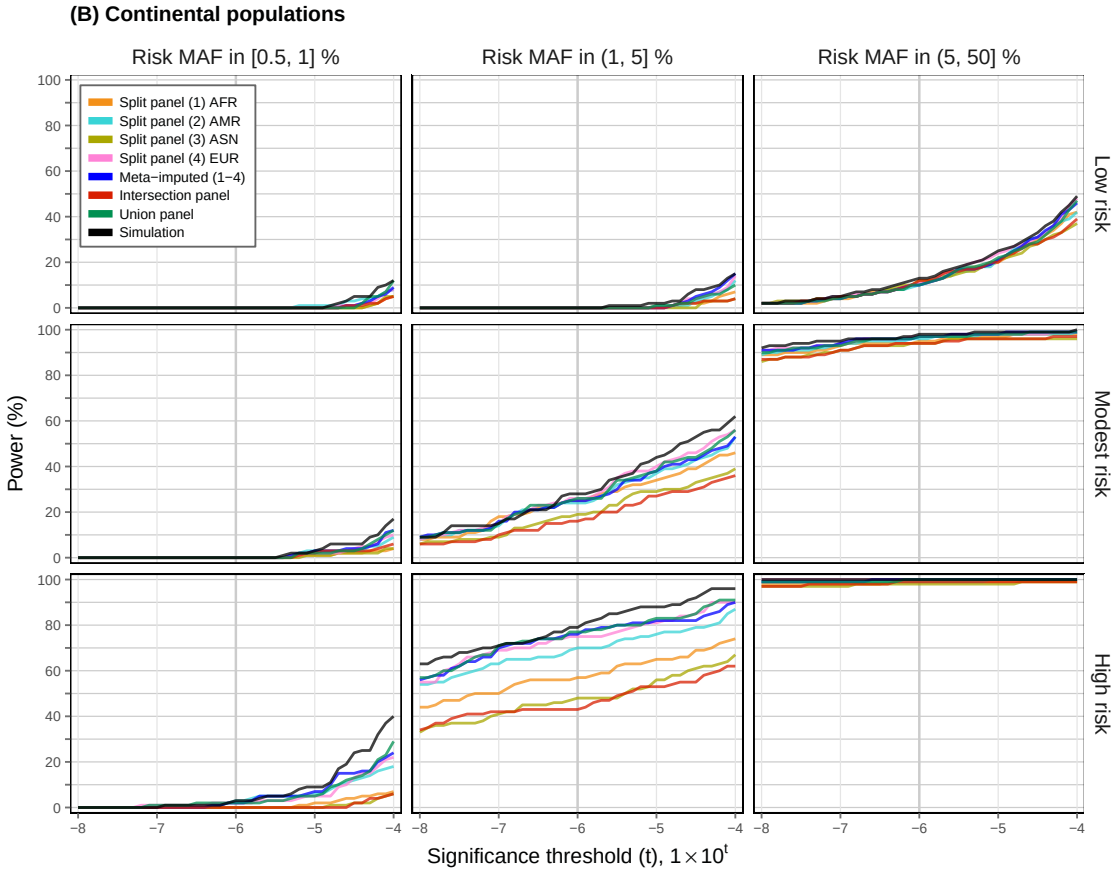


Figure 2.10: Continued.

in Scenarios A and B, 4.85 % (±0.416 % SE) and 2.65 % (±0.248 % SE), respectively, whereas the intersection panel had the highest difference, 9.56 % (±0.846 % SE) and 15.5 % (±1.25 % SE), respectively. Notably, meta-imputation was similarly close as the union panel and outperformed the other imputation strategies in Scenario A, 4.96 % (±0.431 % SE). For example, at a nominal threshold (p-value $\leq 1 \times 10^{-6}$), the union panel reached 71% power and meta-imputation 69% in the high risk category. In Scenario B, the power observed for meta-imputed data was high by comparison, *e.g.* 76% power at high risk, compared to 77% for the union panel and 43% for the intersection panel; however, Δ_P measured for meta-imputation was 3.07 % (±0.295 % SE), which was lower in the EUR split panel, 2.72 % (±0.259 % SE), reaching 76% in the high risk category. In Scenario C, the *Finns* panel showed the lowest difference in power, 3.11 % (±0.343 % SE), and the *ORCADES* panel the highest, 5.76 % (±0.572 % SE); yet, meta-imputation ranked 2nd best among the strategies compared, 3.51 % (±0.357 % SE), but 1st in the high risk category with 74% power (compared to 73% and 68% for *Finns* and *ORCADES*, respectively).

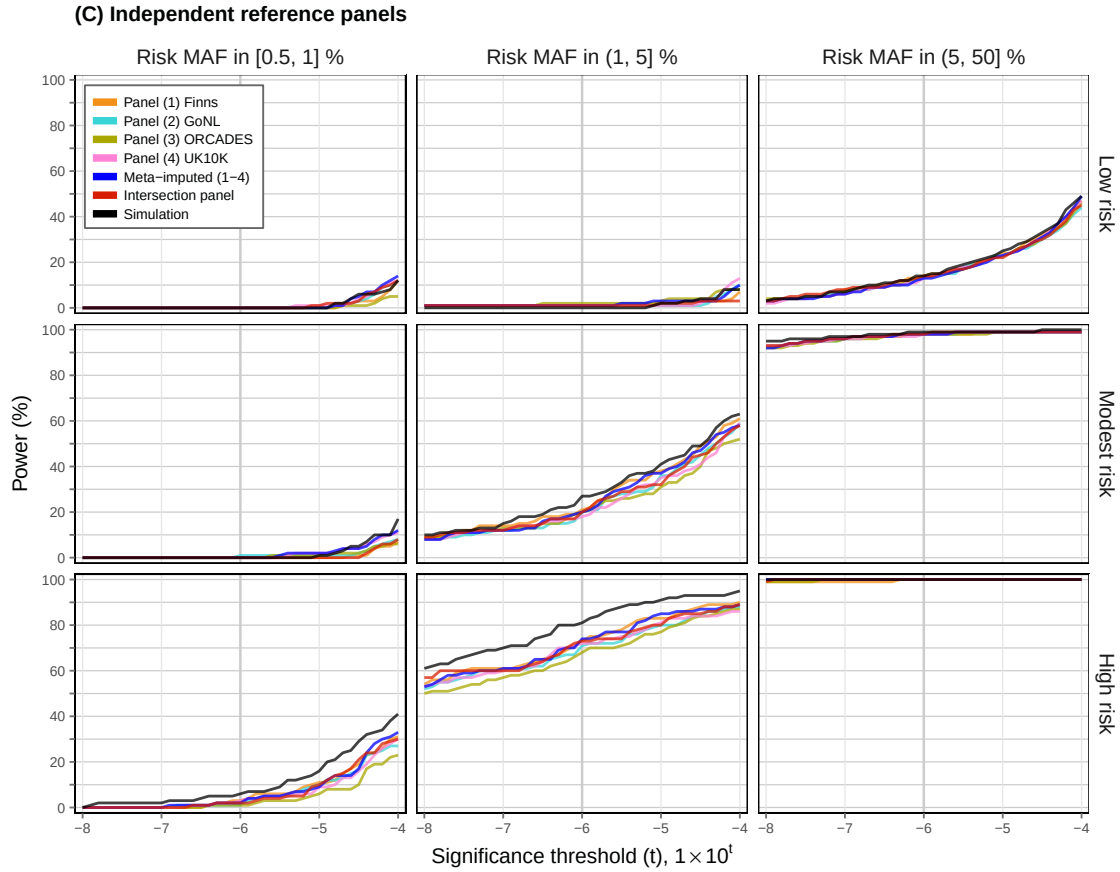


Figure 2.10: Continued.

2.6 Discussion

Meta-imputation was presented as a novel approach to integrate reference data after imputation into a common study sample, but the idea of combining genotype data imputed from different reference panels has been investigated before. Chen *et al.* (2013) used low to high-depth sequencing data as references for imputations into a given study sample, where imputed data have been matched and combined at overlapping sites, but such that variant genotypes imputed from the high-quality panel were included preferentially. They have shown that this approach improved overall accuracy compared to each separately imputed dataset. Here, I considered several variations of this approach which I evaluated using several reference datasets as available in different use case scenarios. Notably, the meta-imputation method does not require prior knowledge to guide the merging process (such as high or low quality of each dataset considered), which instead is determined by summary information derived directly from imputed genotype data.

The results I presented in this chapter showed that the combination of genotype data may indeed result in an increase of accuracy across the allele frequency spectrum, but where the largest improvements were seen for low-frequency variants (*e.g.* 1–5% MAF).

Table 2.6: Estimated power per imputation strategy. Power was estimated as the proportion of significant association signals found among replicate simulation experiments, for which one variant per simulation was selected at random from three MAF intervals (as specified in the table). Each of the selected variants was simulated to act as a causal risk factor, where relative risk was simulated in three categories; low (RR_{het} = 1.2), modest (RR_{het} = 1.6), and high risk (RR_{het} = 2.0). Power at a nominal significance threshold (p-value $\leq 1 \times 10^{-6}$) is reported at each combination of MAF interval and risk category. The average difference (Δ_P) in relation to the non-imputed simulation benchmark is given per MAF interval for each imputation strategy; the lowest average difference is highlighted (**bold**). This table shows the results obtained for imputed and meta-imputed data in Scenario A; results for Scenario B (next page) and Scenario C (page 73) are shown separately.

(A) European sub-populations

Risk MAF (%)	Panel	Power (%), p-value $\leq 1 \times 10^{-6}$			Δ_P (%)*	
		Low	Modest	High	Mean	(± SE)
[0.5, 1]	Split panel (1) CEU	0	0	2	2.537	(0.487)
	Split panel (2) FIN	0	0	0	3.041	(0.520)
	Split panel (3) GBR	0	0	2	2.098	(0.444)
	Split panel (4) TSI	0	1	2	2.309	(0.509)
	Meta-imputed (1-4)	0	0	3	1.553	(0.289)
	Intersection panel	0	0	0	3.366	(0.593)
	Union panel	0	0	3	**1.049**	(0.206)
(1, 5]	Split panel (1) CEU	0	24	62	8.561	(0.729)
	Split panel (2) FIN	0	25	66	6.374	(0.543)
	Split panel (3) GBR	0	25	66	7.431	(0.670)
	Split panel (4) TSI	0	26	65	7.122	(0.605)
	Meta-imputed (1-4)	0	27	69	4.959	(0.431)
	Intersection panel	0	23	60	9.561	(0.846)
	Union panel	0	27	71	**4.846**	(0.416)
(5, 50]	Split panel (1) CEU	11	98	100	0.780	(0.068)
	Split panel (2) FIN	12	98	99	0.894	(0.070)
	Split panel (3) GBR	11	98	100	0.821	(0.083)
	Split panel (4) TSI	11	97	100	0.748	(0.090)
	Meta-imputed (1-4)	10	97	99	0.959	(0.069)
	Intersection panel	11	97	100	0.634	(0.073)
	Union panel	11	97	100	**0.537**	(0.068)

* Average difference in power between simulated and (meta-)imputed association results (Δ_P); averaged over risk category (low, modest, and high risk) and association signals detected at a moving significance threshold; between p-value $\leq 1 \times 10^{-8}$ and p-value $\leq 1 \times 10^{-4}$.

I showed that meta-imputation improved genotype accuracy such that single-reference imputations were outperformed (*e.g.* in Scenario A), but also that meta-imputed genotype data may not further increase accuracy if a reference is highly accurate by itself (*e.g.* the EUR sample in Scenario B). Nonetheless, the inclusion of other, more distantly related reference haplotypes may not affect the accuracy of the resulting meta-imputed dataset (*e.g.* the AFR or ASN samples for imputation into the European sample in Scenario B).

Table 2.6: Continued.

(B) Continental populations

Risk MAF (%)	Panel	Power (%), p-value $\leq 1 \times 10^{-6}$			Δ_P (%)*	
		Low	Modest	High	Mean	(± SE)
[0.5, 1]	Split panel (1) AFR	0	0	0	3.000	(0.541)
	Split panel (2) AMR	0	0	3	1.488	(0.337)
	Split panel (3) ASN	0	0	0	3.203	(0.583)
	Split panel (4) EUR	0	0	2	1.553	(0.296)
	Meta-imputed (1-4)	0	0	2	**1.228**	(0.247)
	Intersection panel	0	0	0	3.049	(0.579)
	Union panel	0	0	2	1.301	(0.247)
(1, 5]	Split panel (1) AFR	0	25	57	9.366	(0.866)
	Split panel (2) AMR	0	24	70	5.220	(0.442)
	Split panel (3) ASN	0	19	48	14.618	(1.197)
	Split panel (4) EUR	0	26	75	2.715	(0.259)
	Meta-imputed (1-4)	0	25	76	3.065	(0.295)
	Intersection panel	0	16	43	15.545	(1.248)
	Union panel	0	26	77	**2.650**	(0.248)
(5, 50]	Split panel (1) AFR	11	95	99	1.984	(0.104)
	Split panel (2) AMR	10	96	99	1.407	(0.109)
	Split panel (3) ASN	12	94	98	2.829	(0.177)
	Split panel (4) EUR	11	97	100	**0.650**	(0.072)
	Meta-imputed (1-4)	10	97	100	0.951	(0.089)
	Intersection panel	12	94	99	2.561	(0.173)
	Union panel	10	97	100	1.138	(0.093)

* See Table 2.6A (page 71).

Meta-imputed genotype data were contrasted with data obtained in imputations from corresponding, larger datasets, which contained the unified sample across the datasets considered in meta-imputation; *i.e*. the intersection and the union of variants present across the other reference datasets, respectively. Although meta-imputation did not perform markedly better in terms of accuracy (measured at the same variant sites), I showed that meta-imputation generally outperformed the intersection panel, in terms of power to detect significant association signals, due to the low coverage retained at the intersection of variants across available reference data. However, note that meta-imputation combined data such that the resulting coverage was identical to the coverage of the union panel; meta-imputed data was overall similar to using the union reference for imputation, with regard to both accuracy and power.

In conclusion, these results suggest that meta-imputation is a viable approach to combine genotype data such that a larger, unified dataset of imputed genotypes is available for association analysis. However, it is unlikely to increase accuracy and power further than possible with imputation from a large, canonical reference; *e.g*. the reference dataset

Table 2.6: Continued.

(C) Independent reference panels

Risk MAF (%)	Panel	Power (%), p-value $\leq 1 \times 10^{-6}$			Δ_p (%)*	
		Low	Modest	High	Mean	(± SE)
[0.5, 1]	Panel (1) Finns	0	0	3	1.919	(0.249)
	Panel (2) GoNL	0	1	1	1.862	(0.290)
	Panel (3) ORCADES	0	0	1	2.805	(0.416)
	Panel (4) UK10K	0	0	3	1.683	(0.305)
	Meta-imputed (1-4)	0	0	2	**1.317**	(0.248)
	Intersection panel	0	0	2	1.854	(0.258)
(1, 5]	Panel (1) Finns	1	21	73	**3.114**	(0.343)
	Panel (2) GoNL	1	20	71	4.943	(0.436)
	Panel (3) ORCADES	2	20	68	5.764	(0.572)
	Panel (4) UK10K	1	18	72	4.789	(0.428)
	Meta-imputed (1-4)	1	20	74	3.512	(0.357)
	Intersection panel	1	20	73	4.195	(0.387)
(5, 50]	Panel (1) Finns	14	98	100	0.683	(0.067)
	Panel (2) GoNL	13	98	100	0.821	(0.103)
	Panel (3) ORCADES	14	98	100	0.911	(0.096)
	Panel (4) UK10K	13	98	100	0.780	(0.079)
	Meta-imputed (1-4)	13	98	100	0.748	(0.079)
	Intersection panel	14	98	100	**0.545**	(0.092)

* See Table 2.6A (page 71).

provided by the Haplotype Reference Consortium (HRC). Yet, future GWA studies may benefit from meta-imputation, for example, in situations when researchers have to choose from a collection of available reference datasets, or to increase the coverage of imputed data in general. The meta-imputation algorithm, as presented in this chapter, is available as a computational tool which I implemented in C++.*

* Meta-imputation software (`meta-impute`): https://github.com/pkalbers/meta-impute

"Begin at the beginning," the King said gravely,
"and go on till you come to the end: then stop."

— Lewis Carroll, *Alice in Wonderland*

3

Using rare variants to detect haplotype sharing and identity by descent

Contents

3.1 Introduction

Identity by descent (IBD) is a fundamental concept in genetics that describes the genealogical relation between individuals (Malécot, 1948). Two chromosomes are said to be identical by descent, or rather to share a haplotype by descent, if they have inherited the same genetic material from a common ancestor (*e.g.*, see Browning and Browning, 2012; Thompson, 2013). Over generations, the length of an ancestral haplotype is broken down through meiotic recombination, as the genetic material is blended with haplotypes that derive from different ancestral lineages. Consequently, any random sample of two different chromosomes carries a unique pattern of relatedness, with different ancestries at different loci, arising as the result of historical recombination events. The underlying

structure of pairwise relatedness can be thought of as a mosaic of segments at which two chromosomes share a haplotype by descent, but where each of these IBD segments traces back to a different most recent common ancestor (MRCA).

In general, knowledge about relatedness, haplotype sharing by descent, or the recombination history of a sample is of importance in a variety of statistical operations that are used in both population and medical genetics research (Milligan, 2003; Albrechtsen *et al.*, 2009; Gusev *et al.*, 2009); for example, to provide insights into the demographic history of a population (Harris and Nielsen, 2013), to inform methods for genotype phasing and imputation (Kong *et al.*, 2008), to map disease loci using linkage analysis (Purcell *et al.*, 2007; Albrechtsen *et al.*, 2009), as well as to reveal patterns of population stratification and to identify unreported relatedness among individuals in disease association analysis (Freedman *et al.*, 2004; Price *et al.*, 2006; Choi *et al.*, 2009; Mathieson and McVean, 2012).

The entire IBD structure of a sample can be represented by the ancestral recombination graph (ARG) (Griffiths, 1991; Griffiths and Marjoram, 1996, 1997b), which is straightforward to generate in coalescent simulations, but inference from observed data is limited (Rasmussen *et al.*, 2014). This is because even complete data is unlikely to provide sufficient information to explicitly infer the ARG, in addition to the problem that inference becomes computationally expensive for larger sample sizes. Most methods for IBD discovery operate on summary statistics to make inference computationally tractable.

In practice, IBD discovery is largely dependent on the length of a shared haplotype and the genetic similarity between compared sequences. Co-inherited haplotypes that are separated by only a few meioses are expected to cover relatively long tracts, because recombination had less time to break down the length of the region shared between the two chromosomes (Thompson, 2008, 2013). Likewise, as mutations are accumulated along different genealogical lineages, the similarity between shared segments is expected to decrease over time. Thus, for most purposes, the detection of *recent* IBD is of primary interest (Browning and Browning, 2010).

Numerous approaches for the detection of IBD segments have been proposed, most of which attempt to infer IBD based on measures of genetic similarity or through use of statistical models to determine salient patterns of linkage disequilibrium (LD). Commonly employed tools are `PLINK` (Purcell *et al.*, 2007), `GERMLINE` (Gusev *et al.*, 2009), `fastIBD` (Browning and Browning, 2011), and `Refined IBD` (Browning and Browning, 2013), to

name a few. The methodological diversity of existing approaches emphasises the central role of IBD in genetics, but also indicates that there is a need for an accurate as well as efficient method to detect IBD in larger samples of purportedly unrelated individuals.

Due to the growing magnitude of available genomic datasets, IBD discovery is becoming more computationally expensive. Note that alternate approaches exist, for example methods to perform long range phasing (LRP) implicitly harness long IBD regions among related individuals (Kong *et al.*, 2008; Palin *et al.*, 2011; Loh *et al.*, 2016a), which employ computationally efficient methods to match relatively long (*e.g.* >10 cM) haplotypes even in very large datasets. But in a general context, as IBD describes a pairwise relationship between two haplotypes, a search algorithm may visit each of the possible pairs of chromosomes in a sample to determine IBD status from patterns of shared genetic variation observed along the full length of the chromosome. For instance, in a sample of n chromosomes, there are $\binom{n}{2} = n(n-1)/2$ possible pairs that need to be scanned to resolve IBD status if done in an exhaustive manner. To reduce this search space, it would be convenient if a pairwise approach could be targeted to regions and individuals for whom it is more likely to find recent haplotype sharing by descent.

In this chapter, I present a non-probabilistic method to detect IBD segments in pairs of diploid individuals, which utilises rare variants as indicators of recent relatedness. The computational burden of IBD detection is thereby reduced due to the relative low number of individuals that share a given rare or low-frequency allele. In each pair, the regions to each side of a focal allele are scanned, so as to infer the "breakpoints" of historical recombination events that delimit the underlying IBD segment. The inference of recombination is based on the *four-gamete test* by Hudson and Kaplan (1985), for which haplotype information is required, but which is extended, following Mathieson and McVean (2014), such that recombination breakpoints can be inferred in genotype data.

In the following section, I highlight the genealogical properties of rare variants which make them useful for the inference of recent and relatively long haplotypes by descent. I then describe the method by which IBD segments are detected, conditional on variation observed at a focal rare variant. For the evaluation of the methodology presented, I generated a large dataset using coalescent simulations, so as to measure the accuracy of the IBD detection method in comparison to the true IBD structure (determined from simulation records). These results are also compared to IBD detected using an alternate method. Lastly, I apply the method presented in this chapter to data from the 1000 Genomes Project (1000G).

3.2 Rare variants as indicators of haplotype sharing by descent

One of the properties of rare variants is their presumed young age, as a low frequency is indicative of a recent origin through mutation (Kimura and Ota, 1973; Griffiths and Tavaré, 1998). Chromosomes sharing a rare allele are therefore likely to have inherited a relatively long haplotype segment from a common ancestor. For example, genetic markers tend to be in high LD with alleles at lower frequencies, because the alleles near a rare variant site are likely to segregate together on the same haplotype (Kruglyak, 1999; Slatkin, 2008b).

Consider a focal site at which two haplotypes are shared by descent. The length of the IBD segment is defined by the nearest ancestral recombination events that occurred to either side of the focal position; *i.e.* haplotype sharing is broken down by recombination on both sides independently. The expected length of a haplotype segment is determined by the number of meioses that separate two haplotypes in relation to the MRCA who lived t generations in the past; hence, the pair is separated by $2t$ meioses. In each meiosis, recombination is modelled as a Poisson process with rate of 1 per unit of genetic distance (*Morgan*). It follows that the recombination process over $2t$ meioses is Poisson distributed with rate equal to $2t$. The expected length can be expressed as the sum of two independent random variables that are exponentially distributed, and which describes the distance to either side of the focal position (see Wakeley and Wilton, 2016); *i.e.* the length, L, is gamma-distributed with shape 2 and rate $2t$, namely $L \propto \Gamma(2, 2t)$.

Given this exponential "decay" of IBD length over time, rare or low-frequency variants are useful for identifying genomic regions in which individuals are likely to share recent and relatively long IBD tracts. For example, Mathieson and McVean (2014) selected doubletons (alleles that are present only twice in a sample), which they refer to as f_2 variants, to identify the shared haplotype in the two individuals sharing the allele. To borrow from this notation, henceforth, f_k is used to denote a variant at which k allele copies are found in a sample.

To emphasise the utility of rare variants, see the example shown in Figure 3.1 (next page). Using coalescent simulations, a sample of $N = 5{,}000$ chromosomes was generated.* A rare variant was randomly selected (frequency $\leq 0.5\%$), as well as two of the chromosomes which share the focal allele. The underlying IBD structure for the given pair of chromosomes was determined from simulation records and shown in Figure 3.1a. IBD segments are distinguished by the time to the most recent common ancestor (T_{MRCA})

* See Section 3.4.1 (page 89) for a description of how data were simulated.

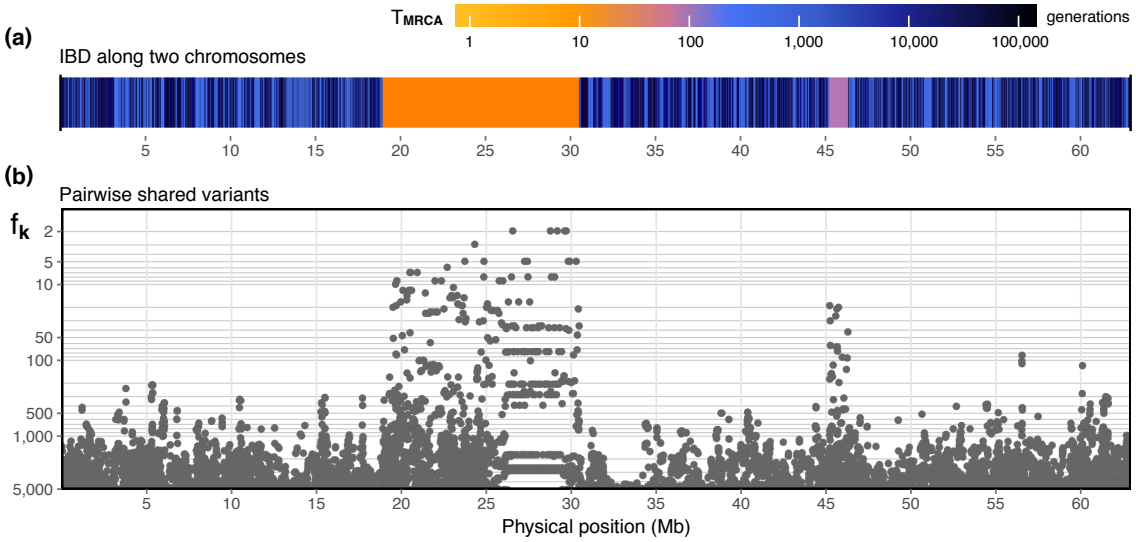
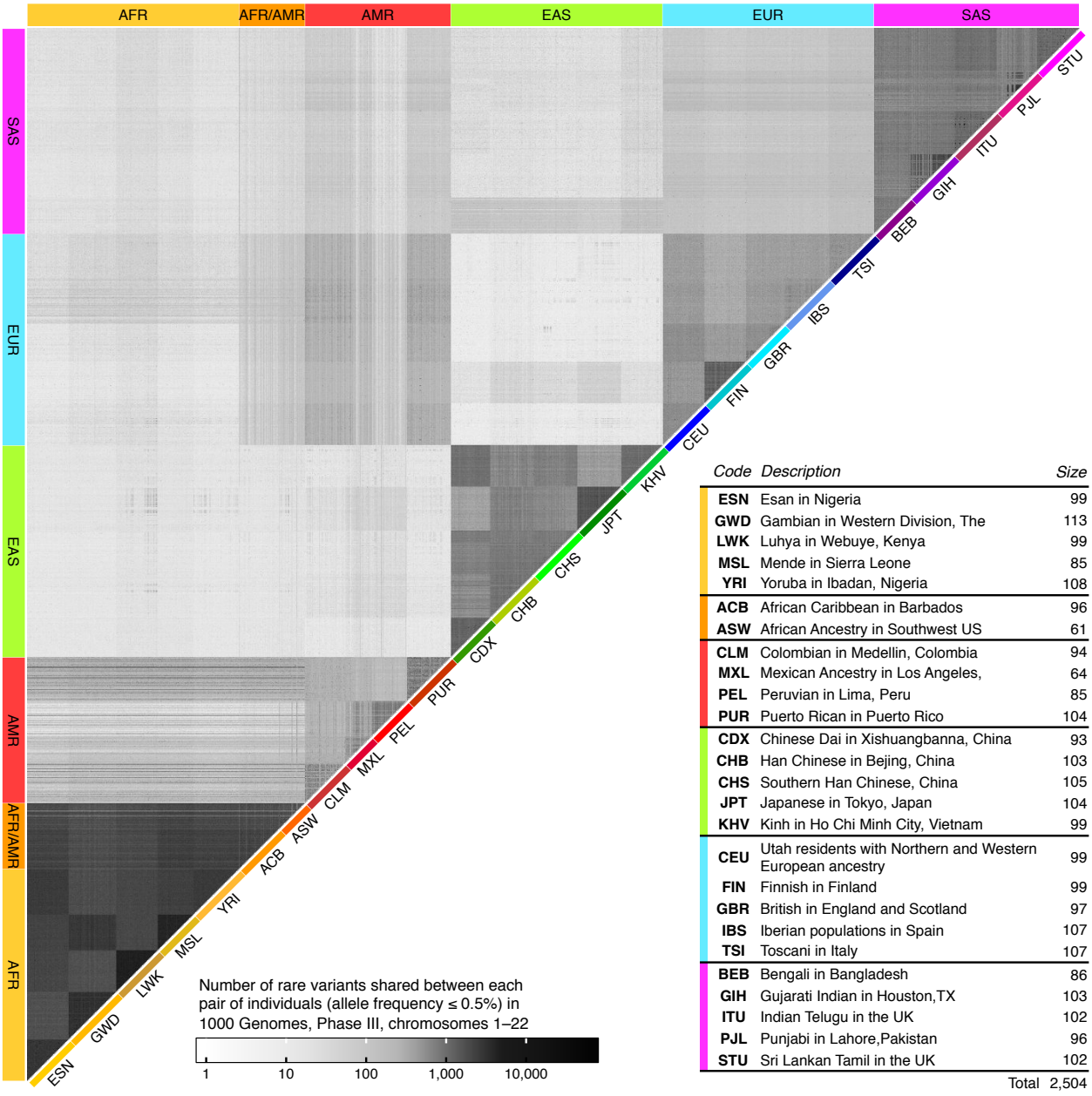


Figure 3.1: IBD structure and pairwise variant sharing. A dataset of N = 5,000 haplotypes was simulated under the coalescent using `msprime` (Kelleher *et al.*, 2016). IBD status was determined from simulated genealogies for a pair of chromosomes selected at random from the set of chromosomes that shared a rare allele (frequency ≤ 0.5%). Panel **(a)** shows the "mosaic" of IBD segments along the full length of the simulated region for the two selected chromosomes. The length of a given IBD segment is defined by the chromosomal interval over which the MRCA of the selected pair does not change. The colour of each segment indicates the time to the most recent common ancestor (T_{MRCA}) for the selected pair. Panel **(b)** shows the physical position of f_k variants shared by the two chromosomes, ranging from very low allele frequency at the top (f_2) to very high frequency at the bottom (*e.g.* $f_{>500}$). Note that the simulation was carried out under variable recombination rates using the genetic map for human chromosome 20 from the International HapMap Project (HapMap) Phase II Build 37. The pattern of extended shared variation seen at positions around 25–30 Megabase (Mb) arises from a low recombination rate at the region of the centromere.

at each position along the sequence. To illustrate pairwise allele sharing, the frequency of each allele shared by the two haplotypes is shown by chromosomal position in alignment with the IBD structure above; see Figure 3.1b. As suggested in the figure, the majority of low-frequency variants align with IBD segments that are more recent.

The majority of variants observed in the human genome are low in frequency or rare. For example, there are 84.7 million single-nucleotide polymorphisms (SNP) in the final release dataset of the 1000 Genomes Project (1000G) Phase III (N = 2,504), of which 71.9 % are below 1% allele frequency and 64.2 % are below 0.5% (after removing singletons and monomorphic sites), suggesting that there are ample opportunities to find rare allele sharing. This is illustrated in Figure 3.2 (next page), which indicates the number of alleles shared between each pair in the dataset (chromosomes 1–22), at allele frequency ≤ 0.5%. Notably, the sharing pattern highlights population structure, as the number of shared alleles is generally larger within a sub-population.



Code	Description	Size
ESN	Esan in Nigeria	99
GWD	Gambian in Western Division, The	113
LWK	Luhya in Webuye, Kenya	99
MSL	Mende in Sierra Leone	85
YRI	Yoruba in Ibadan, Nigeria	108
ACB	African Caribbean in Barbados	96
ASW	African Ancestry in Southwest US	61
CLM	Colombian in Medellin, Colombia	94
MXL	Mexican Ancestry in Los Angeles,	64
PEL	Peruvian in Lima, Peru	85
PUR	Puerto Rican in Puerto Rico	104
CDX	Chinese Dai in Xishuangbanna, China	93
CHB	Han Chinese in Bejing, China	103
CHS	Southern Han Chinese, China	105
JPT	Japanese in Tokyo, Japan	104
KHV	Kinh in Ho Chi Minh City, Vietnam	99
CEU	Utah residents with Northern and Western European ancestry	99
FIN	Finnish in Finland	99
GBR	British in England and Scotland	97
IBS	Iberian populations in Spain	107
TSI	Toscani in Italy	107
BEB	Bengali in Bangladesh	86
GIH	Gujarati Indian in Houston,TX	103
ITU	Indian Telugu in the UK	102
PJL	Punjabi in Lahore,Pakistan	96
STU	Sri Lankan Tamil in the UK	102
	Total	2,504

Figure 3.2: Rare variant sharing in the 1000 Genomes dataset. The plot shows the upper triangle of a pairwise sharing matrix in which the number of variants shared in each pair of individuals is indicated by tones of grey (log-scaled), ranging from *light* (low number) to *dark* (high number); see legend. Pairwise rare variant sharing was determined for all shared alleles observed at frequency ≤ 0.5%, across chromosomes 1–22, and in each pair of the 2,504 individuals present in the final release dataset of the 1000 Genomes Project Phase III. The dataset comprises sample data from six continental populations (or *super-populations*) which are further subdivided in 26 populations of different ethnic background. Each group is abbreviated using a three-letter code. The six continental populations are defined as follows; African (AFR), African-American (AFR/AMR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). The table in the lower right corner shows the code and description of each population sample, as well as the number of individuals in each group.

3.3 IBD detection around rare variants

In the following sections, I describe the methodology by which IBD segments are detected around rare variant sites. I then describe the implementation of each of the two tests for the detection of IBD segments in large sample data. Lastly, I conclude this section by highlighting certain caveats of the implemented method before its evaluation using simulated data.

3.3.1 Inference of historical recombination events

Two approaches for a non-probabilistic inference of recombination events are described below; these are the *four-gamete test* (Hudson and Kaplan, 1985), which requires haplotype information, and the criterion of *inconsistent homozygote genotypes* (see Mathieson and McVean, 2014), which requires genotype data; henceforth referred to as the *discordant genotype test*.

Note that the aim of this implementation is to detect recombination in pairs of diploid individuals, relative to a given target position in the genome, where it is attempted to delimit the shared haplotype segment of the two haplotypes sharing the target allele. This is further explained in Section 3.3.2 (page 84), with examples provided in Section 3.3.3 (page 86).

Four-gamete test (FGT). Given four haplotypes in two diploid individuals, a recombination event is inferred between two loci if all four possible gametes are observed. This holds true under the infinite sites model (Kimura, 1969), where mutation events may only occur once per site in the history of a sample, such that at most two allelic states can be observed at a given site. It follows that for a pair of sites there are four possible allelic state configurations; $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$, where 0 and 1 denote the ancestral and derived type, respectively. If all four configurations are observed, genealogies at the two sites are incompatible and the observation can only be explained by a recombination event that occurred in the history of the sample. Because recurring mutations or back mutations are assumed to have zero probability, at least one recombination event must have occurred in the interval between the two sites. In the following, the term *breakpoint* is used for either of the two sites that together delimit the interval. An example configuration is shown in Figure 3.3 (next page).

Notably, private or *de novo* mutations appearing as singletons in the sample cannot lead to the observation of the four required configurations. Although the exact location of recombination (*i.e.* chromosomal crossover) cannot be retrieved from the data, the

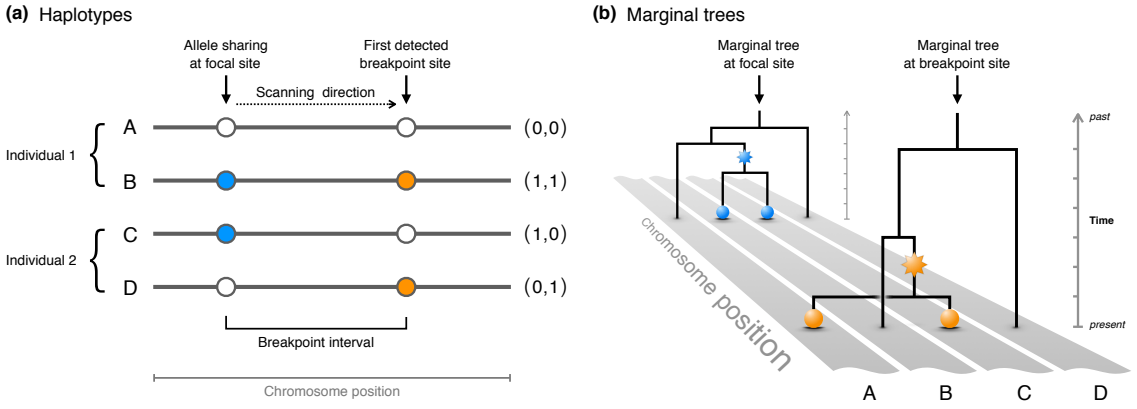


Figure 3.3: Breakpoint detection using the four-gamete test (FGT). Panel **(a)** shows the four haplotypes (gametes) in a pair of two diploid individuals (*horizontal lines*). The focal allele (*blue*) on haplotypes B and C is shared by both individuals. A breakpoint interval is detected if all four possible allelic state configurations are observed at two variant sites along the sequence. Beginning at a given focal site, which is heterozygous in both individuals, the sequences are scanned independently to the left and right hand-side (only right hand-side is shown) until a breakpoint is inferred. The interval delimits the region in which at least one recombination event must have occurred in the history of the sample (given the assumptions of the infinite sites model). The four allelic state configurations are shown on the *right* to each sequence. The alleles are shown at the two breakpoint sites; indicated as ancestral (*hollow* circle) and derived state (*solid*). Note that the order of gametes is ignored. Panel **(b)** shows the corresponding marginal trees at the focal site and the detected breakpoint, where *stars* indicate a mutation event and *spheres* the derived alleles.

FGT can be used to find the smallest interval in which at least one recombination event occurred. However, it is important to note that this test cannot determine which of the four haplotypes recombined, and where it is also possible that there have been multiple recombination events pertaining to different haplotypes within the detected interval.

Discordant genotype test (DGT). In absence of haplotype information, data are represented as genotypes, where genotypic states are encoded as 0, 1, and 2, for variants that are homozygous for the ancestral allele, heterozygous, and homozygous for the derived allele, respectively. Given the genotype sequences of two diploid individuals, recombination is inferred between two sites; one being heterozygous in both individuals (*i.e.* the genotypes 1 and 1) and another with opposite homozygous genotypes (0 and 2). In the latter case, it follows that the two individuals cannot share a haplotype at that locus.

The DGT is a special case of the FGT, as the same composition of alleles is implied. For example, if the allelic configurations $(0, 1)$ and $(0, 0)$ are seen in individual 1, and configurations $(1, 0)$ and $(1, 1)$ in individual 2, the corresponding genotypic configurations are $(0, 1)$ and $(2, 1)$, respectively, which satisfies the breakpoint condition in both the FGT and DGT. However, because genotype data result from haplotype occurrence in

individuals, not all breakpoints detectable under the FGT can be found using the DGT. At sites where the FGT detects a breakpoint interval, the DGT cannot break if both sites are heterozygous in the same individual. As a consequence, it can be expected that the DGT is more restrictive than the FGT; *e.g.* if breakpoints are found, they are likely to sit farther apart.

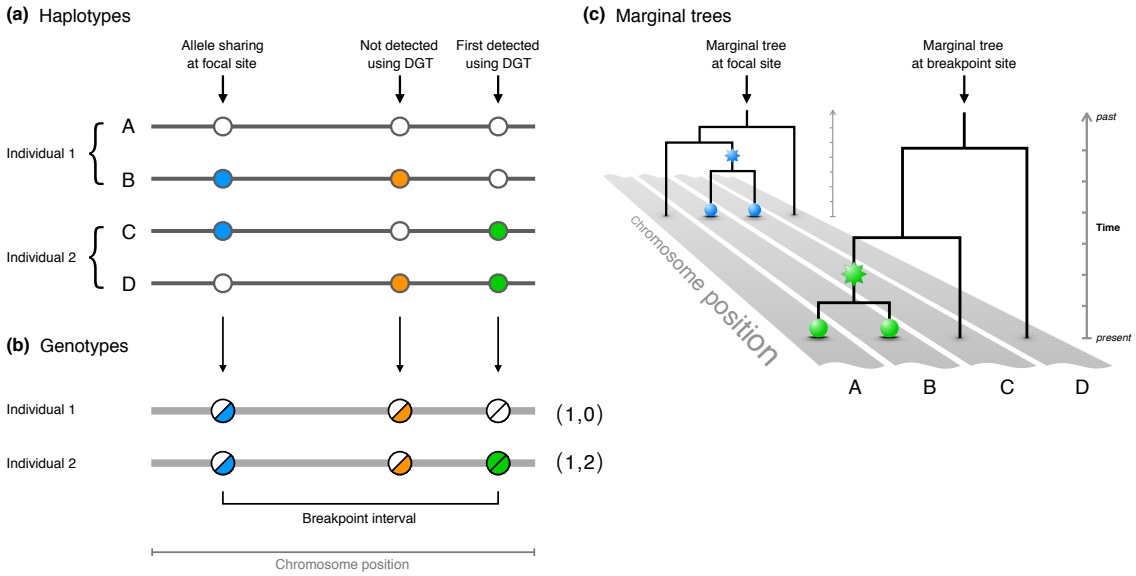


Figure 3.4: Breakpoint detection using the discordant genotype test (DGT). Unlike the FGT, which requires haplotype information, the DGT identifies a breakpoint interval using genotype data. This representation extends the example shown for the FGT in Figure 3.3 (page 82). For comparison, Panel **(a)** shows the four gametes of the two individuals involved. Panel **(b)** shows the two corresponding genotype sequences per individual (*thick* horizontal lines) from which a breakpoint interval is inferred using the DGT. The genotypic states of the breakpoint sites are given on the *right*. Genotypes can either be homozygous for the ancestral allele (*hollow* circle), heterozygous (*semi-solid*), or homozygous for the derived allele (*solid*). The site indicated between the focal site and the detected breakpoint would satisfy the breakpoint condition under the FGT, but is missed under the DGT. Panel **(c)** shows the corresponding marginal trees at the focal site and the detected breakpoint, where *stars* indicate a mutation event and *spheres* the derived alleles.

The DGT thereby simplifies the breakpoint condition to observing opposite homozygote genotypes in two diploid individuals, but where allele sharing at a given target site (that is heterozygous in both individuals) is required such that the conditions of the FGT would be satisfied. This is further exemplified in Figure 3.4 (this page), which highlights the difference to the FGT by comparison to the example shown in Figure 3.3 (page 82).

3.3.2 Description of the algorithm

The FGT and DGT provide the means for non-probabilistic inference of recombination breakpoints from either haplotype or genotype data, respectively. This methodology is implemented such that the full length of an IBD segment can be found around a given target site in a pair of diploid individuals. The allele at a target site serves as an indicator for haplotype sharing by descent; hence, to detect recent IBD, rare variants are used as primary targets. The aim of this method is to infer breakpoint intervals independently on both sides of the target position along the sequence, so as to infer the two recombination events that delimit the underlying IBD segment. As such, the target variant is set as the *focal* breakpoint. The algorithm is described below; a more intuitive example is illustrated in Figure 3.5 (next page).

Let M be the number of variant sites observed in a sample of N diploid individuals. At the target site, b_i, where $i \in \{1, 2, \ldots, M\}$, the subset of individuals sharing the derived allele is identified and compared in a pairwise fashion. Importantly, the allele at this site is used as an identifier for haplotype sharing, on which inference is conditioned in either the FGT or DGT. Thus, individuals are only considered if they are heterozygous for the focal allele, as the breakpoint condition in either test cannot be satisfied otherwise. However, note that this restriction arises from the variant-centric focus on a given rare allele; *e.g.* the condition of the FGT could be satisfied for individuals homozygous for a given allele, but without that the allele is shared by the other individual (hence, defying the purpose of this implementation). In each pair, chromosomes are scanned to the left and right-hand side from the target site until the first site is found that, together with the allelic or genotypic states observed at b_i, satisfies the breakpoint condition, which is done independently on each side. Detected breakpoints are labelled as b_L and b_R on the left and right-hand side, respectively, such that the intervals $[b_L, b_i]$ and $[b_i, b_R]$ delimit the chromosomal regions in which recombination events occurred, respectively; where $L, R \in \{1, 2, \ldots, M\}$. Hence, the underlying IBD segment is enclosed in $[b_L, b_R]$.

The allelic or genotypic states at b_L or b_R provide only the first indication of recombination found along the sequence on either side of the focal allele, but may not mark the points of the actual crossover events. The detected interval is therefore inclusive of the breakpoints such that the full length of the underlying IBD segment is covered. In cases where the end of a chromosome is reached without detecting any evidence of recombination, the terminal site is recorded to capture the length of the segment; this is hereafter referred to as a *boundary case*.

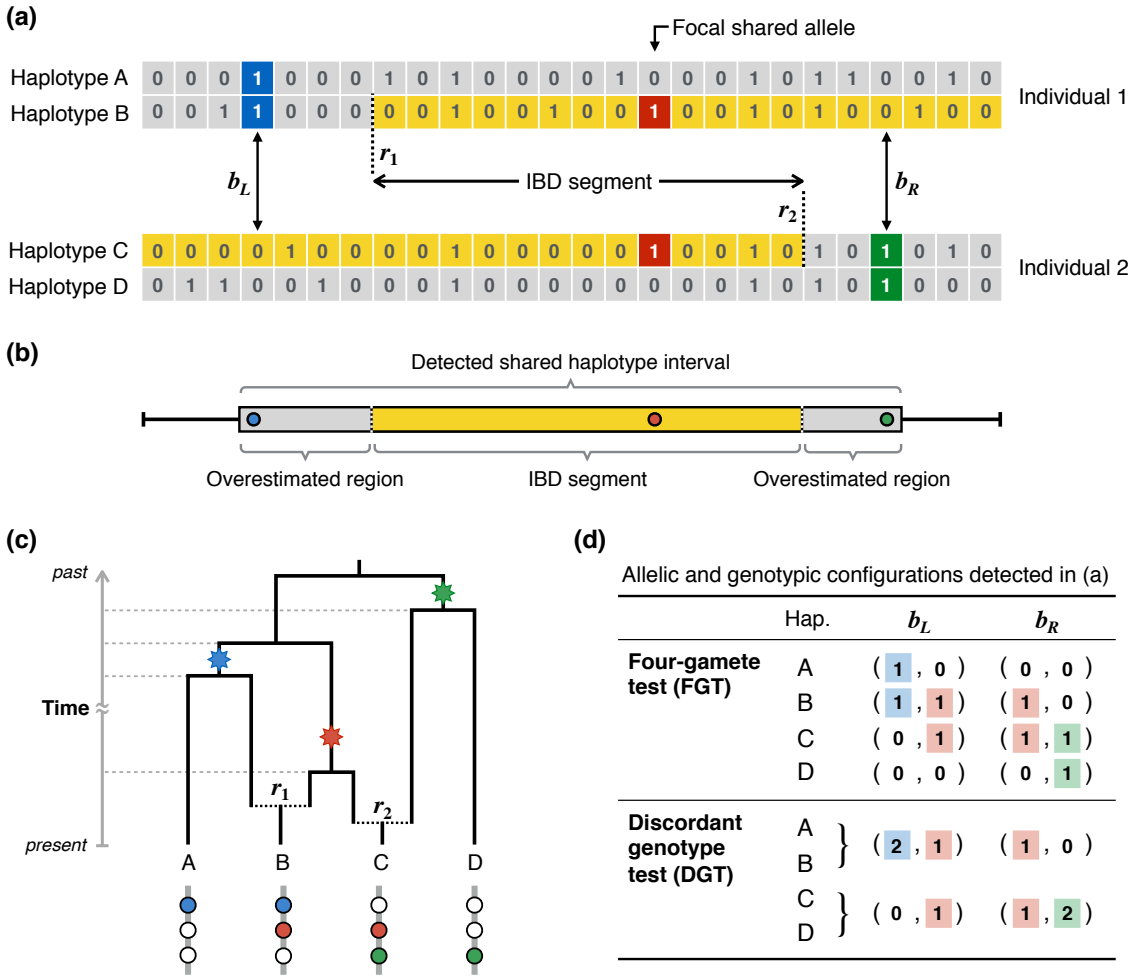


Figure 3.5: Illustration of shared haplotype detection in a pair of diploid individuals. Panel **(a)** shows two individuals composed of haplotypes A and B, and haplotypes C and D, respectively. Each haplotype is represented as a sequence of observed allelic states, where 0 and 1 denote the ancestral and derived allele, respectively. Breakpoints are detected by independently scanning to the left and right-hand side from the target position. The two individuals share a haplotype by descent (highlighted in *yellow*) which is tagged by the focal allele (*red*), for which the two individuals are heterozygous Two sites (*blue* and *green*) mark the first sites at which a breakpoint condition is satisfied, such that b_L and b_R are detected. The IBD segment shared by both individuals is indicated by r_1 and r_2 (*dashed lines*). Panel **(b)** shows the detected breakpoint interval, delimited by b_L and b_R (inclusive). Note that detected breakpoints are only the first indication of recombination found distal to the focal site, but may not mark the points of the actual crossover events; thus, it is expected that the length of the detected segment is overestimated, dependent on available data. Panel **(c)** represents the history of the sample as an ancestral recombination graph (ARG). Mutation events are indicated on the tree (*stars*) and gave rise to the alleles highlighted in (a); *blue*, *red*, and *green*. The dotted *grey* lines indicate the time of coalescent events in the history of the sample; dotted *black* lines indicate recombination events. Panel **(d)** provides a table outlining the configurations of allelic and genotypic states at breakpoint sites as considered in the FGT and DGT, respectively. Notably, in the example shown, both the FGT and DGT detect breakpoints at indicated sites. But, for example, if individual 1 was composed of haplotypes A and C, and individual 2 of haplotypes B and D, the breakpoints would be detected as shown under the FGT, but not the DGT.

3.3.3 Anticipated limitations

As noted by Hudson and Kaplan (1985), not all recombination events in the history of a sample are found by the FGT, and are therefore also missed by the DGT. In the implementation presented, a breakpoint is found by performing a scan along the sequence away from a target position. Provided that the neighbouring haplotype regions derive from different ancestral lineages, in the general case, it is likely that a breakpoint will be found eventually (or the boundary of the chromosome is reached).

The main limitation to the accuracy of the detected breakpoints is the overestimation of the interval, in relation to the underlying true IBD length; as shown in Figure 3.5b. While the underlying IBD segment is enclosed in the interval, it can be expected that breakpoints are detected at sites some distance away from where recombination occurred, thus overestimating the true length of the underlying IBD tract. The extent of overestimation is dependent on the number and density of observed variant sites in the sample. Because the rate of mutation is directly proportional to the expected number of segregating sites (Watterson, 1975), a higher mutation rate can generally be expected to decrease the overestimation of segment length.

When using whole-genome sequencing (WGS) data, strategies for variant calling and filtering may affect the accuracy of detecting recombination events as not all variant sites might be captured correctly. It cannot be expected that genotyping arrays provide sufficient marker density to infer breakpoints with high accuracy (with regard to the true length of the underlying shared haplotype segment).

Conversely, it is also possible that segment length is underestimated. A recombination event can occur with chromosomes outside the sub-tree of the lineages deriving from the focal mutation, such that a detected breakpoint may pertain to recombination occurring on either of the "unshared" haplotypes. Given the four chromosomes required, there are $\binom{4}{2} = 6$ possible pairs of chromosomes which may share extended haplotype regions by descent. A segment that was more recently co-inherited by one of the two other haplotypes that do not share the focal allele may result in detection of an "unwanted" breakpoint, where recombination did not break the genealogical relationship between the haplotypes sharing the allele. The FGT or DGT may correctly infer a recombination event, but which would result in an underestimation of the length of the focal IBD segment. This is illustrated in Figure 3.6 (next page), which shows two examples generated using coalescent simulations.

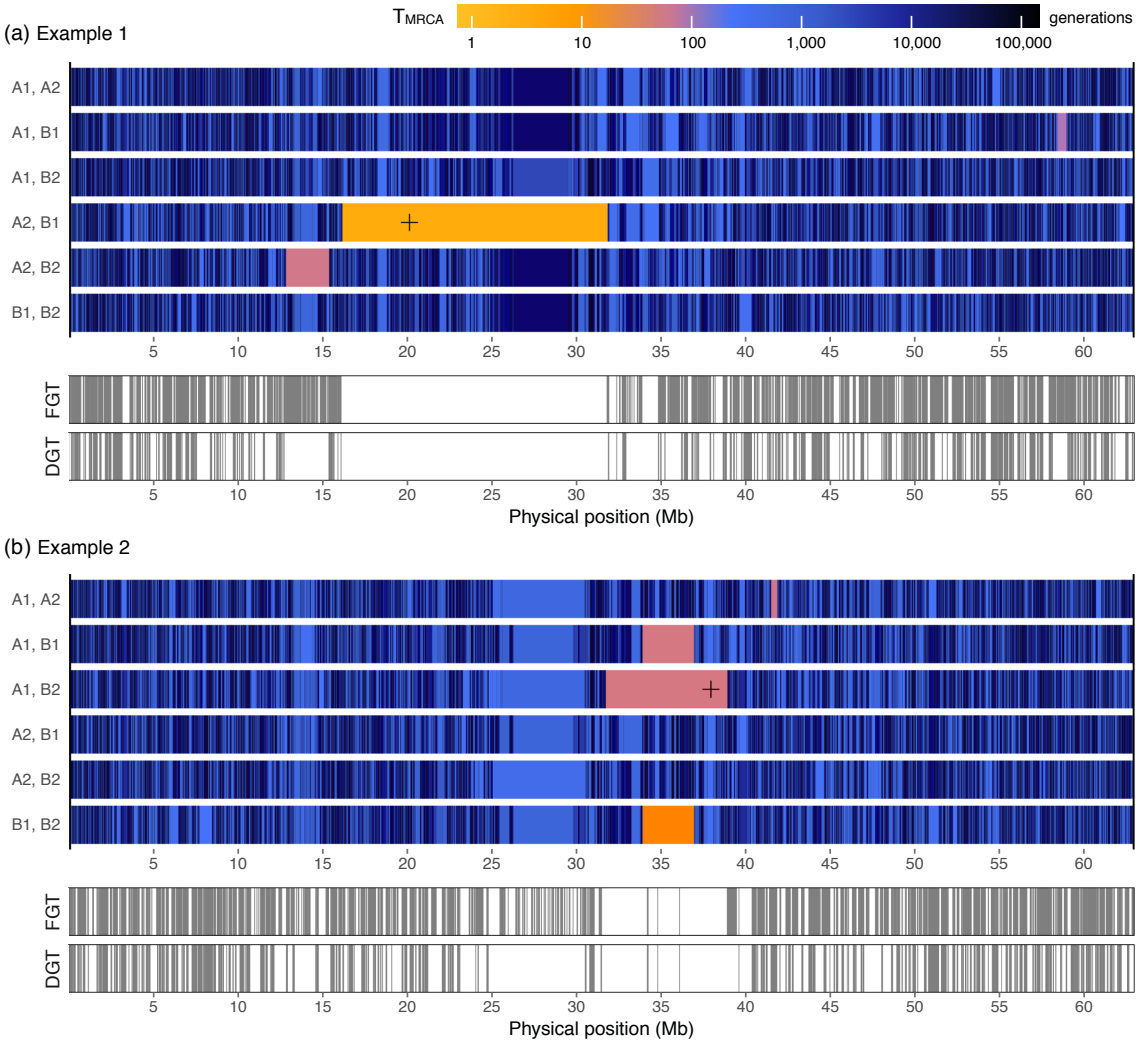


Figure 3.6: Examples of the underlying IBD structure in each pair of four chromosomes. The true, underlying IBD structure is shown for each possible pair among four chromosomes in two diploid individuals; two examples are shown. Each chromosome is labelled by its occurrence in individuals A and B, where chromosomes 1 and 2 are distinguished. The "mosaic" of IBD segments per pair was determined from coalescent records produced in simulations using `msprime` (Kelleher *et al.*, 2016); see Section 3.4.1 (page 89). Each segment defines the region that was co-inherited from a most recent common ancestor (MRCA), and is colour-coded by the number of generations separating the two chromosomes from their shared MRCA in that region. The *cross* marks the position of the focal allele in the pair that shares it. Below, all breakpoints detected relative to the focal variant along the simulated region are indicated, using the FGT (*top*) and DGT (*bottom*). Panel **(a)** shows that the innermost breakpoint intervals (relative to the target position) detected in the FGT or DGT align closely with the true termini of the IBD segment. The extent of overestimation appears to be negligible in relation to the length of the detected segment. Panel **(b)** shows that the innermost intervals are underestimated, due to an overlap of recently co-inherited haplotypes on different chromosome pairs.

In Example 1 (3.6a), a rare allele target site was randomly selected, as well as the two individuals sharing the focal allele. The true IBD structure was determined from simulation records for each pair of the four chromosomes in the two individuals. Each pair is represented by a mosaic of IBD segments along the sequence, where each segment is distinguished by time to the most recent common ancestor (T_{MRCA}). Both the FGT and DGT were applied, but where all consecutive breakpoints after the first detection were also recorded along the sequence on both sides of the focal variant. The innermost interval delimits the detected shared haplotype segment around the target site. Example 1 illustrates the case in which a rare allele identifies the underlying co-inherited haplotype segment, which may stand out as being much younger due to recent shared ancestry.

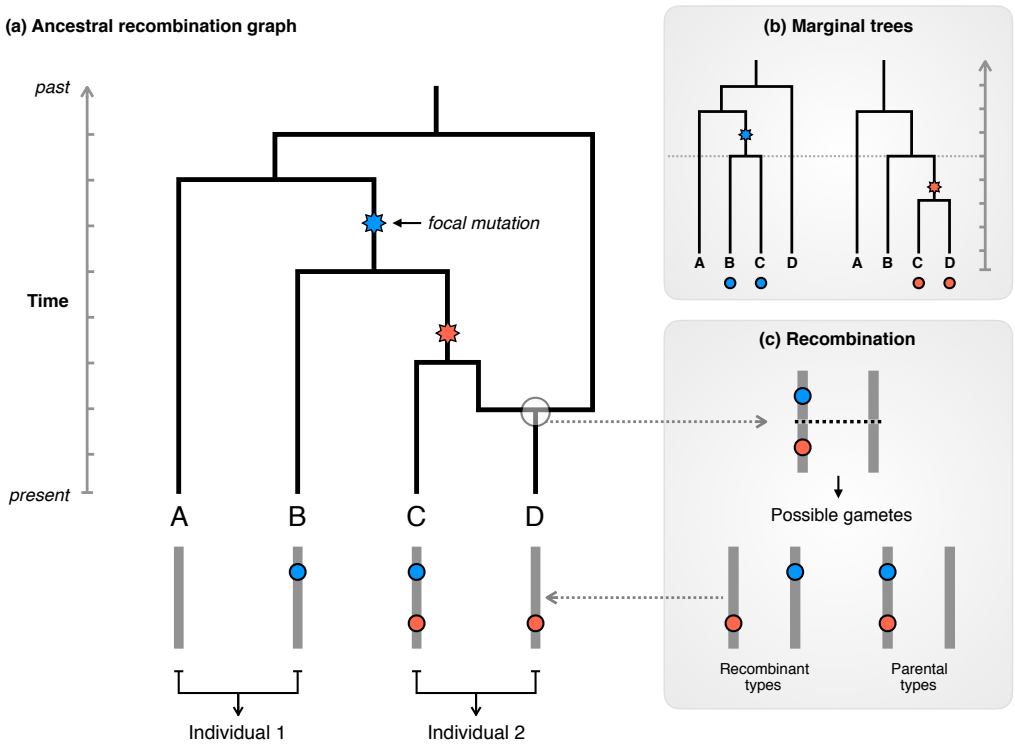


Figure 3.7: Recombination outside the focal sub-tree. The example shows two diploid individuals that share a focal allele (*blue*) on haplotypes *B* and *C*, which form a sub-tree within the larger genealogy. Panel **(a)** shows the ancestral recombination graph (ARG) of a possible recombination history that would result in an "unwanted" detection of a breakpoint, due to recombination event outside the focal sub-tree. Both the FGT and DGT would correctly detect a recombination event between the two sites. However, because recombination occurred neither with haplotype *B* nor *C*, the co-inheritance relationship between both haplotypes (relative to the focal allele) remained unbroken. Panel **(b)** shows the marginal trees at the two sites, corresponding to the ARG shown in Panel (a). Note that the T_{MRCA} of haplotypes *B* and *C* in both marginal trees is identical. Panel **(c)** shows the possible gametes that can result from a recombination occurring between the two sites indicated.

The same was done in Example 2 (3.6b), but here the target site and the pair of individuals was chosen because it was found that the length of the detected IBD segment was underestimated. As can be seen, this underestimation is due to a recombination event on an "unshared" haplotype (which does not carry the focal allele), which occurred more recently than the mutation event giving rise to the focal allele. Figure 3.7 (page 88) shows a possible genealogy that could give rise to a variation pattern where the focal shared haplotype segment is underestimated. Such a result may be expected in cases of inbreeding, where the maternal and paternal chromosomes in an individual are more closely related to each other than to other chromosomes in the population. Note that in the simulations conducted, the generated haplotypes were randomly paired to form diploid individuals.

3.4 Evaluation

The IBD detection method presented in this chapter was evaluated using simulated data. This allowed assessment of the accuracy of detected breakpoint intervals in relation to the known genealogy of the simulated sample. For comparison, an alternate IBD detection method was applied to the same data. Lastly, the method presented was applied to data from the 1000 Genomes Project.

3.4.1 Data generation

The coalescent simulator used to generate data was `msprime` (version 0.4.0), which simulates the exact coalescent with recombination, and where mutations are generated under the infinite sites model (Kelleher *et al.*, 2016).* The software is a reimplementation of the classic `ms` algorithm by Hudson (2002), but allows efficient simulation of extended chromosomal regions for very large sample sizes, where the entire history of the simulated sample can be stored and queried for further analysis. Notably, `msprime` allows simulation under variable recombination rates, for example by using established recombination maps of the human genome.

* Coalescent simulator `msprime`: https://github.com/jeromekelleher/msprime [Date accessed: 2016-11-12]

3.4.1.1 Demographic model

A demographic model was defined following Gutenkunst *et al.* (2009), who used intergenic data from four global populations to estimate parameters from diffusion approximations of expected allele frequency spectra. Accordingly, here, data were simulated with an ancestral population size of $N_e = 7{,}300$ (denoted by N_A in the model) and under the assumption of a generation time of 25 years. The mutation rate was set to a constant $\mu = 2.35 \times 10^{-8}$ per site per generation, which was estimated from the human-chimp divergence in Gutenkunst *et al.* (2009). Note that recent studies have estimated the human mutation rate to be slightly lower; for example, Scally and Durbin (2012) have estimated $\mu \approx 1.2 \times 10^{-8}$ from analyses of genome-wide *de novo* mutations using recent sequencing technologies. The mutation rate used here is two-fold higher, but still in the same order.

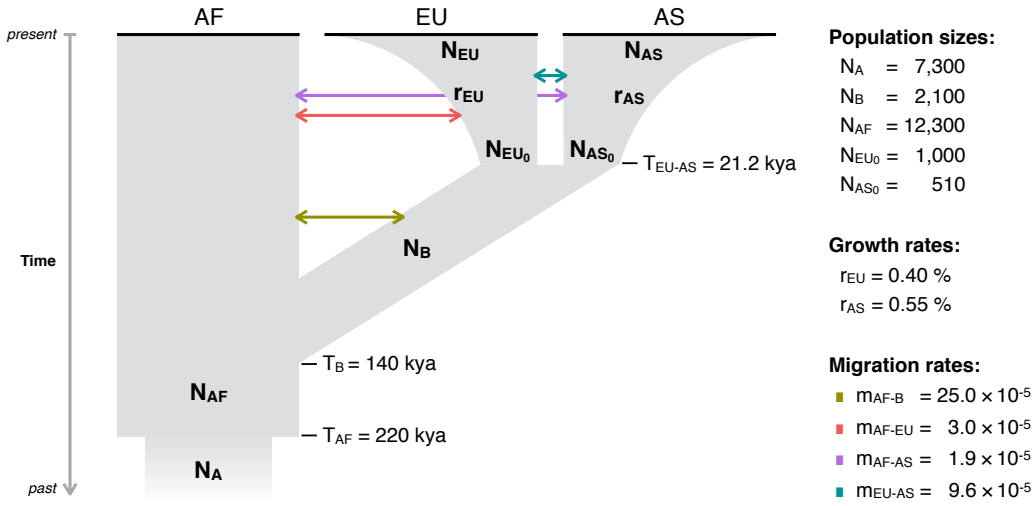


Figure 3.8: Demographic model used in simulations. Three populations were modelled, African (AF), European (EU), and Asian (AS), which derive from an ancestral population (A). Both EU and AS experienced a bottleneck with subsequent exponential growth following the out-of-Africa expansion of a founder population (B) that split from the ancestral population. Modified from Gutenkunst *et al.* (2009), Figure 2 (see doi:10.1371/journal.pgen.1000695.g002), with parameter values taken from Table 1 (see doi:10.1371/journal.pgen.1000695.t001).

The demographic history as defined in the simulation model is illustrated in Figure 3.8 (this page); parameter values of the model are specified therein. The model recapitulates the human expansion out of Africa, for which three populations were considered; African (AF), European (EU), and Asian (AS). The African population was included with a constant population size, while EU and AS experienced exponential growth after divergence and split from an ancestral African population. Population sizes of EU and AS were calculated as $N = {}^{N_0}/_{e^{-rt}}$, where N is the size at present, N_0 the initial size at EU-AS divergence, r the growth rate, and t the time since divergence (in years).

3.4.1.2 Simulated dataset

A sample of 5,000 haplotypes was simulated, where the set of generated chromosomes represented a sample taken from the EU population. To reproduce realistic distributions of recombination variability along the simulated sequence, the simulation was performed using recombination rates from human chromosome 20, as provided in Build 37 of the International HapMap Project (HapMap) Phase II (International HapMap Consortium *et al.*, 2007; International HapMap 3 Consortium *et al.*, 2010).* Note that the ratio between genetic and physical length on human chromosome 20 is $\approx 1.7\,\frac{\text{cM}}{\text{Mb}}$, which differs from the average observed for the human genome at $\approx 1.2\,\frac{\text{cM}}{\text{Mb}}$. The resulting dataset consisted of 0.673 million segregating sites observed over a chromosomal length of 62.949 Mb (108.267 cM). The history of the simulated sample was stored separately to derive genealogical information in subsequent analyses.

The simulated chromosomes were used to generate three datasets. In the first, haplotypes were randomly paired to construct a sample of 2,500 diploid individuals. From this, second, a corresponding genotype dataset was generated by forming genotypes (encoded as 0, 1, and 2) as the sum of alleles (encoded as 0 and 1) along the sequence in each individual. This dataset was then used to generate a third dataset in which haplotypes were estimated from genotype data; *i.e.* resulting data consisted of phased haplotypes. Phasing was conducted using `SHAPEIT` version 2 (Delaneau *et al.*, 2008, 2013), using default parameters without a reference panel.†

3.4.2 Accuracy analysis

The detection of IBD was evaluated in relation to the underlying true IBD structure of the sample, which was determined from the stored simulation records. Given a target site and the two haplotypes sharing the focal allele, the genealogy was scanned along the sequence of variant sites observed in the sample, in both directions from the target position. The MRCA of the pair was identified at each variant site and a breakpoint was defined as the first site at which a different MRCA was found. This returned the smallest interval detectable from available data around the nearest recombination events that delimit an IBD segment.

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

† Phasing software SHAPEIT: https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html [Date accessed: 2016-11-12]

Accuracy was measured in terms of the physical distance between a given breakpoint site and the focal position of the segment. Two measurements were considered; the squared Pearson correlation coefficient, r^2, which measures the strength of the linear relation between detected and true distance, and the root mean squared logarithmic error (RMSLE);

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\log_{10}\left(\frac{\hat{d}_i + 1}{d_i + 1}\right)\right]^2} \tag{3.1}$$

where d_i and $\hat{d}_i$ are the distances of the true and detected breakpoints, respectively, and n is the overall number of comparisons. The RMSLE is similar to the root mean squared error (RMSE), which measures the variance and bias in the set of compared values, and is equal to the standard deviation when there is no bias. As such, the RMSLE can be interpreted as a score metric for the magnitude of error. Here, this is useful because larger departures from the actual values are penalised more than smaller ones. A lower score value indicates a lower magnitude of error, where RMSLE = 0 indicates that true and inferred values are identical. Also, note that the RMSLE is usually defined using the natural logarithm; here, $\log_{10}$ was used as a more intuitive representation of error magnitude.

Note that the same breakpoint interval may be inferred from multiple shared alleles in a given haplotype pair. Below, the number of detected segments was reduced to the set of "uniquely" detected segments per pair in each approach. The remaining "duplicate" segments were removed by sorting identical intervals by the frequency of target alleles, where only the segment detected around the allele of the lowest frequency was retained (and randomly sampled if multiple focal alleles occurred at the same frequency). Detected segments were thereby tagged by the presumably youngest shared allele within a given interval. This enabled the analysis to measure breakpoint accuracy conditional on the frequency of the target allele, which otherwise may have produced biased results due to sharing of higher-frequency alleles that are presumed to be older than the T_{MRCA} of the underlying shared haplotype.

The performance of the proposed IBD detection method was assessed for the described haplotype and genotype-based tests on the three datasets derived from coalescent simulations. The following approaches were distinguished:

(a) FGT on haplotype data as simulated; *i.e. true* haplotypes,

(b) FGT on *phased* haplotypes, and

(c) DGT on genotype data.

3.5 Results

IBD detection was carried out on a large set of target sites, for which all f_k variants found at $k \in \{2, \ldots, 25\}$ were selected, *i.e.* alleles shared at frequency $\leq 0.5\%$. This threshold was chosen arbitrarily, but such that the considered frequency range was expected to be sufficiently low to identify recent IBD given the size of the sample. The set of target sites comprised 0.317 million SNPs that were heterozygous in the individuals sharing a focal allele. This resulted in 11.598 million pairwise analyses and an equal number of IBD segments detected using the FGT on the true and phased haplotypes in Approaches (a) and (b), respectively, and the DGT on genotype data in Approach (c).

The number of uniquely identified segments differed slightly in Approaches (a), (b), and (c); 2.983 million (25.723 %), 3.091 million (26.654 %), and 2.978 million (25.679 %), respectively. For the corresponding true IBD segments, the number of unique segments was 3.001 million (25.876 %). These data were further reduced to the intersection of retained target sites across approaches, so as to enable direct comparisons on the same set of targets, which resulted in 2.978 million (25.679 %) unique intervals. The results obtained from these data are summarised in Table 3.1 (next page).

The proportion of breakpoints that were overestimated (in relation to the corresponding true IBD breakpoints) was noticeably high overall; using the FGT, 97.390 % and 95.666 % were overestimated in Approaches (a) and (b), respectively. However, overestimation was highest when the DGT was used (98.362 %) in Approach (c). Conversely, the proportion of underestimated breakpoints was lowest in (c), 1.543 %, and highest when haplotypes were phased in (b), 4.147 %. In (a), 2.418 % of breakpoints were underestimated. The proportion of detected breakpoints that coincided with the corresponding true breakpoints was 0.192 %, 0.188 %, and 0.095 % in (a), (b), and (c), respectively.

The highest overall accuracy was found for the FGT on true haplotypes, followed by the analysis on phased haplotypes, which had $r^2 = 0.926$ and $r^2 = 0.892$ in Approaches (a) and (b), respectively. The accuracy achieved by the DGT was lower, but still considerably high with $r^2 = 0.847$ in Approach (c). This was also reflected in the measured magnitude of error (RMSLE), which was 0.400, 0.434, and 0.569 in (a), (b), and (c), respectively. The measurement of accuracy was further broken down by the allele frequency of target variants (f_k category); results are shown in Table 3.1 (next page). Accuracy decreased towards higher allele frequency in each approach. For example, for f_2 variants, r^2 was

Table 3.1: Accuracy of detected breakpoints per f_k category. The accuracy of detected IBD breakpoints was measured using the squared Pearson correlation coefficient, r^2, and the RMSLE in relation to the true IBD segments determined from simulation records; measured in terms of the distance between breakpoint site and the corresponding focal position per segment. The analysis included of 317,020 target sites around which IBD was detected in Approaches (a), (b), and (c). In each, accuracy was computed after data were reduced to identical sets of unique IBD segments (n = 2,978,220). The table specifies the allele frequency (%) corresponding to each f_k category, as well as the number of target sites identified.

f_k	Freq. %	Targets	r^2			RMSLE		
			FGT*	FGT**	DGT†	FGT*	FGT**	DGT†
2	0.04	76,515	**0.998**	0.895	0.995	**0.219**	0.598	0.317
3	0.06	46,138	**0.989**	0.957	0.978	**0.243**	0.516	0.359
4	0.08	31,658	**0.963**	0.959	0.941	**0.256**	0.463	0.379
5	0.10	23,581	**0.975**	0.963	0.929	**0.276**	0.429	0.408
6	0.12	19,241	**0.954**	0.938	0.904	**0.281**	0.409	0.421
7	0.14	15,869	**0.955**	0.944	0.892	**0.298**	0.403	0.447
8	0.16	13,175	0.898	**0.918**	0.813	**0.320**	0.398	0.469
9	0.18	10,966	**0.932**	0.927	0.827	**0.314**	0.375	0.467
10	0.20	11,142	0.879	**0.887**	0.773	**0.332**	0.387	0.494
11	0.22	9,392	**0.895**	0.892	0.758	**0.344**	0.401	0.513
12	0.24	7,751	0.835	**0.848**	0.733	**0.358**	0.398	0.526
13	0.26	6,933	**0.842**	0.835	0.721	**0.361**	0.405	0.532
14	0.28	5,767	**0.816**	0.816	0.679	**0.367**	0.391	0.540
15	0.30	5,062	**0.871**	0.860	0.712	**0.381**	0.406	0.556
16	0.32	4,711	**0.839**	0.830	0.701	**0.373**	0.395	0.546
17	0.34	4,210	0.829	**0.832**	0.681	**0.387**	0.410	0.566
18	0.36	3,913	0.813	**0.832**	0.670	**0.380**	0.397	0.561
19	0.38	3,684	**0.801**	0.798	0.642	**0.381**	0.401	0.566
20	0.40	3,214	0.831	**0.837**	0.685	**0.401**	0.416	0.587
21	0.42	3,333	0.773	**0.778**	0.603	**0.399**	0.413	0.584
22	0.44	2,863	0.753	**0.795**	0.571	**0.399**	0.406	0.586
23	0.46	2,595	0.732	**0.745**	0.596	**0.414**	0.425	0.599
24	0.48	2,653	**0.784**	0.780	0.581	**0.396**	0.406	0.583
25	0.50	2,654	0.701	**0.730**	0.560	**0.400**	0.408	0.585

* Approach (a), using the FGT with true haplotypes
** Approach (b), using the FGT with phased haplotypes
† Approach (c), using the DGT with genotype data

0.998, 0.895, and 0.995 in (a), (b), and (c), respectively, which was reduced for f_{25} variants where r^2 = 0.701 in (a) and r^2 = 0.730 in (b), but where (c) was seen to decrease more rapidly by comparison (r^2 = 0.560).

Notably, when haplotype data were phased in Approach (b), accuracy was highest at f_5 variants (r^2 = 0.963), indicating that accuracy was decreased at lower frequencies. Similarly, RMSLE scores reflected the same general pattern, but where the magnitude of error in (b) was at a maximum at f_2 variants. One explanation is that relatively long haplotype regions are more likely to be affected by phasing errors, in particular switch errors, to which the FGT but not the DGT is susceptible. Hence, because f_2 haplotypes

are likely to be longer compared to haplotypes tagged by higher-frequency alleles (as a higher frequency is indicative for an older age), phasing errors are more likely to fall within longer shared haplotype regions and thereby may facilitate an underestimation of shared haplotype lengths.

A more intuitive representation of results is provided in Figure 3.9 (next page), which compares true and detected breakpoint distances in two ways. First, in Figure 3.9A, breakpoint densities are shown in separate scatterplots for breakpoints detected on the left and right-hand side of focal positions. For example, a clear difference in the proportion of underestimated breakpoints can be seen between Approaches (a) and (b), *i.e.* where the FGT was used on true and phased haplotypes, respectively. In Approach (c), where the DGT was used on genotype data, breakpoint densities indicate a higher proportion of overestimated distances compared to (a) or (b). Second, in Figure 3.9B, the relative distance was calculated as $x = \hat{d}_i/d_i$, where $\hat{d}$ and d denote detected and true distances, respectively. By doing so, detected breakpoint distances were "mapped" relative to the corresponding true distances, such that $0 < x < 1$ indicates underestimation and $x > 1$ indicates overestimation. The CDF of the relative distance is shown separately per f_k category. For example, it can be seen that a larger proportion of f_2 variants (15.215 %) contributed to the overall underestimation found in Approach (b), *e.g.* compared to f_5 (5.544 %) and f_{25} variants (0.755 %).

The distribution of physical and genetic IBD length is shown in Figure 3.10 (page 97). These results were obtained after boundary cases were removed in each approach (*i.e.* discarding segments where the end of a chromosome was reached without detecting a breakpoint), so as to ensure that observed IBD length was delimited by recombination on both sides of a segment; 1.449 %, 1.400 %, and 1.637 % were removed in (a), (b), and (c), respectively, and 1.340 % in the set of true IBD segments. Data were then intersected again to retain the same set of target sites in each approach; as a result, 2.929 million unique segments were retained (98.363 %).

Median physical length (and median genetic length) over the set of retained segments was computed for each approach. A small difference was seen for the FGT, where median length was 0.417 Mb (0.800 cM) on true haplotypes in Approach (a), and 0.413 Mb (0.791 cM) on phased haplotypes in Approach (b). For the DGT on genotype data, median length was longer by comparison, 0.570 Mb (1.094 cM). The median of true IBD length was 0.328 Mb (1.573 cM), which was shorter than detected in each approach. But as seen in Figure 3.10, the distribution of IBD lengths in (a), (b), and (c) closely followed the true

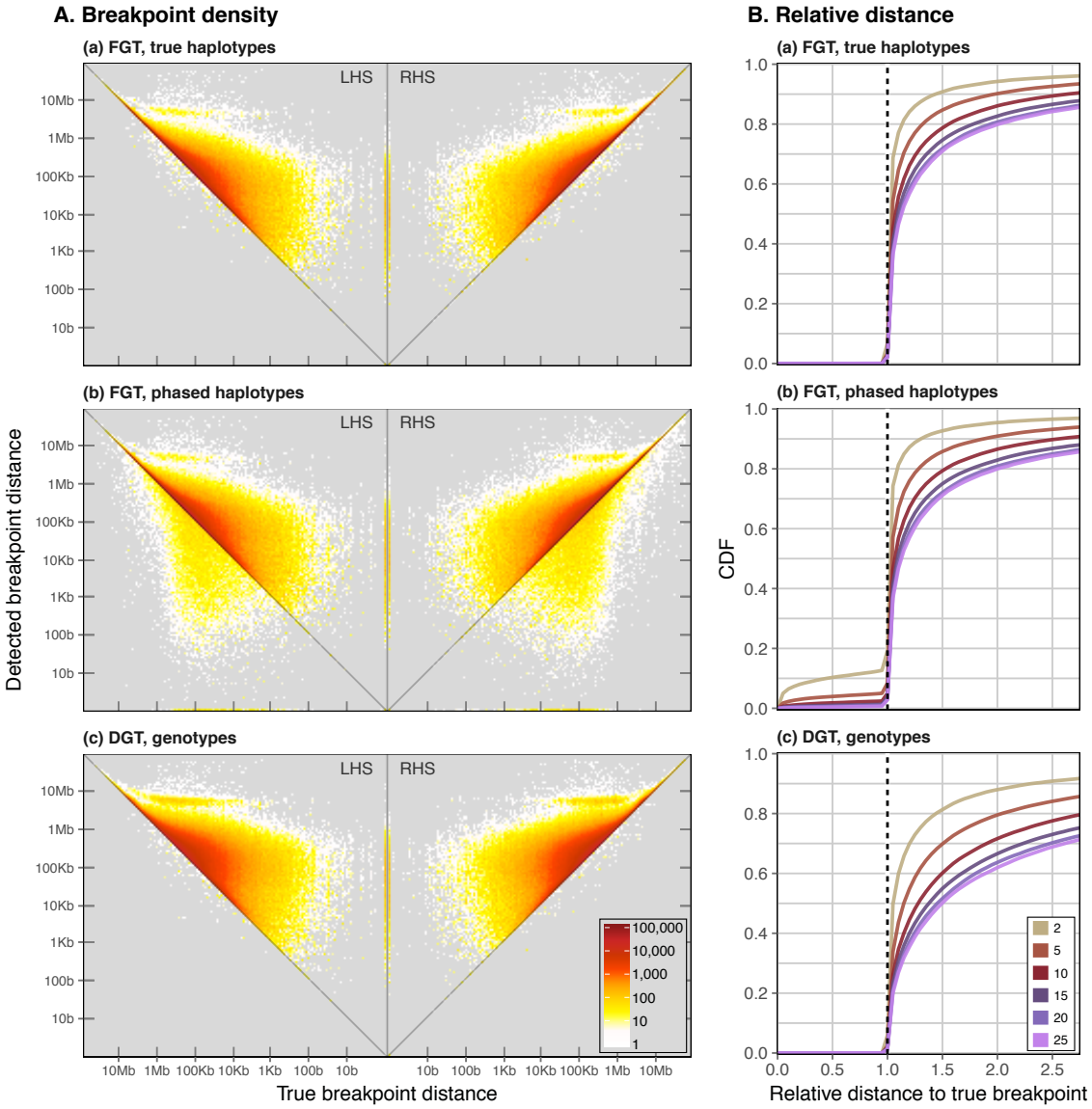


Figure 3.9: Accuracy of breakpoint detection in simulated data. Breakpoints detected in f_k pairs at $k \in \{2, \ldots, 25\}$ are compared to true IBD breakpoint sites, after removing boundary cases in either the detected or true dataset. Segments were inferred using the FGT on true haplotypes **(a)** and phased haplotypes **(b)**, as well as the DGT on genotype data **(c)**. Panel **(A)** illustrates the relationship between each detected breakpoint and the corresponding true breakpoint, measured as the physical distance to the focal site. Along each axis, distances were pooled into 200 bins (on log scale) and cells in the resulting 200^2 grid were colour-coded for the number of intersecting true and detected breakpoints, where grey indicates zero. Segment breakpoints to the left (*LHS*) and right-and side (*RHS*) of the focal position are shown separately. Panel **(B)** shows the cumulative distribution function (CDF) of detected breakpoints in relative distance to the focal and true breakpoint sites. The physical distance between detected breakpoint and focal position was divided by the distance between true breakpoint and focal position, such that values < 1 indicate underestimation and > 1 overestimation relative to the true distance (*dashed line*).

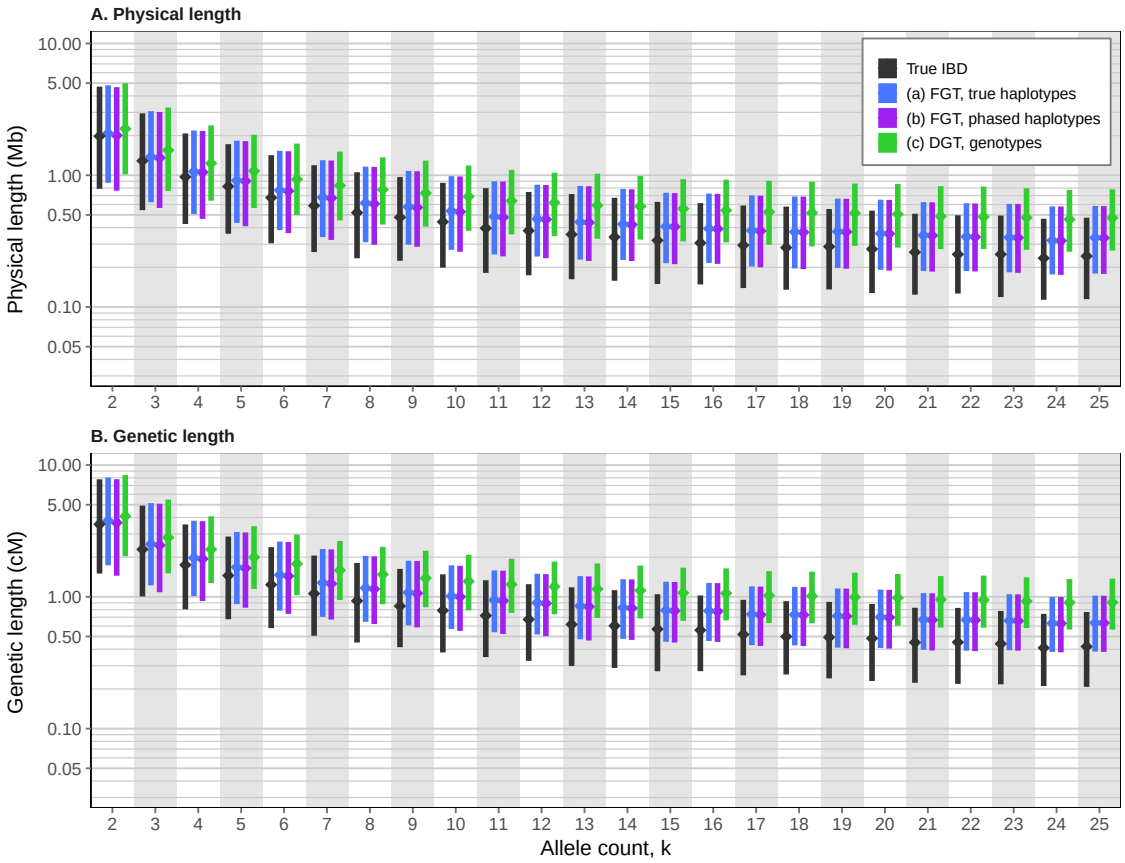


Figure 3.10: IBD segment lengths inferred in simulated data. The distribution of median physical and genetic length of detected IBD segments is shown by allele frequency of the focal variant ($f_{[2,25]}$). IBD detection was performed using the FGT on true and phased haplotypes, as well as the DGT on genotype data; Approaches (a), (b), and (c), respectively. The true IBD length is shown for comparison. Bottom and top of each bar indicate 1st and 3rd quartiles, respectively, between which the median (2nd quartile) is marked (*diamonds*).

lengths along the allele frequency range. However, the gap between true and detected lengths increased towards higher allele frequencies. For example, for f_2 variants, median length of true IBD segments was 1.978 Mb (3.551 cM), which is only marginally shorter compared to 2.079 Mb (3.773 cM), 2.005 Mb (3.652 cM), and 2.256 Mb (4.085 cM) in (a), (b), and (c), respectively. For f_{25} variants the difference was more pronounced; *i.e.* median length of true IBD segments was 0.243 Mb (0.419 cM), compared to 0.336 Mb (0.636 cM), 0.335 Mb (0.634 cM), and 0.475 Mb (0.907 cM) in (a), (b), and (c), respectively.

In summary, the FGT on true haplotype data in Approach (a) overall achieved the highest levels of accuracy while maintaining low error. This was particularly seen in comparison to Approach (b), which differed only in the additionally included phasing step. Since genomic datasets were typically composed of phased haplotypes, Approach (b) can be seen as being the realistic approach. However, the higher error rate at lower

frequency variants may pose a problem for analysis for rare variants. As an alternative, the DGT on genotype data, Approach (c), can be used to detect IBD with high accuracy and comparatively low error rates. However, the larger proportion of overestimated IBD breakpoints may result in additional error, *e.g.* if it is assumed that the genealogy is consistent along the sequence of inferred IBD segments.

3.5.0.1 IBD detection using the *Refined IBD* method

Simulated data were additionally analysed using the `Refined IBD` algorithm implemented in `Beagle` version 4.1 (Browning and Browning, 2013).* The method is based on the non-probabilistic `GERMLINE` algorithm (Gusev *et al.*, 2009), which identifies putative IBD segments from short exact matches between haplotype pairs. Candidate segments are then found by extending identified regions to longer inexact matches. In `Refined IBD`, an additional probabilistic approach is included to assess candidate segments conditional on the likelihood ratio (LR) of the data, calculated under IBD and non-IBD models. A logarithm of odds (LOD) score is calculated as $\log_{10}(\text{LR})$, and segments are reported as IBD if the LOD score is above a specified threshold. This approach has been found to achieve greater accuracy than `GERMLINE` alone or `fastIBD`, which is a non-probabilistic method that detects IBD based on haplotype frequency (Browning and Browning, 2011, 2013).

The method requires haplotype data and cannot be executed with genotype information alone. In the following, Approach (a) refers to the analysis conducted on true haplotypes and Approach (b) on phased haplotypes. The analysis was performed using default parameters in `Refined IBD` (retaining candidate segments at LOD > 3.0) and after conversion of simulated data into Variant Call Format (VCF).†

The purpose of the following analysis was to evaluate whether `Refined IBD` could be used as a method to detect recombination breakpoints and thereby the length of the underlying shared haplotype. This was done by reference to the set of true IBD segments that was determined from simulation records for the set of previously analysed target sites at all $f_{[2,25]}$ variants found in the data (allele frequency $\leq 0.5\%$). However, note that the detection approach employed by `Refined IBD` reports all segments inferred for a given pair of haplotypes, such that detected and true intervals cannot be matched by direct reference to a particular target site. Hence, for each pair of haplotypes present in both sets (detected and true), it was necessary to match segments based on their intervals. As a consequence, it was not possible, for example, to make statements about IBD segments falsely identified by `Refined IBD`, as these would be among the segments removed in the matching process.

* Beagle 4.1: https://faculty.washington.edu/browning/beagle/beagle.html [Date accessed: 2016-11-22]
† Variant Call Format: http://vcftools.sourceforge.net/VCF-poster.pdf [Date accessed: 2016-11-22]

In Approach (a), the analysis returned 13.689 million IBD segments at 6.911 million haplotype pairs. A similar number was returned in Approach (b), where 13.647 million segments were found at 6.856 million pairs. The haplotype pairs at which IBD could be inferred differed between these results; for example, 4.378 million pairs were present in both datasets. The set of available true IBD segments contained 11.598 million intervals at 2.638 million pairs. The number of pairs matched between each detected set and the true set was 2.332 million in (a) and 1.661 million in (b).

The lower number of matched pairs in Approach (b) was due to mismatched haplotypes resulting from the phasing process. Note that it was straightforward to account for haplotype mismatches in the previous analysis, where the focal haplotype could be identified from a given target allele, but which was not possible in the present analysis. It would be possible, for example, to match segments by pair of individuals (instead of haplotype pair) to avoid haplotype mismatches due to phasing. This would introduce a bias when evaluating the accuracy of detected haplotype breakpoints due to possible overlaps of shared haplotypes within the same pair of individuals. To circumvent this bias, such segments were randomly sampled per individual pair if an overlap was found in the set of results returned in the analysis on the phased dataset, so as to allow the identification of the correct true IBD segment based on matching pairs of individuals. This removed 0.145 million detected IBD segments (1.06 %) in Approach (b), but increased the number of segments that could be matched to the set of true segments (1.887 million pairs).

In the following, two analyses were performed. First, the sets of detected IBD segments in Approaches (a) and (b) were matched to the set of true segments based on interval overlap. The proportion of overlap was measured relative to both the inferred and true segments, where segments were ignored if none of the inferred intervals overlapped with any of the true segments available for a given pair. Second, to facilitate comparisons to the targeted IBD detection method evaluated in the previous section, where the accuracy of breakpoint detection was measured in relation to a given target site, inferred segments were matched to the set of true segments based on the inferred interval containing a given target site.

When intervals were matched by overlap, on average, 97.8 % of an inferred interval overlapped with a true shared haplotype in Approach (a), but only 43.0 % of a given true interval was covered by an inferred segment on average. This was similar in (b); 97.4 % and 41.6 %, respectively. The density of overlap measured relative to both the inferred and true segments is shown in Figure 3.11 (next page). These results indicate that the length

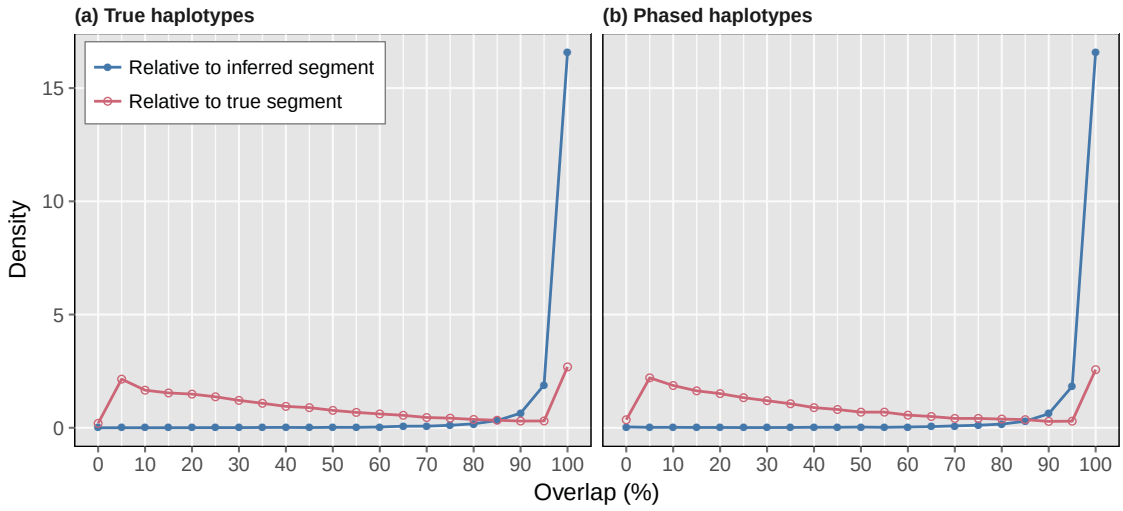


Figure 3.11: IBD segment overlap inferred using *Refined IBD* in Beagle 4.1. Each of the inferred IBD segments were aligned with each of the true segments determined for a given pair to measure the proportion of total base overlap; interval comparisons with zero overlap were ignored. The results shown were generated on a random subset of 10,000 pairs in Approaches (a) and (b). The reported densities refer to the proportion of overlap with respect to the inferred segment (*blue*) and the true segment (*red*).

of segments detected by `Refined IBD` were more likely to be underestimated, but where it is possible that the underlying shared haplotype may be covered by multiple, shorter segments. For example, an average of 1.137 unique segments per pair was known from simulation records, but 1.981 and 1.971 segments were inferred per pair on average in (a) and (b), respectively.

Next, the sets of inferred IBD segments returned in Approaches (a) and (b) were matched to the set of true intervals by the condition that a given target site was contained in the inferred interval. The matching process resulted in 9.173 million segments in (a), but which were reduced to 2.108 million by removing duplicate intervals per pair. Likewise, 8.959 million segments were matched in (b), which was reduced to 2.084 million.

In reference to the matched true IBD intervals, 47.0 % and 46.0 % of the detected breakpoints were overestimated in Approaches (a) and (b), respectively, while 49.9 % and 51.1 % were underestimated. Differences due to phasing were seen at lower frequency target alleles, *e.g.* at f_2, for which 48.4 % were underestimated in (a) but 58.8 % (b). The accuracy of breakpoint detection using `Refined IBD` is illustrated in Figure 3.12 (next page).

The density of true and detected breakpoints (Figure 3.12A) suggests that the breakpoints inferred using `Refined IBD` were closely distributed around the corresponding true breakpoints. However, overall accuracy was low in both (a) and (b), reaching

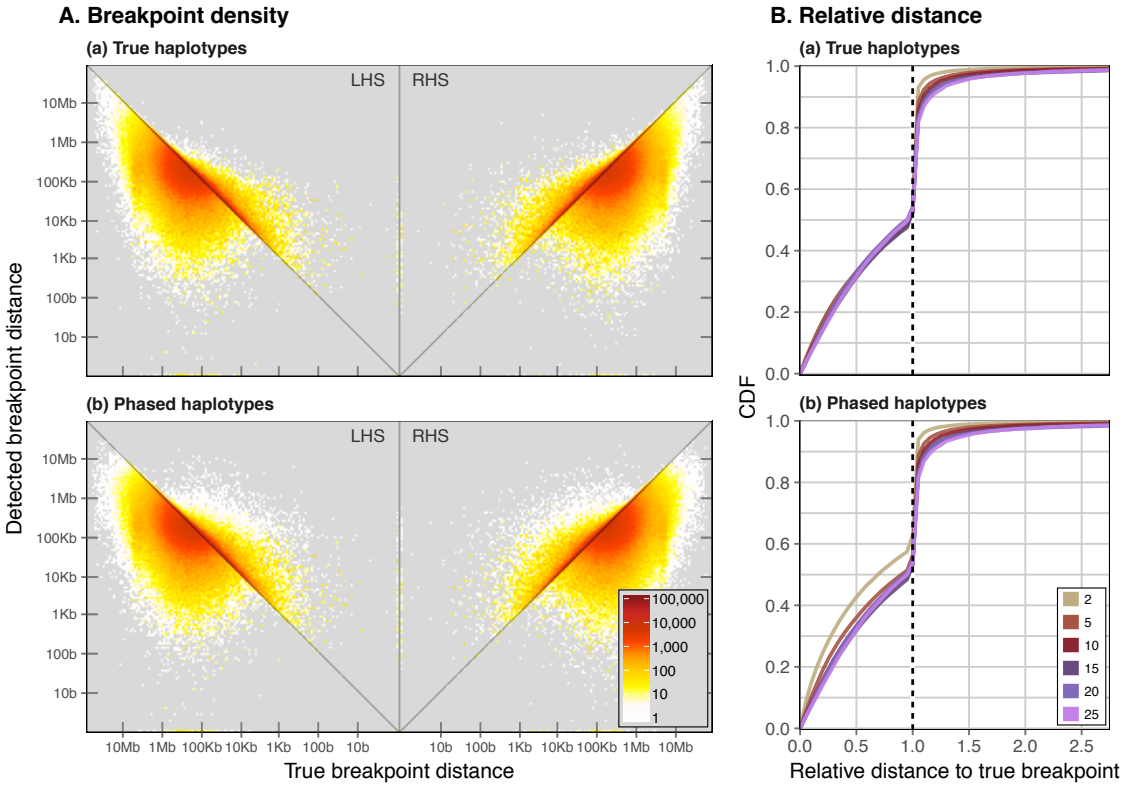


Figure 3.12: Accuracy of breakpoint detection using *Refined IBD* in Beagle 4.1. Results are shown for shared haplotype segments inferred using the `Refined IBD` method; after the detected segments were matched to the set of true segments based on a given target site being contained within the interval of the detected segment. IBD was inferred on true (simulated) haplotype data **(a)** and phased haplotypes **(b)**. Panel **(A)** provides a heatmap representation of a scatter plot, comparing physical distances between focal site and true breakpoint (x-axis) and detected breakpoint (y-axis). Segment breakpoints to the left (*LHS*) and right-and side (*RHS*) of the focal position are shown separately. Panel **(B)** shows the CDF of detected breakpoints in relative distance to the focal and true breakpoint sites.

$r^2 = 0.354$ and $r^2 = 0.209$, respectively. The magnitude of error, RMSLE, was similar in both (a) (b); 0.534 and 0.548, respectively. When true haplotypes were analysed, Approach (a), accuracy decreased steadily towards higher allele frequencies. For example, accuracy was highest for f_2 variants ($r^2 = 0.522$) but lowest for f_{25} variants ($r^2 = 0.080$). The magnitude of error was similar across allele frequencies, *e.g.* at f_2 (RMSLE = 0.571) and f_{25} variants (RMSLE = 0.523). When haplotypes were phased, Approach (b), error was increased at f_2 variants (RMSLE = 0.706) in comparison to f_{25} variants (RMSLE = 0.527). The higher error at lower allele frequencies was also reflected in r^2 values; *e.g.* accuracy was low at f_2 ($r^2 = 0.164$) and highest at f_3 ($r^2 = 0.215$), but lowest at f_{25} ($r^2 = 0.072$). The difference between true and phased datasets is further highlighted in Figure 3.12B, where a higher proportion of f_2 variants is seen to be underestimated in Approach (b).

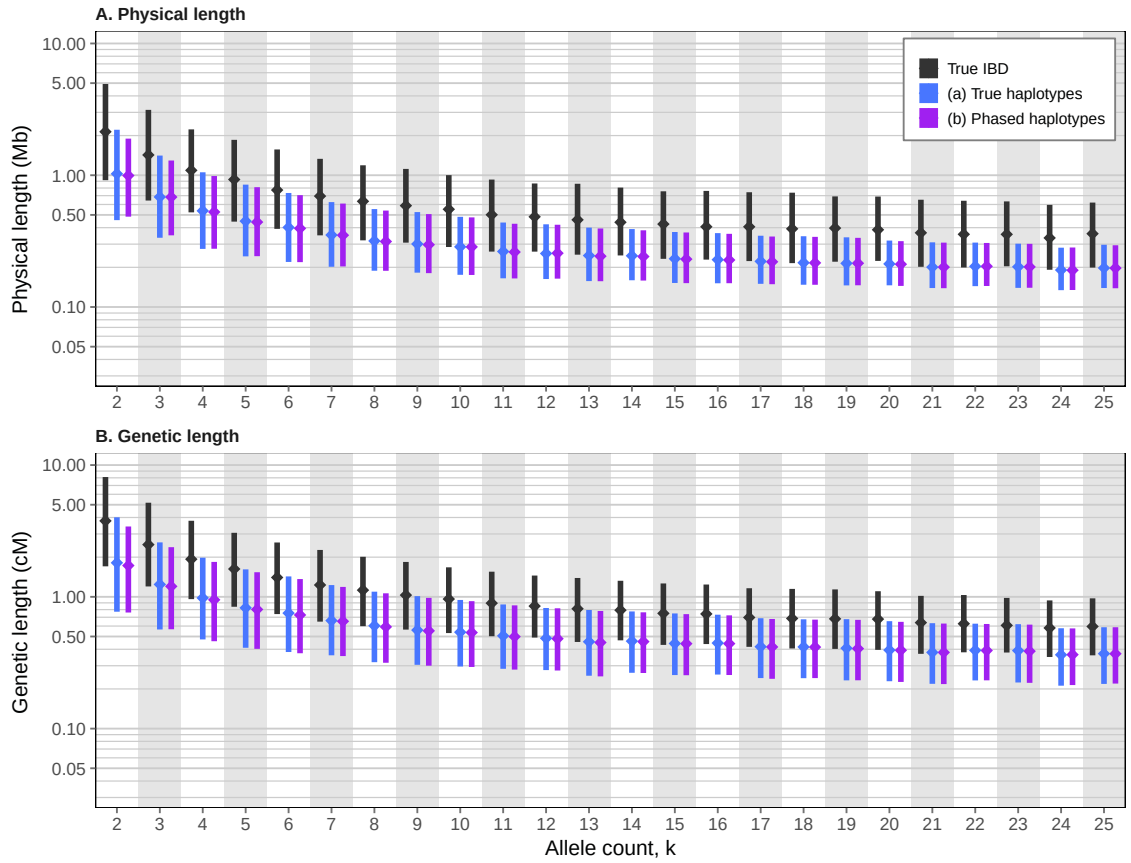


Figure 3.13: IBD segment lengths inferred using *Refined IBD* in Beagle 4.1. The distribution of median physical *(A)* and median genetic *(B)* segment length is shown by allele count (f_k category). IBD segments were inferred using the `Refined IBD` algorithm implemented in `Beagle 4.1`, using true (simulated) haplotype data (a) and phased haplotype data (b). Bottom and top of each bar indicate 1st and 3rd quartiles, respectively, between which the median (2nd quartile) is marked (*diamonds*).

The distribution of physical and genetic lengths for the segments retained in Approaches (a) and (b) are in shown in Figure 3.13 (this page), in relation to the true IBD lengths at each f_k category. Because (a) and (b) were compared on different sets of detected segments, the reported lengths of true segments were computed from the set matched to (a). Boundary cases were removed to avoid potential bias in length comparisons; 1.68 % and 1.67 % in (a) and (b), respectively.

Overall median physical length (and median genetic length) was 0.240 Mb (0.458 cM) in (a) and 0.238 Mb (0.453 cM) in (b), but both were shorter in comparison to the set of true IBD segments at 0.450 Mb (0.781 cM). At f_2 variants, the median length of IBD segments inferred in (a) was 1.026 Mb (1.812 cM), which was longer compared to (b), where median length was 0.995 Mb (1.728 cM). However, both were considerably shorter in comparison to the true segments around f_2 variants at 2.138 Mb (3.770 cM). This

difference persisted towards higher allele frequencies; *e.g.* for f_{25} variants; 0.198 Mb (0.370 cM) in (a) and 0.197 Mb (0.370 cM) in (b), compared to the median length of the matched true segments at 0.360 Mb (0.596 cM).

These results suggested that the lengths of inferred breakpoint intervals are likely to be shorter than the underlying haplotype region shared by descent. Thus, the `Refined IBD` algorithm is less accurate with regard to the inference of the recombination breakpoints that delimit the underlying IBD tract. It was not possible in this analysis to evaluate whether IBD was inferred incorrectly, as incorrect segments would have been removed in the matching process between the sets of inferred and true shared haplotype segments that were determined from simulation records. However, by also evaluating the relative proportion of total base overlap between inferred and true intervals, it was indicated that IBD detection using `Refined IBD` is likely to result in multiple, shorter segments along the length of the underlying shared haplotype.

3.5.0.2 IBD detection in real data: 1000 Genomes, chromosome 20

The IBD detection methodology developed in this chapter (FGT and DGT) was applied to the final release dataset of the 1000 Genomes Project Phase III (1000 Genomes Project Consortium *et al.*, 2012, 2015), which included $N = 2{,}504$ individuals. IBD detection was performed for each autosome (chromosomes 1–22), where selected target sites comprised all variants found at allele frequency $\leq 0.5\%$; *i.e.* f_k where $k \in \{2, \ldots, 25\}$. However, to facilitate a closer comparison to the results obtained on the simulated dataset, which used a variable recombination rate as provided by the Built 37 HapMap Phase II genetic map for chromosome 20 (see Section 3.4.1.2, page 91), the following results focus on chromosome 20 only. A summary of the IBD detection results for chromosomes 1–22 is given in Table 3.2 (next page).

Data were available as phased haplotypes, which enabled the analysis using both the FGT and DGT; *i.e.* the results produced can therefore be seen as being analogous to Approach (b) and Approach (c), respectively. In each, 18.0 million IBD segments were inferred, of which 43.2 % were unique for the FGT, and 39.4 % for the DGT. After removal of boundary cases (0.194 % and 0.285 % for the FGT and DGT, respectively), data were intersected to retain a common set of target sites in the analysis, which retained 7.069 million unique segments.

As there is no "truth" dataset that could serve as a reference to measure accuracy, the following analysis was limited to the quantitative description of the inferred IBD lengths. These results are shown in Figure 3.14 (page 105). Median physical length

Table 3.2: Inferred IBD length per chromosome in 1000 Genomes. Shared haplotype segments in 1000G Phase III were inferred using the FGT and DGT, on data from 2,504 individuals across all autosomes. Pairwise shared segments were identified from rare variants at allele frequency ≤ 0.5% ($f_{[2,25]}$). Median genetic and physical lengths over all inferred segments were calculated per chromosome, after removing boundary cases and retaining unique segments only.

Chr.	SNPs	Targets	Segments	Unique (%)		Length (Mb)		Length (cM)	
				FGT*	DGT**	FGT*	DGT**	FGT*	DGT**
1	6,196,151	2,126,720	64,449,399	40.3	35.9	0.125	0.237	0.150	0.300
2	6,786,300	2,323,889	70,274,554	38.1	33.8	0.136	0.248	0.143	0.280
3	5,584,397	1,893,872	57,220,884	37.1	33.2	0.138	0.243	0.154	0.290
4	5,480,936	1,847,521	57,598,118	36.6	32.8	0.138	0.247	0.150	0.283
5	5,037,955	1,716,580	53,055,802	36.4	32.8	0.139	0.245	0.158	0.293
6	4,800,101	1,625,828	50,544,859	37.0	33.0	0.133	0.238	0.148	0.280
7	4,517,734	1,546,940	47,303,666	39.2	34.8	0.119	0.218	0.139	0.270
8	4,417,368	1,519,028	46,250,487	37.3	33.4	0.119	0.212	0.140	0.268
9	3,414,848	1,171,960	35,718,922	40.6	36.6	0.110	0.203	0.156	0.296
10	3,823,786	1,313,699	40,488,078	39.6	35.3	0.114	0.210	0.154	0.299
11	3,877,543	1,318,559	39,668,383	38.3	34.2	0.128	0.228	0.148	0.283
12	3,698,098	1,255,880	38,116,079	39.4	35.3	0.124	0.221	0.164	0.311
13	2,727,881	919,222	28,252,993	38.9	35.2	0.126	0.222	0.166	0.305
14	2,539,149	861,549	25,955,712	39.5	35.6	0.119	0.214	0.157	0.299
15	2,320,474	795,882	23,977,630	42.6	38.2	0.100	0.183	0.153	0.304
16	2,596,072	901,185	26,907,909	43.5	38.3	0.081	0.153	0.140	0.286
17	2,227,080	775,133	22,914,233	44.5	39.8	0.096	0.175	0.150	0.300
18	2,171,378	739,822	22,405,301	41.5	37.7	0.109	0.193	0.169	0.311
19	1,751,878	607,451	18,033,860	46.1	41.3	0.079	0.146	0.147	0.293
20	1,739,315	599,065	18,040,053	43.2	39.4	0.102	0.180	0.182	0.339
21	1,054,447	365,330	11,051,666	44.7	40.4	0.090	0.172	0.162	0.312
22	1,055,454	363,748	10,748,355	47.2	42.5	0.070	0.133	0.145	0.291
Total	77,818,345	26,588,863	808,976,943						

* 1000G data are available as phased haplotypes; hence, results are analogous to Approach (b).
** Conducted on genotype data; hence, results are analogous to Approach (c).

(and median genetic length) over the whole set of retained IBD segments was 0.101 Mb (0.188 cM) using the FGT and 0.180 Mb (0.339 cM) using the DGT. As was seen in the analysis of simulated data, the DGT generally is more likely to overestimate breakpoint distance, leading to the discovery of longer intervals. This discrepancy in length was more pronounced for f_2 variants, for which median length was 0.124 Mb (0.195 cM) using the FGT and 0.253 Mb (0.428 cM) using the DGT. Notably, IBD lengths were more than twice as long in half of the detected segments using the DGT, compared to the FGT. The length of segments identified at lower frequencies was longer in comparison to higher frequencies; *e.g.* for f_{25} variants, median length was 0.084 Mb (0.165 cM) and 0.149 Mb (0.292 cM) using the FGT and DGT, respectively. However, the IBD lengths were highest at $f_{[3,5]}$ when the FGT was used, but which was not the case for the DGT.

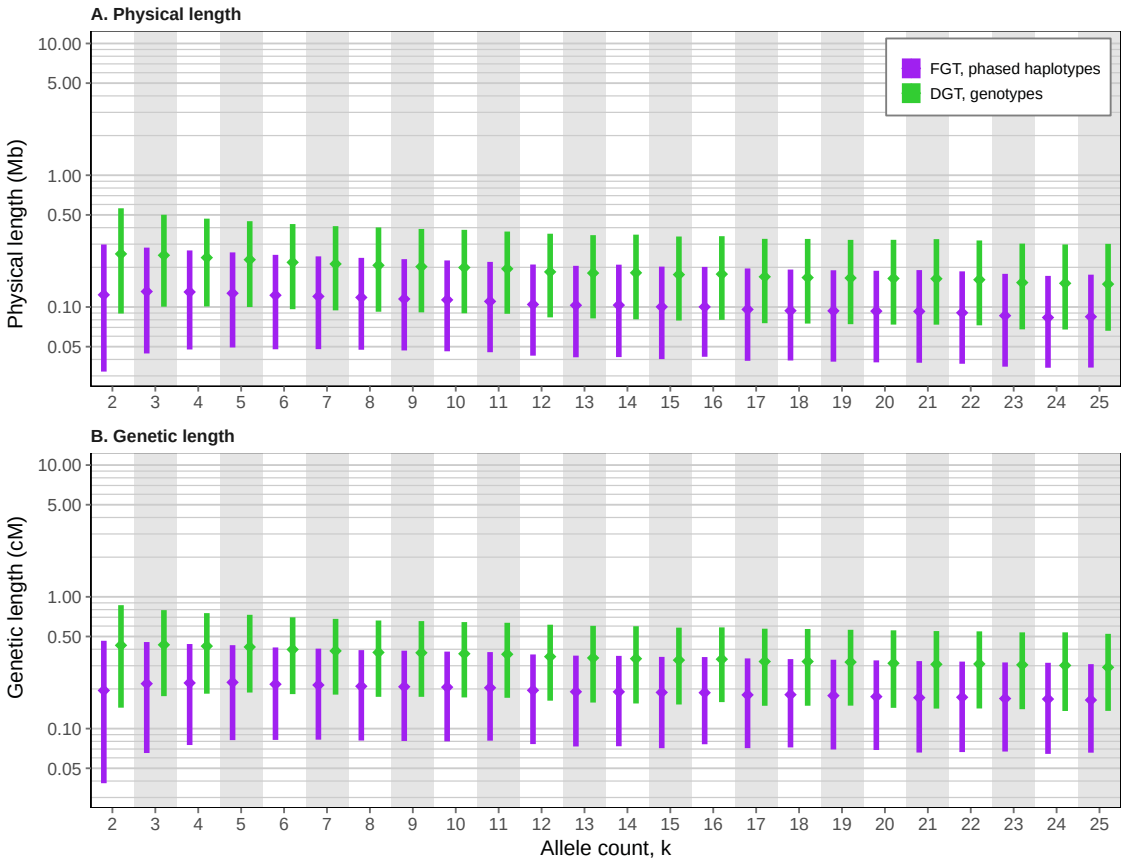


Figure 3.14: Distribution of inferred IBD lengths in 1000 Genomes data, chr. 20. Results are shown for the detected physical and genetic lengths of shared haplotype segments by f_k, using chromosome 20 in the final release dataset of 1000 Genomes Project Phase III, including $N = 2{,}504$ individuals. IBD segments were detected using the FGT (on phased haplotypes) and the DGT (on genotype data). Bottom and top of each bar represent the 1st and 3rd quartile, respectively, between which the median (2nd quartile) is marked (*diamonds*).

The main observation from applying the FGT and DGT to real data is that the detected shared haplotype segments appear to be shorter than suggested by the previous analysis on simulated data. It is possible that a large number of segments were underestimated due to the detection of false positive breakpoints; *e.g.* through violations of the infinite sites model or other sources of error, which can be expected to be present in real data such as 1000G. The example shown in Figure 3.15 (next page) may support this notion. One of the selected target sites was chosen at random and each pair of individuals sharing the focal allele were re-analysed to record all positions at which the a breakpoint was found relative to the focal allele. In each pair, it can be seen that a few, isolated breakpoints appear within longer, unbroken regions, followed by a quick succession of detected breakpoints. This pattern can be compared to Figure 3.6a (page 87), which showed a similar example from the simulated dataset. Clusters of breakpoints were (generally)

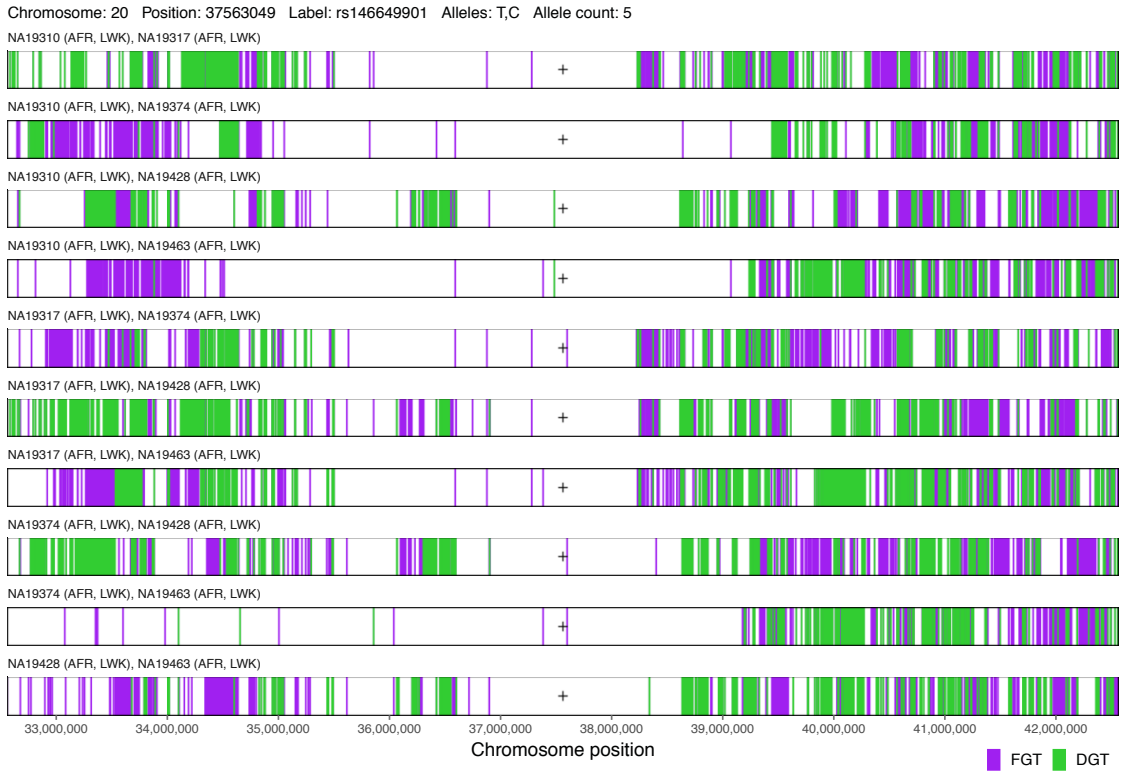


Figure 3.15: Example of breakpoints detected in 1000 Genomes, chr. 20. One of the target sites analysed was randomly selected and breakpoint detection was performed for all pairs of individuals sharing the focal allele. The plot shows all breakpoints detected using the FGT and DGT relative to the allelic configuration observed at the target site (*cross*), within a 10 Mb region around the target position. Sample IDs (and population codes) as found in 1000 Genomes data are shown on the top of each pairwise analysis. Note that any breakpoint detected using the DGT also implies detection through the FGT.

detected at some distance away from the true, underlying recombination breakpoint, whereas single, isolated breakpoints were rarely observed in simulated data. It is therefore possible that the detection of short shared haplotype intervals in 1000G data may indeed be due to false positive breakpoints.

3.6 Discussion

In this chapter, I presented a novel IBD detection method which is able to infer recombination events in both haplotype and genotype data. To be able to apply this method on a larger scale, I implemented the IBD detection algorithm described in this chapter as a computational tool written in `C++`; called **`tidy`** (targeted IBD detection done thoroughly).*

* Targeted IBD detection done thoroughly, `tidy`: https://github.com/pkalbers/tidy

Although the FGT showed overall high levels of accuracy, phasing error was identified as a problem. Current phasing methods such as `SHAPEIT2` typically show very low error rates (O'Connell *et al.*, 2014). However, occasionally, alleles are placed on the wrong haplotype. This may happen at single loci (*flip errors*) or such that longer haplotype stretches are exchanged (*switch errors*). Both types of error can affect breakpoint detection under the FGT as both flip and switch errors may change the configuration of alleles observed in relation to a given focal variant. As an alternate solution to using phased haplotypes, the DGT can be used on genotype data, as it is not affected by phasing error. However, the lengths of detected IBD segments tend to be overestimated.

The IBD results obtained from analysis of the 1000G dataset suggested that the FGT was similarly affected by phasing error as seen in the simulation analysis. However, the DGT was also affected by additional sources of error. One consideration is that both the FGT and DGT assume the infinite sites model, but which is only an approximation to the conditions observable in nature. In particular, back mutations and recurrent mutations are excluded in the model, but these are prevalent in the (human) genome. For instance, recurrent mutations can produce patterns of variation that would otherwise only be observable if recombination had occurred (McVean *et al.*, 2002). Thus, false positive breakpoints may be inferred, such that IBD length is underestimated. Nonetheless, the infinite sites model is usually seen as a reasonable approximation to reality, as the number of variant sites in a sample is typically much smaller than the number of nucleotides in the chromosomal sequence (Hein *et al.*, 2004).

The presence of error in real data cannot be ruled out; in particular, given the known error rates in current sequencing and genotyping technologies, imperfect statistical methods used in different pipelines for data generation, such as genome assembly, variant calling, and filtering strategies, as well as human error in data processing. Hence, error in 1000G data is practically guaranteed to be present and likely to have had an impact on the accuracy of the IBD detection methodology presented in this chapter. It is therefore unlikely that the rule-based approach presented here could be used reliably when working with biological datasets. In the following chapter, I focus on the analysis of error in different datasets, and I use this information to develop a novel, probabilistic method for shared haplotype detection around target sites.

The first principle is that you must not fool yourself –
and you are the easiest person to fool.

— Richard Feynman

4

Consideration of genotype error in the inference of haplotype sharing by descent

Contents

4.1 Introduction

Recent advancements in genotyping and next-generation sequencing (NGS) technologies have enabled us to study the human genome in unprecedented detail and scale. The availability of high-throughput methods to survey large samples has led to successful identification of thousands of disease causing risk factors, which in particular was driven by genome-wide association (GWA) studies. This explosion of human genetic data has

further enabled collaboration initiatives through the setup of genetic databases, which can be queried by research groups worldwide. However, because no technology is perfect, acquired data are likely to contain undetected amounts of error, which may affect statistical inference in many ways.

Statistical tests often rely on the assumption that genotype data (retained after quality control) are correct, or that error quantities are negligible. Yet, the effects of misclassification in genotype data are well documented. For example, it has been shown that even minor amounts of genotype error can distort estimated distances in linkage mapping studies (Buetow, 1991; Shields *et al.*, 1991; Sobel *et al.*, 2002), result in a substantial loss of linkage information in quantitative trait analyses (Douglas *et al.*, 2000; Abecasis *et al.*, 2001), decrease power in association studies (Kang *et al.*, 2004), and can substantially increase type I (false positive) error in haplotype-based case-control analyses (Moskvina *et al.*, 2005).

Identification of incorrectly typed or called genotypes remains a difficult problem, which becomes more challenging as the magnitude of data increases. But, for example, as shown by Cox and Kraft (2006) and independently by Moskvina and Schmidt (2006), genotype error theoretically does not always affect the distributions of genotypes to the extent that Hardy-Weinberg equilibrium (HWE) can be violated. Given the common practise to exclude presumably incorrect genotypes based on departures from HWE, it therefore remains difficult to catch falsely called or typed variants.

In this chapter, I explore the impact of genotype error on the detection of identity by descent (IBD) segments and, based on these results, I implement a new approach for targeted IBD inference using a Hidden Markov Model (HMM). First, I introduce a generic model for genotype error; see section below. The remainder of this chapter is then divided into two main parts. In the first part (Section 4.2), I characterise the distribution of genotype error in data obtained on different genotyping and sequencing platforms, to construct empirical error profiles. I use this information to integrate realistic error rates in simulated data, such that the effects of error can be observed in practice. In particular, I evaluate the non-probabilistic IBD detection method presented in Chapter 3. The insights gained from this analysis enabled a probabilistic extension of the targeted IBD detection method, which I implemented using a HMM; I present this new method in the second part of this chapter (Section 4.4). This HMM-based method is incorporated in the previously presented `tidy` algorithm for the targeted detection of IBD segments (see Chapter 3).

4.1.1 Probability of genotype error

Consider a biallelic locus with alleles a or b, which respectively occur at frequency p and $q = 1 - p$ in a population. Genotypes are formed by combination of two alleles in diploid organisms (therefore sometimes referred to as *diplotypes*). There are four possible combinations of alleles, *i.e.* aa, ab, ba, and bb, but of which genotypes ab and ba are indistinguishable. It is convenient to recode the two alleles as 0 and 1 to denote the reference and alternate allele, respectively. By introducing k to count the number of alternate alleles, let g_k denote a genotype, where $k \in \{0, 1, 2\}$. If all combinations of the two alleles are statistically independent, *e.g.* in a randomly mating population, sample genotype frequencies, $f_g(k)$, are multinomially distributed with expectations given by HWE proportions (Hardy, 1908; Weinberg, 1908); *i.e.* such that $(p + q)^2 = p^2 + 2pq + q^2 = 1$. The general form of the expected genotype frequency is given in Equation (4.1), where n refers to the number of chromosome copies (ploidy); *e.g.* $n = 2$ for diploid organisms.

$$f_g(k) = \binom{n}{k} p^{n-k} \, q^k \tag{4.1}$$

In presence of genotype error, the actual, *true* genotype is distinguished from the *observed* genotype, $\tilde{g}_k$, and the observed frequency, $f_{\tilde{g}}(k)$, is different from the true (but unknown) genotype frequency, dependent on the rate of error. More precisely, let the rate at which genotype g_j is classified as $\tilde{g}_i$ be denoted by ε_{ij}, where $i, j \in \{0, 1, 2\}$. The value of ε_{ij} is often referred to as the *penetrance* of a genotype and represents the probability of observing genotype $\tilde{g}_i$ given the true genotype g_j (Ott, 1999; Gordon *et al.*, 2002). In the following, I use the term *error rate* to refer to genotype penetrance. For convenience, error rate parameters can be represented in a 3×3 confusion matrix, $\mathcal{E}$, below.

$$\mathcal{E} = \begin{bmatrix} \varepsilon_{00} & \varepsilon_{01} & \varepsilon_{02} \\ \varepsilon_{10} & \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{20} & \varepsilon_{21} & \varepsilon_{22} \end{bmatrix} \tag{4.2}$$

Considering the relation $\sum_{i=0}^{2} \varepsilon_{ij} = 1 \; \forall \; j$, where $0 \leq \varepsilon_{ij} \leq 1$, it follows that the expected observation frequency of a genotype is

$$f_{\tilde{g}}(k) \; = \; \begin{cases} f_g(0)\, \varepsilon_{00} \; + \; f_g(1)\, \varepsilon_{01} \; + \; f_g(2)\, \varepsilon_{02} & \text{if } k = 0 \\ f_g(0)\, \varepsilon_{10} \; + \; f_g(1)\, \varepsilon_{11} \; + \; f_g(2)\, \varepsilon_{12} & \text{if } k = 1 \\ f_g(0)\, \varepsilon_{20} \; + \; f_g(1)\, \varepsilon_{21} \; + \; f_g(2)\, \varepsilon_{22} & \text{if } k = 2 \end{cases} \tag{4.3}$$

where $i = j$ indicates correct classification and $i \neq j$ misclassification of the true genotype; see Moskvina and Schmidt (2006).

4.1.1.1 Genotype error models

Equations (4.2) and (4.3) provide a generic framework for the error rate of genotypes and the calculation of genotype frequencies after error. Two error models are presented below which provide formulations for the calculation of model parameters ε_{ij}.

Douglas *et al.* (2002) introduced a genotype-based model with parameters γ and η, denoting the probability of a homozygous genotype to be misclassified as a heterozygous genotype and vice-versa, respectively. The intuition behind this model is based on technical error in the polymerase chain reaction (PCR) amplification process, which is used in both genotyping and sequencing methods for the replication of DNA fragments. However, note that observed genotypes $\tilde{g}_0$ and $\tilde{g}_2$ both have equal probability to arise from misclassification of g_1, and the probability that a homozygous genotype appears as the opposite homozygote, g_0 as $\tilde{g}_2$ or g_2 as $\tilde{g}_0$, is zero.

As an alternative, misclassification of genotypes can be modelled as a consequence of errors that occur at random and independently in each of the two alleles. An explicit formulation of an allele-based model was proposed by Gordon *et al.* (2001), where ϵ_0 was defined as the probability that allele 0 (h_0) was observed as allele 1 (h_1), and ϵ_1 the probability that h_1 was observed as h_0.

Table 4.1: Penetrance functions in genotype and allele-based error models. Error probability (or *penetrance*) is denoted by ε_{ij}, which is the probability of observing genotype i given the true genotype j. Two models are presented which are genotype-based and allele-based, respectively. In each model, equations refer to the probability that a true genotype, g_j, was observed as any of the possible genotypes, $\tilde{g}_i$, such that ε_{ij} is calculated from the corresponding row-by-column expression.

Model	Observed genotype	True genotype		
		g_0	g_1	g_2
Genotype-based[1]	$\tilde{g}_0$	$1-\gamma$	$\frac{1}{2}\eta$	0
	$\tilde{g}_1$	γ	$1-\eta$	γ
	$\tilde{g}_2$	0	$\frac{1}{2}\eta$	$1-\gamma$
Allele-based[2]	$\tilde{g}_0$	$(1-\epsilon_0)^2$	$\epsilon_1(1-\epsilon_0)$	${\epsilon_1}^2$
	$\tilde{g}_1$	$2\epsilon_0(1-\epsilon_0)$	$\epsilon_0\epsilon_1+(1-\epsilon_0)(1-\epsilon_1)$	$2\epsilon_1(1-\epsilon_1)$
	$\tilde{g}_2$	${\epsilon_0}^2$	$\epsilon_0(1-\epsilon_1)$	$(1-\epsilon_1)^2$

[1] Douglas *et al.* (2002); $\gamma = P(\text{hom.} \rightarrow \text{het.})$, $\eta = P(\text{het.} \rightarrow \text{hom.})$
[2] Gordon *et al.* (2001); $\epsilon_0 = P(h_0 \rightarrow h_1)$, $\epsilon_1 = P(h_1 \rightarrow h_0)$
Table modified from Gordon *et al.* (2002), Table 2.

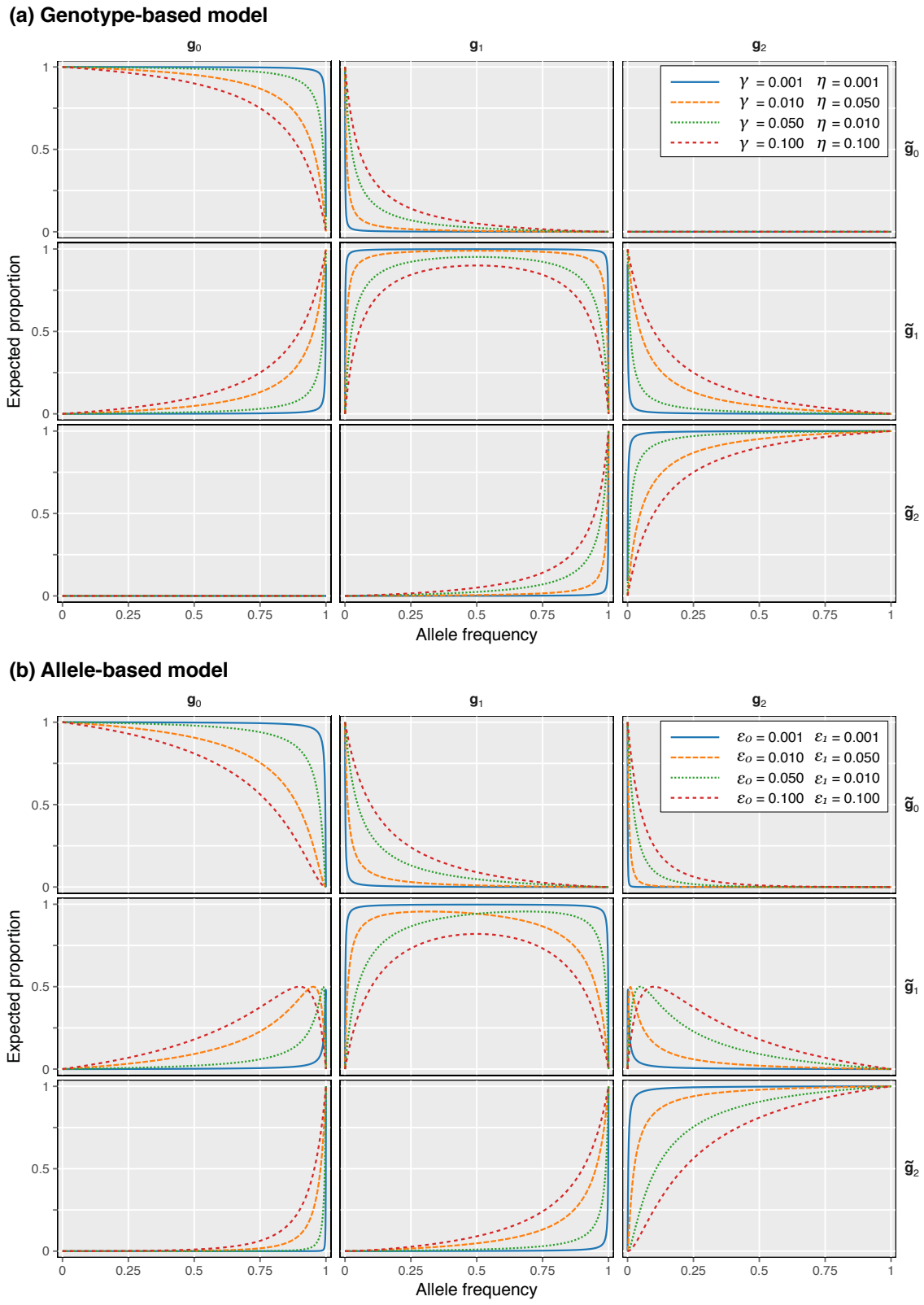


Figure 4.1: Expected error distributions in genotype and allele-based models. The graphs show the expected proportion of true genotype g_j observed genotype as ($\tilde{g}_i$) given the population allele frequency; calculated at different, nominal error rates (see legend). The error functions provided in Table 4.1 (page 112) were used as in Equation (4.3) (page 111) and results were normalised to sum to 1 per true genotype class (columns).

Error functions for both models are given in Table 4.1 (page 112); note that these are arranged as in error matrix $\mathcal{E}$ in Equation (4.2). To illustrate the expectations arising from these models, Figure 4.1 (page 113) shows the expected proportion of a true genotype observed as the same or another genotype, at different, nominal error rates, for both the genotype and allele-based models. It should be noted that the error parameters specified in each model may not apply equally to each variant in genomic data, due to differences arising from technical bias in the sequencing or genotyping process and variations in the accessibility of DNA along the genome (*e.g.* chromatin structure variations near telomeric or centromeric regions).

In the following section, I estimated genotype error rates in different datasets. In each, error was computed from the proportions of correctly and incorrectly classified genotypes, such that error parameters were estimated for each model.

4.2 Generation of platform-specific genotype error profiles

Assessment of genotype accuracy requires the existence of an error-free "gold standard" dataset against which data generated on other platforms can be compared; provided that data were obtained on the same biological sample. In reality, however, the possibility of undetected genotype error cannot be excluded, but it can be reduced, for example, based on pedigree information and the laws of Mendelian inheritance. In the section below, I describe the dataset which I used as a reference for high-confidence genotype data. These were compared to several publicly available datasets generated using different genotyping and sequencing technologies, which included individuals also present in the reference dataset.

4.2.1 High-confidence genome data as benchmark for comparisons

The analysis was based on data from the Illumina Platinum Genomes Project (IPG),* which comprises a 17-member, three-generation family of European ancestry; CEPH pedigree 1463.† This dataset has been generated using recent state-of-the-art sequencing technologies and methods for variant calling, where a total of 5.43 million variants were identified genome-wide (Eberle *et al.*, 2016); this included 4.73 million single-nucleotide polymorphisms (SNP). Individuals had been sequenced to a depth of 50× on

* Illumina Platinum Genomes: http://www.illumina.com/platinumgenomes/ [Date accessed: 2016-11-16]
† Centre d'Etude du Polymorphisme Humain (CEPH), Utah family pedigree 1463: https://catalog.coriell.org/0/Sections/Collections/NIGMS/CEPHFamiliesDetail.aspx?fam=1463 [Date accessed: 2016-11-16]

Illumina HiSeq 2000, and variants were called in concordance to several variant calling methods. Notably, due to the availability of pedigree information, artefacts such as genotype errors had been excluded based on deviations from Mendelian inheritance. The dataset comprises 11 children from two parents, who themselves are the children of the four founders of the pedigree; see Figure 4.2 (this page). Thus, inheritance constraints were most informative for the two parents, labelled NA12877 and NA12878 (Coriell ID), which were additionally sequenced to 200× depth, and for which high-confidence variant calls were made available.

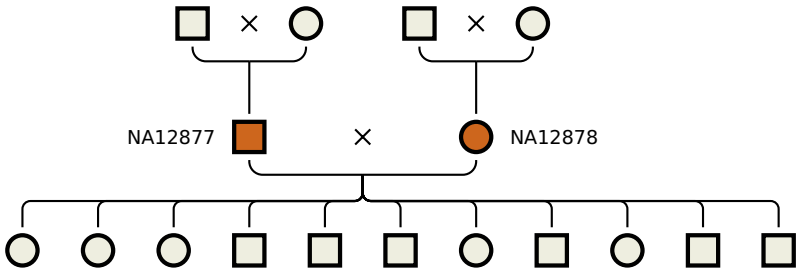


Figure 4.2: CEPH pedigree 1463. The pedigree of the family sequenced in the Illumina Platinum Genomes Project. Genotype data of individuals NA12877 and NA12878 (indicated) were used as reference against which data obtained on other genotyping or sequencing platforms were compared. Figure modified from Eberle *et al.* (2016), Figure 1.

Genotype (SNP) data from IPG for individuals NA12877 and NA12878 were used as reference or *truth* for comparison to concordant data obtained on other platforms. Although the possibility of genotype error in IPG data cannot be excluded, it is assumed that error rates in NA12877 and NA12878 are sufficiently low to allow proportional estimation of genotype misclassification rates based on observations over thousands of variant sites.

Due to the imperfection of even high-standard sequencing technologies, not all chromosomal regions are equally accessible, which affects the power to determine variants in the calling process along the length of the sequence. The confidence of variant calls is derived from the depth of mapped sequence reads and quality scores. To maintain high levels of confidence in the data, accessibility masks provided by IPG were applied such that only sites in high-confidence regions were retained in the analysed datasets. This retained a sum of 3.407 million and 3.605 million SNPs for NA12877 and NA12878, respectively, across chromosomes 1–22.

4.2.2 Selection and preparation of datasets from different platforms

Because cell lines from CEPH pedigree 1463 are a well-characterised model system, either NA12877 or NA12878, or both, have been assessed in several studies. For example, CEPH pedigree 1463 was genotyped in the International HapMap Project, which was one of the first large-scale catalogues of human genetic variation (International HapMap Consortium, 2003; International HapMap Consortium *et al.*, 2007; International HapMap 3 Consortium *et al.*, 2010). Considering a more recent example, the 1000 Genomes Project provides data obtained on several platforms, including whole-genome sequencing (WGS) and high-density genotyping technologies (Altshuler *et al.*, 2010; 1000 Genomes Project Consortium *et al.*, 2012, 2015).

It must be noted that the process of acquiring data is substantially different for genotyping and sequencing methods. The established approach for genotyping is to use chip or array-based methods, which are designed to target, or "type" specific molecular markers at predetermined regions and require prior knowledge about mapped locations in the genome. On the other hand, sequencing determines the contiguous nucleotide sequence, either genome-wide or for a region of the genome. Sequence data are aligned against a reference genome and further processed. Eventually, variants are "called" at nucleotides that differ from the reference at each position along the sequence.

Genotype error profiles were generated for both sequencing and genotyping data, which were taken from available resource data of the 1000 Genomes Project. The following *test* datasets were included:

- Low-coverage sequencing data from the final release of 1000 Genomes Project Phase III (***1000G***), generated on Illumina HiSeq 2000 and HiSeq 2500 platforms (2-4x), and consisting of 78 million SNPs in total.

- Genotyping data generated on Illumina HumanOmni2.5 BeadChip (***Omni2.5***) with 2.46 million SNPs.

- Genotyping data generated on Affymetrix Genome-Wide Human SNP Array 6.0 (***Affy6.0***) with 0.91 million SNPs.

To acknowledge differences arising from the variant calling and filtering process in sequencing data, two *1000G* profiles were created; one that included all variant sites (***1000G.A***), and one containing only sites within high-confidence regions (***1000G.B***). For the latter, the "strict" accessibility mask provided by 1000 Genomes Project Phase III

was used (see 1000 Genomes Project Consortium *et al.*, 2015, supplementary information 9.2).* Note that the sample of the final release dataset of *1000G* included NA12878, but not NA12877. The other two datasets, *Omni2.5* and *Affy6.0*, which were part of previous releases of the 1000 Genomes Project, included both NA12877 and NA12878.†

Figure 4.3: Illustration of the matching process in the generation of error profiles. Variant data were reduced to SNPs and matched per chromosome by variant position and both alleles (A_0 and A_1) as recorded for either NA12877 or NA12878. Gaps shown in the truth dataset indicate the regions removed after filtering using the accessibility mask provided by IPG, such that only high confidence variant calls were retained. Note that the truth dataset did not contain SNPs homozygous for the reference allele, but which were assumed from high-confidence regions if present in the test dataset. This is indicated by left-pointing arrows.

Misclassification of SNP genotypes was determined by comparison of each test dataset to the truth dataset, which was done for chromosomes 1–22. Genotype data were matched by chromosome and variant position (GRCh37/hg19). As a precaution, sites where reference or alternate nucleotides did not match between test and truth datasets were removed, although only genotypes were compared.

* Accessible genome masks in 1000G:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/
[Date accessed: 2016-11-27]

† High-density genotyping data, Omni2.5 and Affy6.0 in 1000 Genomes Project (1000G):
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/
[Date accessed: 2016-11-17]

It is important to note that IPG data did not contain variants called as being homozygous for the reference allele (g_0). Eberle *et al.* (2016) identified high-confidence regions in the 13 individual call sets (`NA12877`, `NA12878`, and their 11 children) by collating sites that were called as being homozygous for the reference allele and monomorphic in the sample. Monomorphic variants that were homozygous for the alternate allele were included. Therefore, the following assumption was made. If the position of a variant site in a given test dataset was within high-confidence regions of the IPG accessibility mask provided by Eberle *et al.* (2016), but not reported in the truth dataset, the true state was assumed to be the g_0 type. This relies on the expectation that the high-confidence intervals comprised data which would have otherwise been reported as a different type. This matching process is illustrated in Figure 4.3 (page 117).

At each matched site, the population frequency was assigned as recorded in the full sample of the final 1000 Genomes Project Phase III dataset, which contained 2,504 individuals from several continental populations worldwide. Sites for which no frequency information was available were removed. Then, the retained genotypes in the matched datasets were used to measure the rate at which a true genotype (g_0, g_1, or g_2) was observed as the same or another genotype ($\tilde{g}_0$, $\tilde{g}_1$, or $\tilde{g}_2$). This was done to obtain estimates for error rate parameters ε_{ij} in matrix $\mathcal{E}$.

4.2.3 Rate of genotype error in sequencing and genotyping data

The total number of matched variant sites was 76.859 million in *1000G.A*, but of which 73.435 million (≈96%) were assumed as homozygous reference genotypes. Recall that this assumption applied only to sites found within the high-confidence regions as specified in the IPG accessibility mask. A lower amount was available in *1000G.B*, where 59.234 million genotypes were retained, but of which 56.739 million (≈96%) were assumed.

This large proportion of sites at which a true g_0 genotype was assumed may not come as a surprise, because there is a high chance that a considerable fraction of the variants present in either test dataset may fall within the lengths covered by high-confidence regions. However, because g_0 genotypes were removed in IPG data, it is a necessary assumption that those genotypes can be recovered from high-confidence regions. Otherwise, error could not be determined for g_0 genotypes. Overall, 0.079% of genotypes were misclassified in *1000G.A*, and 0.025% in *1000G.B*. If assumed g_0 genotypes are ignored, thus only considering true genotype classes $j \in \{1, 2\}$, overall error was increased; reaching 0.538% and 0.183% in *1000G.A* and *1000G.B*, respectively.

Due to the comparatively lower number of available sites in genotyping data (*Omni2.5* and *Affy6.0*), the matched `NA12877` and `NA12878` datasets were merged. Together, the total number of matched genotypes was 4.234 million in *Omni2.5* and 1.716 million in *Affy6.0*, and where 1.361 million (≈32%) and 0.794 million (≈46%) of true genotypes were assumed as being homozygous for the reference allele, respectively. The proportion of misclassified genotypes was 0.256 % in *Omni2.5*, and 0.139 % in *Affy6.0*. Error decreased in both genotyping datasets if g_0 was ignored, yielding 0.068 % and 0.106 % of misclassified genotypes in *Omni2.5* and *Affy6.0*, respectively.

In the following, the error rate of genotypes was investigated in greater detail, for which matched sites from each true genotype class were considered; first, by exploring the distribution of error along the genome and, second, by true genotype class to obtain empirical error rate estimates, which was then extended to generate frequency-dependent error profiles for each dataset.

4.2.3.1 Genotype accuracy by chromosomal region

Each chromosome was divided into 1 Mb long chunks to depict the rate of misclassified genotypes over the length of the genome; see Figure 4.4 (next page). Error densities were calculated by dividing the number of incorrect genotypes by the number of all genomes within each chunk, where chunks with less than 100 matched sites were removed.

The distribution of error in the *1000G.A* dataset was consistently low on average (0.095 %), but where error densities increased towards telomere regions and near centromeres, reaching error rates above 1%. This is expected, because DNA in the telomeric and centromeric regions is highly repetitive and rich in GC content, which results in difficulties in the amplification process and makes sequence reads difficult to align to the genome. This pattern was less pronounced in *1000G.B*, as sites outside high-confidence regions were excluded, which resulted in a clear reduction of error along the genome on average (0.028 %). However, most chromosomes showed locally increased error rates, but where rates above 1% were rarely observed. Yet, the persistence of error hotspots indicates that not all low-confidence regions were identified from quality assessment of sequencing data.

In genotyping data, error rates showed less variability along the genome, *e.g.* in the *Omni2.5* dataset, but where error rates averaged at 0.274 %, with a few regions of increased error above 5%. Although error was low on average in *Affy6.0*, averaging at 0.308 %, the likewise lower number of sites resulted in sparse coverage. However, a few regions showed error rates above 5%, but which were located near the telomeric or centromeric regions.

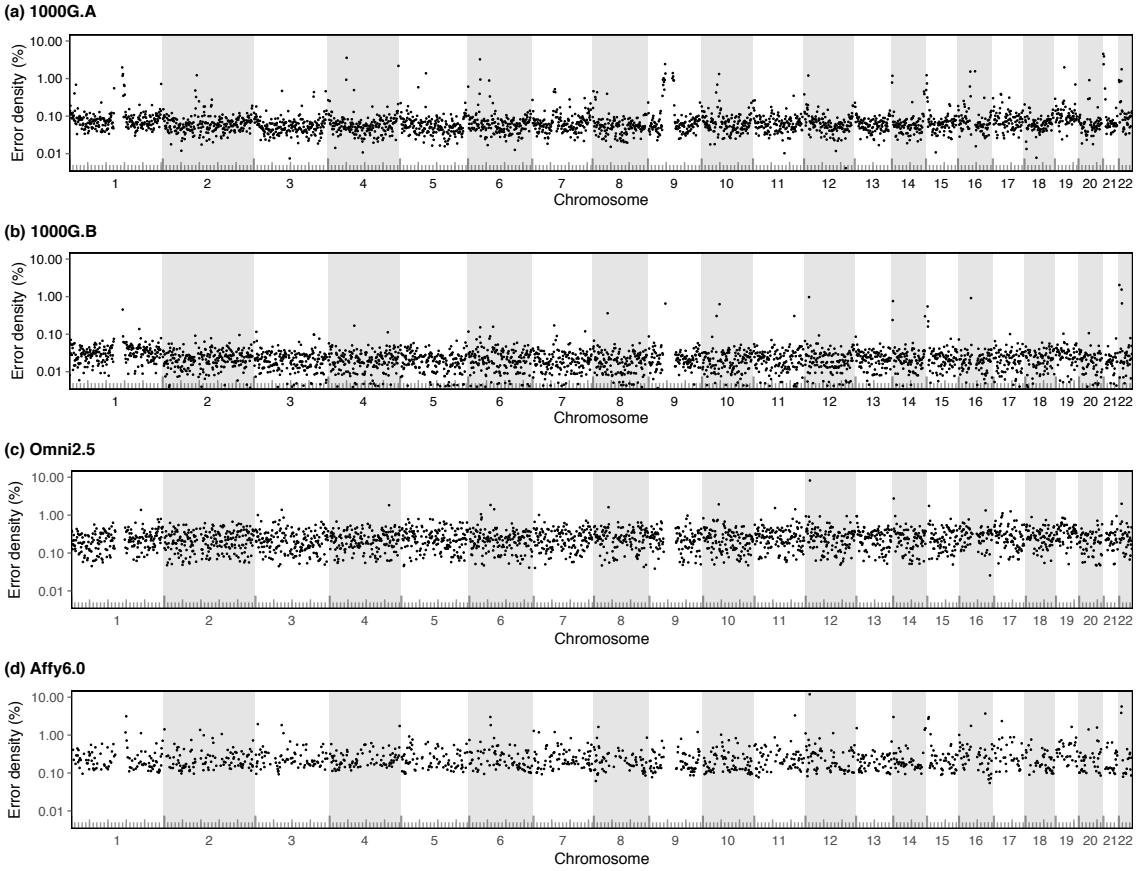


Figure 4.4: Positional genotype error density in sequencing and genotyping datasets. The density of misclassified genotypes was calculated along the length of each chromosome, which were divided into equally sized chunks of 1 Mb size. Error was calculated as the number of misclassified genotypes divided by the total number of genotypes per chunk; percent error shown on log scale. Chunks with less than 100 genotypes were removed. The ruler at the bottom edge of each panel shows physical distance per chromosome, where tick marks sit 10 Mb apart.

4.2.3.2 Empirical estimation of genotype error rate

Estimates for error rate parameters in $\mathcal{E}$ were derived by considering the proportional relation among observed types per true genotype class. For each true genotype class j, the number of genotypes observed in class i was divided by the total number in class j, which gives the empirical value of parameter ε_{ij}; denoted by $\tilde{\varepsilon}_{ij}$. For an exact formulation, let n_{ij} be the number of observed $\tilde{g}_i$ genotypes whose actual type belongs to the true genotype class j. The empirical value is calculated as

$$\tilde{\varepsilon}_{ij} = \frac{n_{ij}}{N_j} \quad \forall j \tag{4.4}$$

where $N_j = \sum_{i=0}^{2} n_{ij}$, *i.e.* the number of all genotypes per true class j, such that $\tilde{\varepsilon}_{0j} + \tilde{\varepsilon}_{1j} + \tilde{\varepsilon}_{2j} = 1 \; \forall \; j$. Results for each dataset are presented in Table 4.2 (next page).

Table 4.2: Measured genotype penetrance in sequencing and genotyping data. Genotypes in each true genotype class (g_0, g_1, and g_2) were distinguished by observed genotype class ($\tilde{g}_0$, $\tilde{g}_1$, and $\tilde{g}_2$), to obtain empirical expectations for genotype penetrances ε_{ij}. Per dataset, proportions sum to 100% by column. The total number of genotypes counted per true class are given in the table.

Dataset	Observed genotype	True genotype		
		g_0	g_1	g_2
1000G.A	$\tilde{g}_0$	99.942%	0.550%	0.033%
	$\tilde{g}_1$	0.041%	99.281%	0.228%
	$\tilde{g}_2$	0.017%	0.169%	99.739%
	Total	73,435,064	2,076,115	1,347,647
1000G.B	$\tilde{g}_0$	99.982%	0.193%	0.003%
	$\tilde{g}_1$	0.013%	99.749%	0.077%
	$\tilde{g}_2$	0.005%	0.057%	99.920%
	Total	56,739,327	1,515,508	978,728
Omni2.5	$\tilde{g}_0$	99.655%	0.048%	0.009%
	$\tilde{g}_1$	0.195%	99.909%	0.021%
	$\tilde{g}_2$	0.149%	0.043%	99.970%
	Total	3,087,037	854,327	522,876
Affy6.0	$\tilde{g}_0$	99.831%	0.081%	0.004%
	$\tilde{g}_1$	0.093%	99.849%	0.040%
	$\tilde{g}_2$	0.075%	0.070%	99.956%
	Total	931,857	463,649	337,649

Note that true genotypes homozygous for the reference allele, g_0, were not present in IPG and assumed from high-confidence regions if present in a given test dataset.

In all four datasets, values for $\tilde{\varepsilon}_{ij}$ were highest when genotypes were classified correctly; *i.e.* $i = j$, the main diagonal in $\mathcal{E}$. Notably, $\tilde{\varepsilon}_{00}$ was highest in all sequencing datasets, whereas $\tilde{\varepsilon}_{22}$ was highest in genotyping datasets. In each dataset, true homozygous genotypes were more likely to be misclassified as heterozygotes than as the opposite homozygote, but the probability to observe opposite homozygotes was non-zero throughout. Misclassification of true heterozygous genotypes showed a preference towards genotypes that are homozygous for the reference allele; except in *Affy6.0* where misclassification rates were nearly equal for $\tilde{g}_0$ and $\tilde{g}_2$. As formulated in Equation (4.3) on page 111, the observed genotype frequency is a function of the true allele frequency and error rates of genotypes. Hence, the frequency-dependent distribution of empirical error rate was assessed; see section below.

4.2.3.3 Frequency-dependent genotype error distribution

For each true genotype class, sites were pooled by their assigned population allele frequencies into 200 frequency bins of equal scope on linear scale; *i.e.* bins were separated in steps of 0.5%. Using Equation (4.4) on page 120, the proportions of observed genotypes were calculated in each bin, to obtain error rate expectations across the frequency spectrum. Additionally, because it can be expected that N_j becomes small at lower genotype frequencies, bins where the number of genotypes dropped below a nominal threshold were marked to indicate less support for estimated error rates. Three nominal levels of support were distinguished;

$$\begin{aligned} \textit{Low support} \quad & \text{if} \quad N_j < 100 \,, \\ \textit{Reduced support} \quad & \text{if} \quad 100 \leq N_j < 1000 \,, \\ \textit{High support} \quad & \text{if} \quad N_j \geq 1000 \,. \end{aligned}$$

For each dataset, the resulting error rate distributions are shown in Figure 4.5 (next page). The proportion of g_0 observed as $\tilde{g}_1$ was low throughout. This effect was seen in all four datasets, regardless of level of support, which was low for bins above 80% frequency in all datasets (except *1000G.A* with reduced support). The most striking observation is the loss of accuracy for true g_0 genotypes at higher allele frequencies. For example in *1000G.A*, the proportion of g_0 genotypes that were misclassified as $\tilde{g}_2$ increased substantially towards 100% alternate allele frequency.

However, this pattern should be interpreted with caution, due to the imperfect matching process between data generated on different platforms and, in particular, because the set of true homozygous reference genotypes had to be assumed from high-confidence regions in IPG data. For example, it is expected that g_0 genotypes are rarely observed at higher allele frequencies in a sample. Given that several hundred g_0 genotypes were seen at relatively high population frequencies makes it likely that a large proportion of g_2 genotypes were falsely assumed as g_0; *e.g.* due to missed or filtered variant calls.

Another explanation for this observation may be seen in somatic mutations in the sampled biological material. Data for both `NA12877` and `NA12878` were generated from lymphoblastoid cell lines created from sampled B-Lymphocyte cells. For example, it has been shown that induced pluripotent stem cells may accumulate genetic modifications (Gore *et al.*, 2011). Although CEPH cell lines are often used as a renewable resource of DNA, the possibility that cell lines undergo further genetic modifications may not be

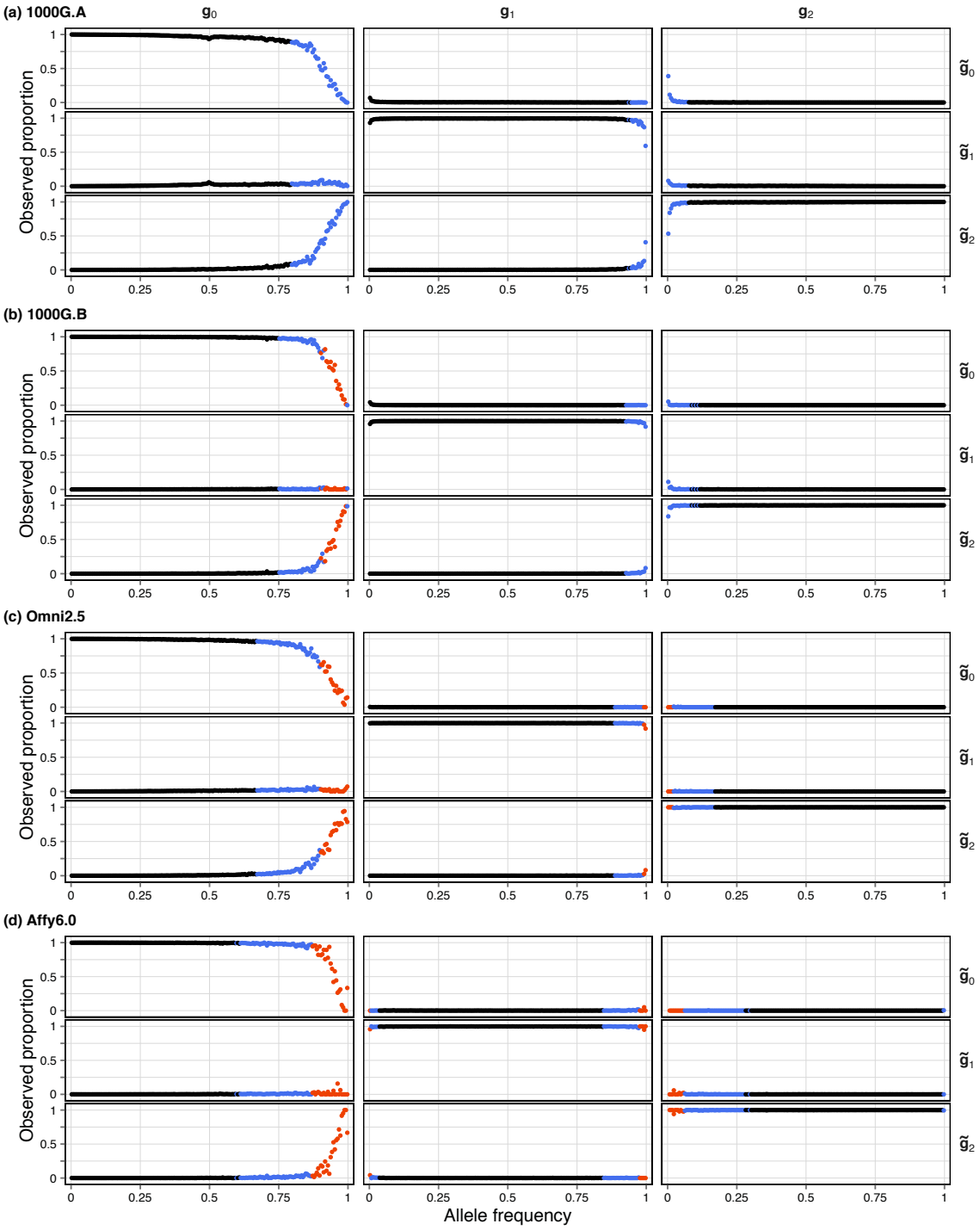


Figure 4.5: Empirical error rates in sequencing and genotyping data. For each true genotype class (columns), the fraction of g_j observed as $\tilde{g}_i$ (rows) was calculated per allele frequency bin, to estimate the frequency-dependent distribution of genotype penetrance ε_{ij}. The set of matched genotypes per true genotype class was divided into 200 bins along the allele frequency spectrum. Allele frequency was assigned to each matched site in a given test dataset, taking the population frequency as recorded in the full sample of the 1000 Genomes Project phase III dataset (2,504 individuals). Colours indicate the number of genotypes per bin, N_j, distinguished at nominal thresholds $N_j < 100$ (*red*), $100 \leq N_j < 1000$ (*blue*), and $N_j \geq 1000$ (*black*). Note that true genotypes homozygous for the reference allele, g_0, were not present in IPG and assumed from high-confidence regions if present in a given test dataset.

excluded. However, here, because the IPG protocol would have excluded sites that showed cell line artefacts, it is assumed that the genotypes had to be consistent with Mendelian laws. Regardless, note that salient patterns of genotype error were most apparent for the assumed subset of the data; *i.e.* at sites not actually contained in the set of reported genotypes. It is therefore possible that not all unobserved homozygous reference genotypes can be assumed from high-confidence regions when sites are only observed in other data.

In the opposite homozygote class, g_2, observed distributions were mirrored, such that the loss of accuracy occurred at lower frequencies; yet, the proportion of misclassified genotypes was markedly lower. Under the allele-based error model, this asymmetry suggests that the probability of the alternate allele to appear as the reference allele was higher than in reverse direction, such that $\epsilon_0 < \epsilon_1$; see Table 4.1 (page 112).

The estimated error distributions were used to reproduce empirical error rates in simulated data. This was done to assess the effect of genotype misclassification on IBD detection, using the method proposed in Chapter 3; *i.e.* targeted IBD detection done thoroughly, or `tidy`. For comparison, an alternate IBD detection method was applied to the same data (`Refined IBD` in `Beagle 4.1`; Browning and Browning 2013).

4.3 Impact of genotype error on IBD detection

One of the genotype error profiles constructed in the previous section was used to induce realistic error patterns in simulated data. Among the four test datasets, both sequencing datasets were recorded with higher levels of support. Although *1000G.B* showed overall lower levels of genotype error, *1000G.A* can be seen as being more representative for data obtained in recent large-scale studies; hence, the integration of error was conducted according to the frequency-dependent error rate distribution in the *1000G.A* profile. The process of error integration in simulated data is described below.

4.3.1 Integration of empirical error distributions in simulated data

The dataset simulated in Chapter 3 was re-used for integration of genotype error, so as to enable a direct comparison to previously obtained results after applying the same methodology for IBD detection; see Section 3.4.1 on page 89 for a description of the simulation process. Briefly, data were simulated using `msprime 0.4.0` (Kelleher *et al.*, 2016), with a sample size of N = 2,500 individuals (*i.e.* 5,000 haplotypes), resulting in 0.673 million variant sites over a length of 62.949 Mb (108.267 cM). Diploid individuals wee formed by pairing haplotypes. From those, data were converted into genotypes,

which were then phased, such that three datasets were generated (true haplotypes, phased haplotypes, and genotype data). The same process was followed here; however, genotype error was evoked on haplotype level before haplotype sequences were combined to form genotypes. By doing so, identically distributed proportions of error were present in both the haplotype and genotype datasets, after conversion of the former into the latter, as well as subsequent phasing.

Haplotypes were randomly assigned into fixed pairs which would later form the genotypes of individuals. Error was included by randomly replacing haplotype pairs dependent on the empirically determined misclassification rates per true genotype class j. This was done by selecting each variant site in turn and indexing each haplotype pair that would form genotype g_j before error. The index ensured that pairs would be drawn without replacement. Then, for each class j, indexed pairs were randomly drawn and assigned to each of the three observed genotype classes, in proportions equal to empirical error rates, as determined for the given allele frequency of the currently selected site. Haplotype pairs were "mutated" according to their assigned class, such that they would form $\tilde{g}_i$ after error.

These haplotype data were then converted into a corresponding genotype dataset, which was then phased using `SHAPEIT2` (Delaneau *et al.*, 2008, 2013); see description in Section 3.4.2 (page 91). The three resulting datasets resembled the original datasets used in the evaluation of IBD inference presented in Chapter 3 (Section 3.5 on page 93), which therefore facilitated assessment in relation to the simulated genealogy and the underlying IBD structure of the sample, as well as a direct comparison to the results generated before error was included.

Recall that, for example, the empirically determined proportions of the true g_0 class per observed $\tilde{g}_2$ were likely to be inflated at higher allele frequencies; as discussed in Section 4.2.3.3 (page 122). Errors reproduced in the simulated dataset may therefore be higher than actually present in the 1000G dataset. A more detailed account of the characteristics and consequences of the integration of error in the simulated dataset is given in Section 4.3.2 (next page).

4.3.1.1 Accuracy analysis

The following briefly describes the analyses performed. Two IBD detection methods were applied to available data; the `tidy` method as proposed in Chapter 3 and the `Refined IBD` algorithm in `Beagle 4.1` (Browning and Browning, 2013). Recall that the `tidy` method is

based on inference of recombination events in pairs of individuals to detect breakpoints distal to a given target site, which is enabled by the four-gamete test (FGT), which requires haplotype data, and the discordant genotype test (DGT), which requires genotype data; see Section 3.3 (page 81). Accuracy was measured in terms of the physical distance between breakpoints and focal target position at which IBD segments were identified; calculated using the squared Pearson correlation coefficient, r^2, and the root mean squared logarithmic error (RMSLE) as defined in Equation (3.1) (page 92). Data were analysed in three approaches: (a) the FGT on the simulated haplotypes and (b) on phased haplotypes, and (c) the DGT on genotype data. Note that the `RefinedIBD` algorithm can only be used with haplotype data and was therefore evaluated in (a) and (b). Again, f_k was used to denote the frequency of shared alleles, where k is the allele count in the sample.

4.3.2 Results

In presence of genotype error, the misclassification of alleles observed to be shared among individuals may pose a problem to the identification of haplotype sharing by descent, in particular for variants that are low in frequency or rare, see Figure 4.6 (next page); recall that the `tidy` method utilises rare allele sharing to identify regions of recent relatedness. Figure 4.6a indicates the rate at which genotype data appear at a frequency different to their true frequency due to genotype error; shown for variants below 5% allele frequency. The figure shows the change between true and observed allele count, depicted as the difference of the observed minus the true count. For example, 68.226 % of f_2 variants remained at the same frequency, but this fraction decreased for alleles found at higher frequencies, *e.g.* 51.140 % for f_{10} and 28.771 % for f_{50} variants.

For IBD detection using rare variants as target sites, this may not pose a problem if identified individuals indeed share a given allele. This is further explored in Figure 4.6b, where the false positive rate (FPR) indicates the proportion of alleles that were falsely identified due to g_0 or g_2 genotypes being observed as $\tilde{g}_1$. Conversely, the false negative rate (FNR) indicates the proportion of shared alleles that were missed due to g_1 being observed as $\tilde{g}_0$ or $\tilde{g}_2$. The risk for both types of error was greatest for f_2 variants, here observed at FPR = 0.094 and FNR = 0.127. On average, FNR (0.009; $\pm 0.665 \times 10^{-3}$ SE) was higher than FPR (0.007 ; $\pm 0.404 \times 10^{-3}$ SE), indicating that more shared alleles were missed than falsely observed.

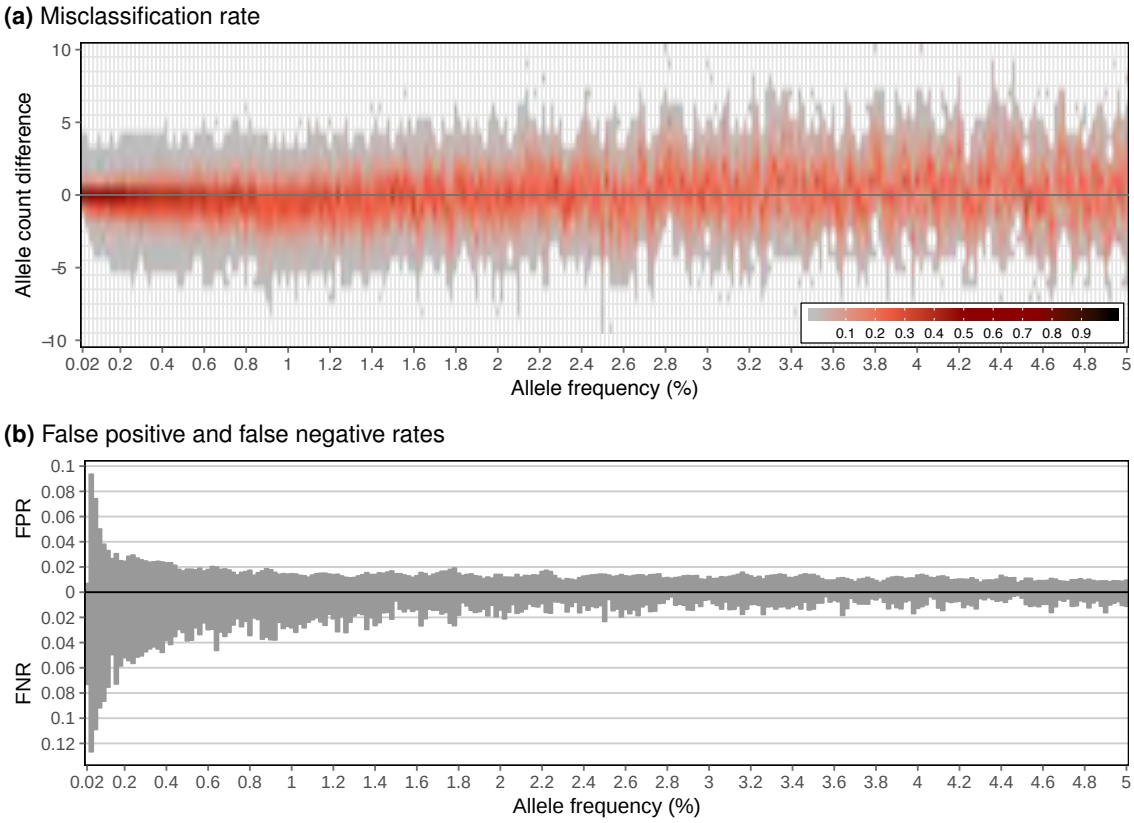


Figure 4.6: Misclassification of target sites in presence of genotype error. Simulated data were modified such that realistic distributions of genotype error were induced. Panel **(a)** indicates the rate at which alleles were observed at different frequencies after the inclusion of error. The proportion of misclassification is indicated by colour intensity. Panel **(b)** distinguishes alleles that were falsely observed (false positive) as well as alleles that were missed after the inclusion of error (false negatives).

4.3.2.1 IBD detection using *tidy*

The set of target sites included all f_k variants found at $k \in \{2, \ldots, 25\}$ (*i.e.* alleles shared at frequency $\leq 0.5\%$). In total, 0.297 million SNPs were available in this frequency range, which represented 0.936 % of the targets previously identified before the inclusion of error. Note that sites were only considered if matched to the set of true IBD segments (as previously determined from simulation records). Hence, false positives were not considered in this analysis. The number of pairs sharing the focal alleles at available target sites was 10.362 million; *i.e.* the total number of IBD segments detected in each approach.

Duplicate segments were removed to retain unique segments after sorting segments by f_k, such that segments were associated with the presumably youngest shared allele within the detected breakpoint interval. Recall that the same IBD interval may be inferred from multiple focal alleles, as these are assumed to sit on the same shared haplotype. The proportion of uniquely identified segments was 48.035 % in Approach (a), 48.554 % in

Approach (b), and 41.094 % in Approach (c), whereas 27.403 % were unique in the set of true IBD segments. These sets were then intersected to measure accuracy on the same set of unique IBD segments, which resulted in 2.824 million (27.256 %) per approach.

The proportion of breakpoints that were overestimated (in terms of the true distance between target position and actual recombination breakpoint) was 50.684 %, 49.691 %, and 63.864 % in (a), (b), and (c), respectively. Recall that before error was included, the vast majority of breakpoints (> 95%) was overestimated in each approach. When the FGT was used, 49.221 % of breakpoints were underestimated and 0.095 % coincided with true breakpoints in Approach (a), which was similar in Approach (b) where 50.217 % and 0.092 % of breakpoints were underestimated and exact, respectively. When the DGT was used, 36.074 % of breakpoints were underestimated, but also only 0.061 % were exact.

Overall accuracy was low in all approaches; r^2 was 0.069, 0.072, and 0.089 in (a), (b), and (c), respectively, which was also reflected in corresponding high error scores (RMSLE), which were 0.714, 0.722, 0.694, respectively. For comparison, accuracy measured on the same set of segments, but without genotype error, was $r^2 > 0.85$ and RMSLE < 0.55 for each approach. When seen per f_k category, all three approaches consistently showed low correlation with true segment breakpoints ($r^2 < 0.2$) and a high magnitude of error (RMSLE > 0.6); see Table 4.4 on page 153, which is shown for comparison to results obtained using the HMM-based approach developed in the second part of this chapter.

To determine the lengths of IBD segments in each approach, boundary cases were removed to ensure that breakpoints were detected on both sides of each segment; 0.622 %, 0.621 %, and 0.924 % were removed in (a), (b), and (c), respectively, but which was noticeably lower compared to boundary cases removed in the set of true IBD segments (1.359 %). Again, sets were intersected, retaining 2.782 million (98.490 %) common segments across approaches.

Median physical length (and median genetic length) was relatively short when the FGT was used in Approaches (a) and (b), yielding 0.200 Mb (0.381 cM) and 0.198 Mb (0.378 cM), respectively. For the DGT, Approach (c), median length was closer to the true length; 0.332 Mb (0.635 cM) and 0.337 Mb (0.585 cM), respectively. However, for f_{25} variants, a clear difference was seen, where the median length was 0.311 Mb (0.527 cM) in (a) and 0.270 Mb (0.430 cM) in (b). But median length was likewise reduced in (c) compared to the true length; 0.543 Mb (0.926 cM) and 2.172 Mb (3.677 cM) respectively. This difference was not seen towards higher frequencies, *e.g.* at f_2 variants, reaching 0.171 Mb (0.354 cM), 0.173 Mb (0.354 cM), and 0.289 Mb (0.578 cM) in (a), (b), and (c), respectively, compared to 0.228 Mb (0.408 cM) in true segments.

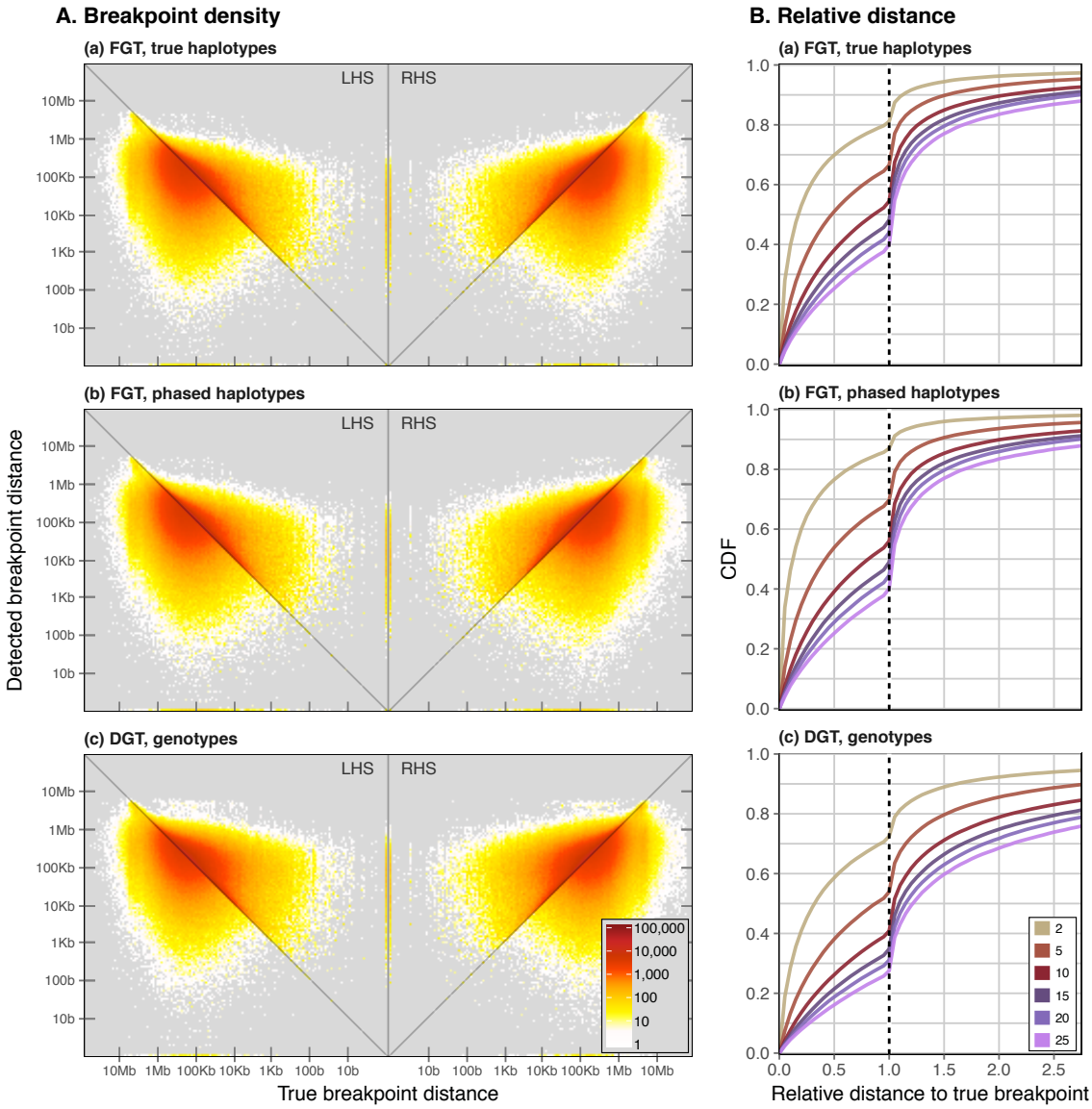


Figure 4.7: Accuracy of IBD detection using *tidy* after inclusion of genotype error. Simulated data after the inclusion of error was analysed using the FGT on true haplotypes (a), phased haplotypes (b), and the DGT on genotype data (c). Panel **(A)** shows the density of true and detected breakpoints in terms of the physical distance between each detected breakpoint and the corresponding focal site; shown separately for breakpoints detected on the left (LHS) and right-hand side (RHS) of a focal position. The number of detected and true breakpoints is indicated by colour intensity. Panel **(B)** shows the physical length in terms of the relative distance between a focal site and the detected breakpoint, $\hat{d}$, normalised by the distance to the true breakpoint, d; *i.e.* relative distance was calculated as $\hat{d}/d$, such that < 1 indicates underestimation and and > 1 overestimation of detected breakpoint distance. This is shown as the cumulative density per f_k variant, for $k \in \{2, 5, 10, 15, 20, 25\}$.

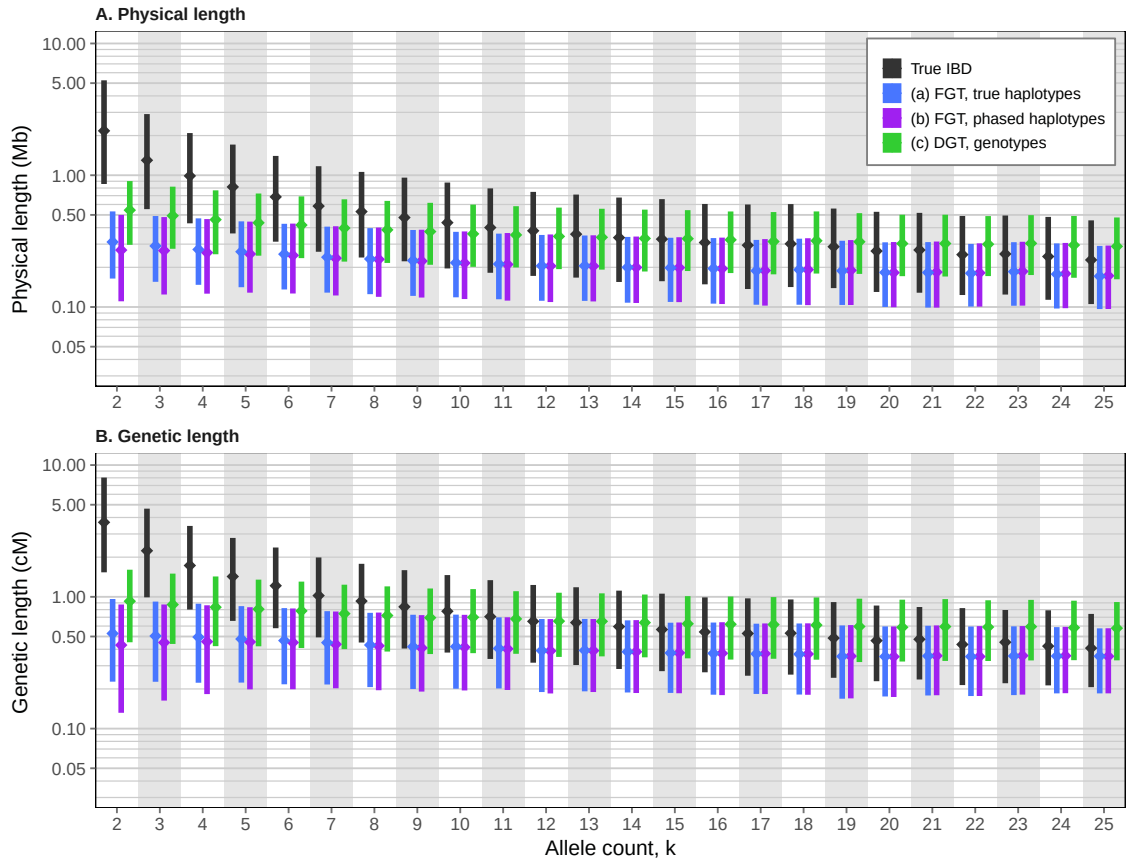


Figure 4.8: IBD length distribution using *tidy* after inclusion of genotype error. The distribution of physical (A) and genetic (B) segment length is shown by allele count (f_k category). Results are shown for three approaches; (a) FGT on true haplotypes, (b) FGT on phased haplotypes, and (c) DGT on genotype data. Corresponding true lengths are shown in for comparison. Bottom and top of each bar indicate 1st and 3rd quartiles, respectively, between which the median (2nd quartile) is marked (*diamonds*).

The length distribution of true and detected IBD segments is shown in Figure 4.8 (this page). Note the similarity to Figure 3.14 on page 105, which was conducted using the FGT and DGT on data from 1000G (chromosome 20). This result suggests that the non-probabilistic IBD detection method implemented in `tidy` is likely to be biased in presence of genotype error. To illustrate the problem, consider the example given in Figure 4.9 (next page), which highlights the effect of error for breakpoint detection using the FGT and DGT. The figure shows the underlying IBD structure for each pair of chromosomes in two individuals sharing a randomly picked rare allele. In Figure 4.9a, the positions of breakpoints detected using the FGT and DGT are indicated as found along the whole chromosome before the inclusion of error. In contrast, Figure 4.9b shows the same analysis but after genotype error was included. Since the innermost breakpoint interval delimits the inferred IBD segment, it can be expected that even small amounts of genotype error are likely to result in underestimation of IBD length.

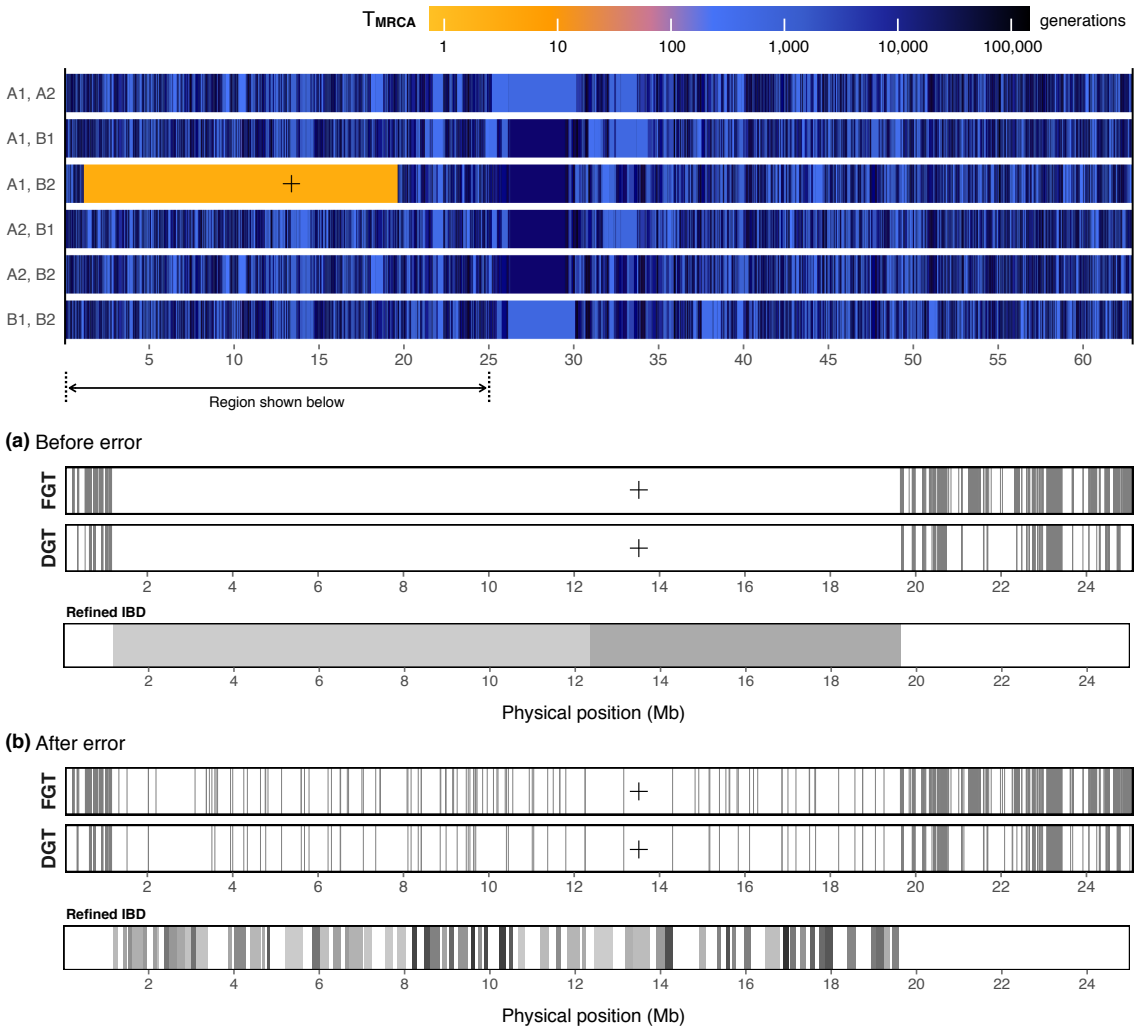


Figure 4.9: Example of the effect of genotype error on IBD detection. One allele and a pair of individuals sharing that allele were randomly selected and the underlying IBD structure for all six possible pairs of the four chromosomes was determined from simulation records (*top*). The figure shows the "mosaic" of IBD segments along the sequence of the simulated chromosome; distinguished by the time to the most recent common ancestor (T_{MRCA}). The focal shared allele is indicated at the pair of chromosomes sharing that allele (*cross*). Data were compared before **(a)** and after **(b)** the integration of empirically determined genotype error. In each dataset, the FGT and DGT were used to detect all breakpoints to the left and right-hand side of the target position. In addition, the IBD segments reported for the focal pair of haplotypes using the `RefinedIBD` method are shown before and after error, where each segment is distinguished by a different colour on grey-scale. Note that these results were produced on true haplotype data but not phased haplotypes, to highlight the impact of genotype error alone. Data were simulated using `msprime` (see Section 3.4.1 on page 89).

4.3.2.2 IBD detection using *Refined IBD* in *Beagle 4.1*

The probabilistic `Refined IBD` method implemented in `Beagle` version 4.1 (Browning and Browning, 2013) was used for IBD detection in data after the inclusion of error.* Because the method requires haplotype data, the analysis was performed on the generated ("true") haplotypes and the set of phased haplotypes; in the following referred to as Approaches (a) and (b). The purpose of this analysis was to determine the impact of genotype error on the detection of contiguous shared haplotype intervals. Because `Refined IBD` attempts to infer IBD for any haplotype pair without reference to a specific target allele, it was necessary to match the set of reported IBD segments to the true shared haplotype intervals that were previously determined from simulation records (*i.e.* all segments around shared $f_{[2,25]}$ alleles). The analysis followed the procedure described in Section 3.5.0.1 (page 98).

Approach (a) returned 12.195 million segments at 5.807 million haplotype pairs. In Approach (b), 12.398 million segments at 5.938 million haplotype pairs were detected. In the latter, 1,382 pairs were removed from the results obtained on phased haplotype data, to enable the analysis to match segments based on the pair of individuals for which IBD was inferred; otherwise the true haplotype pair could not be identified correctly due to the phasing process (as described in Section 3.5.0.1).

The total base overlap between inferred and true shared haplotype intervals was measured, for which all segments inferred in (a) and (b) were aligned to the set of true intervals, after removing duplicate segments in the latter. The proportion of overlap was measured only at segments at which at least one base overlapped. On average, an overlap of 98.9 % and 98.6 % was measured relative to inferred IBD in (a) and (b), respectively, but 18.7 % and 19.0 % on average when measured relative to true IBD, respectively. This suggested that the inferred IBD intervals were likely to be found within the region spanned by the underlying shared haplotype, but such that the region was only partially covered on average. For comparison, before error, average overlap relative to true IBD segments was 44.3 % and 42.3 % in (a) and (b), respectively. The density of overlap measured relative to both the inferred and true segments, before and after error, is shown in Figure 4.10 (next page). Also, the example shown in Figure 4.9 (page 131) highlights the difference in coverage for IBD segments inferred using `Refined IBD` on data before and after the integration of error.

Next, inferred IBD segments returned in Approaches (a) and (b) were matched to true intervals based on finding a given target site within the inferred interval. This was done to facilitate measures of accuracy based on the physical distance between a given

* Beagle 4.1: https://faculty.washington.edu/browning/beagle/beagle.html [Date accessed: 2016-11-22]

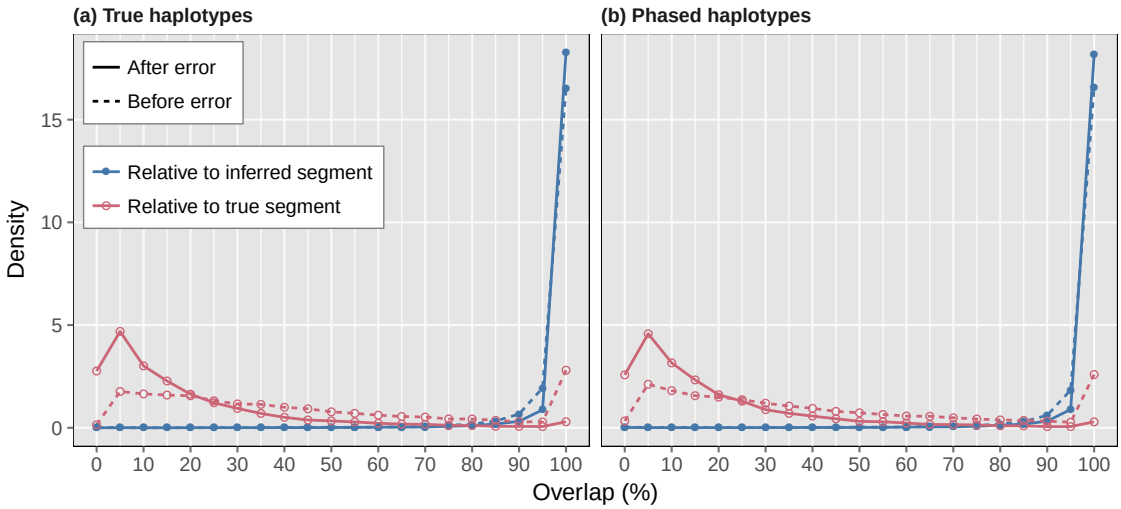


Figure 4.10: IBD segment overlap inferred using *Refined IBD* after integration of error. The proportion of overlap was measured by aligning each inferred IBD segment to the set of unique true segments determined for a given pair. Interval comparisons with zero overlap were ignored. The results shown were generated on a random subset of 10,000 pairs in Approaches (a) and (b). The reported densities refer to the proportion of overlap with respect to the inferred segment (*blue*) and the true segment (*red*). Corresponding densities for the results obtained on data before the integration of error are shown for comparison (*dashed* lines); see Figure 3.11 (page 100).

target site and the position of an inferred breakpoint. To enable the analysis to measure breakpoint accuracy conditional on the frequency of the target allele, the full set of matched segments was further reduced by removing duplicate segment matches per pair, where the segment tagged by the lowest-frequency allele was retained; again, as described in Section 3.5.0.1. As a result, 1.505 million segments and 1.516 million segments were retained, respectively, after removing duplicate segment matches per pair.

In Approach (a), 80.103 % of the breakpoints detected were underestimated relative to the matched target position. This was similar in Approach (b), where 80.2 % were underestimated. A difference between low and high frequency alleles was seen; 93.7 % and 95.2 % of segments matched to f_2 alleles were underestimated, respectively, while 76.0 % and 75.8 % were underestimated at f_{25} alleles, respectively. The density of inferred by true breakpoint distance per match segment, as well as the CDF of the relative distance per f_k category is shown in Figure 4.11 (next page).

The overall accuracy of the breakpoints detected was relatively low; $r^2 = 0.020$ (RMSLE = 0.869) in (a) and $r^2 = 0.021$ (RMSLE = 0.864) in (b), measured by the physical distance between a given target site and the breakpoint on either the left or right-hand side. Again, accuracy decreased towards lower frequencies; for example, at f_2 alleles, $r^2 = 0.006$ (RMSLE = 1.477) and $r^2 = 0.005$ (RMSLE = 1.494), respectively, compared to $r^2 = 0.031$ (RMSLE = 0.753) and $r^2 = 0.033$ (RMSLE = 0.747) at f_{25} alleles, respectively.

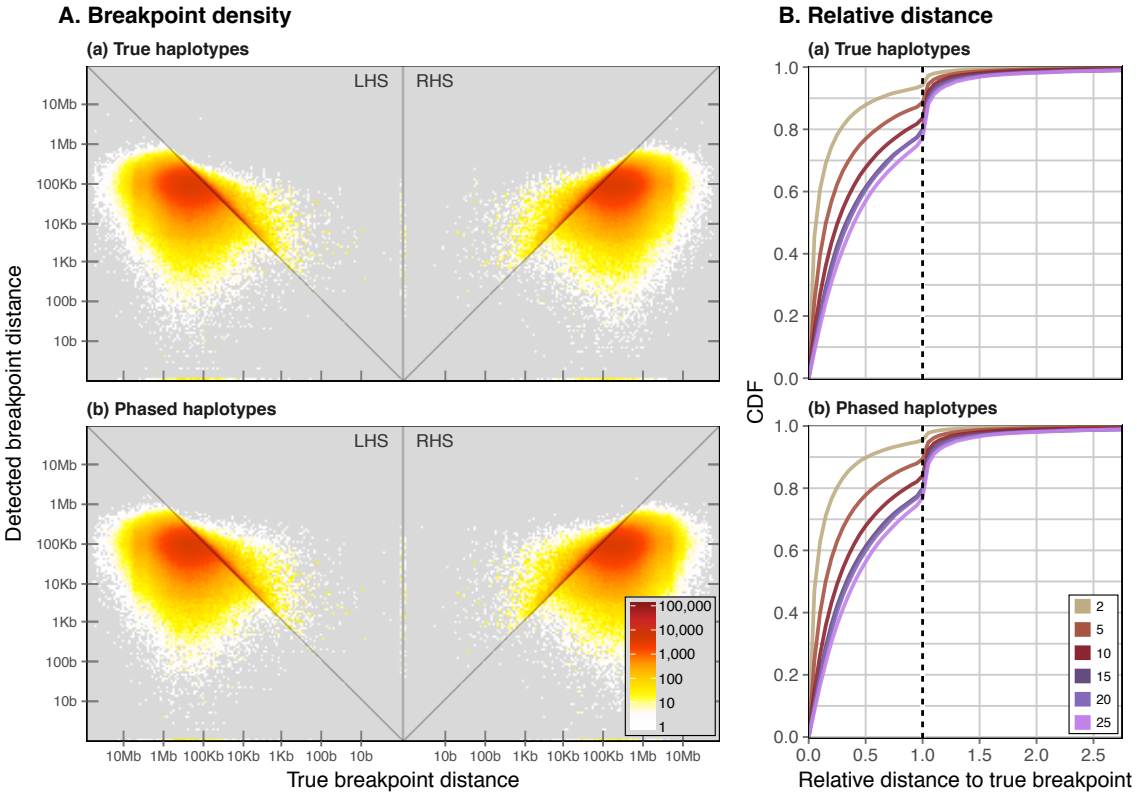


Figure 4.11: Accuracy of IBD detection using *Refined IBD* after integration of error. Panel **(A)** shows the density of true and detected breakpoint distance to the matched target position. Panel **(B)** shows the cumulative distribution function (CDF) of the relative distance by f_k category where the true distance is mapped to a relative distance of 1 (*dashed* line).

The physical (genetic) lengths of inferred segments were measured after removing boundary cases; 26,856 in (a) and 26,939 in (b). Overall median length was 0.129 Mb (0.242 cM) in (a) and 0.131 Mb (0.244 cM) in (b), which was considerably shorter compared to the set of matched true shared haplotype segments at 0.519 Mb (0.878 cM). The median length of segments matched to f_2 alleles was marginally longer compared to higher-frequency alleles; 0.155 Mb (0.272 cM) in (a) and 0.158 Mb (0.271 cM) in (b), where the matched true segments were found at 2.612 Mb (4.475 cM). For f_{25} alleles, median length was 0.123 Mb (0.229 cM) and 0.125 Mb (0.231 cM), respectively, compared to 0.397 Mb (0.637 cM) for the matched true segments. The distribution of segment length inferred in Approaches (a) and (b) is shown in Figure 4.12 (next page).

4.3.3 Discussion

Two conclusions can be drawn from the analysis of simulated data after the integration of (realistic) error rates. First, the distribution of shared alleles is altered in presence of error such that a given shared allele may not correctly identify genomic regions of recent

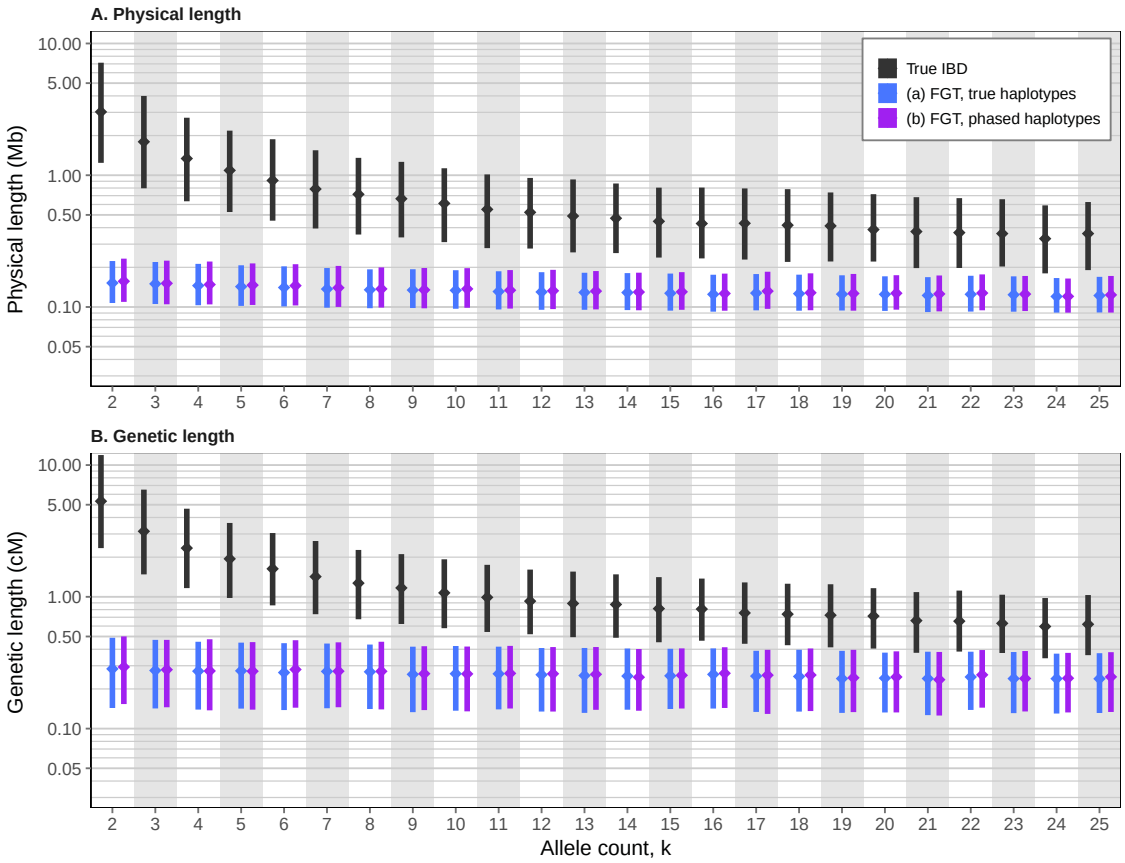


Figure 4.12: IBD length detected using *Refined IBD* after integration of error. The distribution of physical (A) and genetic (B) segment length is shown by allele count (f_k category). Results were obtained using `Refined IBD` in `Beagle 4.1`, on true and phased haplotype data; *i.e.* Approaches (a) and (b) (page 126 and page 126), respectively. Bottom and top of each bar indicate 1st and 3rd quartiles, respectively, between which the median (2nd quartile) is marked (*diamonds*).

shared ancestry between pairs of haplotypes. In cased where a rare allele was missed (or removed through quality control), the underlying shared haplotype segment may still be retrieved from other, nearby rare variants that identify the same shared haplotype in a given haplotype pair. Conversely, in cases where a rare allele was falsely called or typed, it can be expected that the actual relationship between the haplotypes at that position is relatively old, such that the interval detected is likely to be relatively short.

Second, the main insight gained from the analyses using the FGT, DGT, and `Refined IBD` is that the impact of error on the detection of shared haplotype segments is dramatic. As highlighted by the example shown in Figure 4.9 (page 131), neither method was able to infer contiguous intervals that accurately reflect the actual shared haplotype segment. When the FGT or DGT were used, it was suggested that even low rates of error may lead to the observation of false positive breakpoints, because even one false positive would suffice to disrupt the rule-based detection process as it is currently implemented.

However, when using `RefinedIBD`, the low overall accuracy measured may not be surprising given that the inferred segments were likely to be scattered along the full length of the underlying shared haplotype region. The matching process was therefore not straightforward, as only the segment covering a given target site was retained. For example, an additional method could be implemented to concatenate neighbouring segments inferred per pair to approximate the interval of the underlying shared haplotype. This was not attempted here, as it appeared more feasible to extend the targeted IBD detection approach implemented in `tidy`, which was done in the following section.

4.4 A Hidden Markov Model for IBD inference

Despite the high accuracy of the FGT and DGT to detect shared haplotype segments in simulated data, it has emerged from the previous analysis that a non-probabilistic approach may be less suitable for IBD detection if the presence of genotype error cannot be excluded. Because it cannot be assumed that real data is obtained without error, it would therefore be beneficial to devise a fully probabilistic implementation of the IBD detection algorithm, in which observed error rates can be included. Here, this was attempted by constructing a Hidden Markov Model (HMM).

An HMM is a probabilistic sequence model which is widely used in applications of machine learning, likelihood computation, and sequence classification; see Rabiner (1989). In general, a sequence of observations is assumed to be the product of an unobserved Markov process, in which a sequence of underlying, but "hidden" states determines the probability of observing the data. Each state is characterised by a probability distribution over a finite set of possible observations. Although the sequence of hidden states is not known, it can be inferred from the sequence of observations.

A wide range of statistical methods for genetic data analysis are driven by HMM-based algorithms. Notable examples are methods used for genotype phasing and imputation; *e.g.* `SHAPEIT` (Delaneau *et al.*, 2011), `EAGLE` (Loh *et al.*, 2016a,b), and `IMPUTE` (Howie *et al.*, 2009, 2011), to name a few. It is worth to mention that many of the commonly employed methods (above included) are based on the influential Li and Stephens (2003) model, which for a set of observed genotypes reconstructs the unobserved haplotypes as "imperfect mosaics" of known haplotypes in reference data. While this model provides the ability to solve several kinds of problems in statistical genetics, such as phasing or imputation, it is less applicable for inference of IBD.

A variety of different approaches exist for the inference of IBD segments, many of which have not fully adopted the view that observed genetic variation is the product of a genealogical process which, in principle, can be modelled as a Markov process. An example of a rule-based method is the widely implemented `GERMLINE` algorithm (Gusev *et al.*, 2009), which is part of the often employed `Refined IBD` method (Browning and Browning, 2013). This algorithm was designed as an efficient search method through which IBD status is inferred from imperfectly matched haplotypes in large sample data. In contrast, model-based implementations for inference of IBD in samples of seemingly unrelated individuals all rely on HMMs; see review by Thompson (2013). The first to assume that IBD arises from a Markov process (without specifically stating it) was Stam (1980), who extended the idea of recombination breakpoints (or "junctions") introduced by Fisher (1949, 1954) to describe the probability distribution of the fraction of the genome that is identical by descent in a finite and randomly mating population. Later, Leutenegger *et al.* (2003) developed an HMM for inference of inbreeding coefficients from genotype data in individuals of unknown parental relationships. Equivalent models were implemented to detect IBD in phased haplotypes (*e.g.* Purcell *et al.*, 2007; Browning, 2008).

Here, a different IBD-model is proposed which is used for inference of recombination breakpoints around a target position in pairs of individuals. The approach is conceptually similar to the previously presented method for deterministic IBD detection using the FGT or DGT, see Section 3.3 (page 81), but where the detection of breakpoint intervals (*i.e.* the physical start and end points of IBD segments) are determined through sequence classification in the HMM. Notably, the presented method relies on genotype information and does not require haplotype data; it is therefore not affected by phasing error.

The following section describes the algorithm through which target sites in sample data are analysed. This is followed by a detailed description of the model, which includes the theoretical expectations under the assumption of no error. Then, the model is extended to include the empirically determined distributions of genotype error for each of the possible genotype pairs. In the end, the presented HMM-based method for IBD detection was evaluated in the same way as was done for the FGT or DGT in the previous chapter.

4.4.1 The algorithm for probabilistic IBD inference

Consider a sample of N diploid individuals and M variant markers; in particular, SNP data are assumed. To determine the IBD structure around a focal variant site, let this site be denoted by $i \in \{1, \ldots, M\}$ and its physical position by b_i. All individuals sharing

the derived (alternate) allele at this site are identified and analysed in a pairwise fashion. In each pair, the breakpoint interval, $[b_L, b_R]$, is inferred, where b_L and b_R are the chromosomal positions of the most likely recombination breakpoints to the left and right-hand side of the focal position, respectively.

As before, it is convenient to refer to a target site by its frequency in the sample. Thus, f_k variants are distinguished where k is the number of allele copies in the sample, and where $k \geq 2$ must be satisfied. Note that only those individuals are considered that are heterozygous for the focal allele, which is why the subset of identified individuals may be smaller than k, but not smaller than 2 in order to form at least one pair. Also, as described in Chapter 3 (Section 3.2, page 78), rare variants are presumed to derive from relatively recent mutations and are therefore more likely to identify long IBD tracts, as recombination had less time to break down the length of the shared haplotype identity. Hence, this method is primarily intended for inference of IBD around rare variants, where $k \ll 2N$. However, note that in principle any $\mathrm{f}_{\geq 2}$ variant can be analysed using the presented method.

The input data analysed in the HMM is the paired sequence of genotypes in both individuals sharing the focal allele. The observation sequence is composed of the paired genotypes along the chromosomes of the two individuals sharing the focal allele; as such, haplotype data is not required. Since each individual contributes a genotype at a single locus, g_k (where $k \in \{0, 1, 2\}$), to form a genotype pair, denoted by $g_{k_1 k_2}$, it follows that there are six possible observation states; g_{00}, g_{01}, g_{02}, g_{11}, g_{12}, and g_{22}, where the order of genotypes in a pair is ignored. Further, two states are distinguished in which genotype pairs can be observed; either the two individuals share a haplotype identical by descent, or they do not, which is denoted by *ibd* and *non*, respectively. These correspond to the hidden states that are assumed to generate the data.

For a given focal site and a pair of individuals sharing the allele, the sequence of genotype pairs is analysed as two independent Markov chains; *i.e.* one to the left and one to the right-hand side of the focal variant, with the focal site at the start of both chains. For convenience, the index j is defined relative to i and follows the direction of moving from b_i to the last site in the observed sequence, either b_1 to the left or b_M to the right-hand site. Hence, $j = 0$ at the focal site and $j = m$ at the last site, where m is the number of markers to the left or right-handed sequence relative to the focal site (excluding the focal site).

Since the focal allele is assumed to identify the shared haplotype in *ibd*, the first site along the sequence that is classified in the *non* state is taken as a breakpoint, on both sides, such that the inferred IBD segment is enclosed in $[b_L, b_R]$. By definition, the

smallest detectable interval around a focal variant at site i is therefore $[b_{i-1}, b_{i+1}]$. If the chain remains in *ibd* until the end of the sequence, the last position is taken as a breakpoint (referred to as a *boundary case*).

The following section describes the underlying model through which each site in the observation sequence is classified into either *ibd* or *non*.

4.4.2 Description of the model

Identity by descent is modelled as a first-order Markov process in a two-state HMM, where the observed genotypes in a pair of diploid individuals are emitted from either the *ibd* or the *non* state. Given the Markov property, the following assumptions are made. First, the probability of the hidden state at site j only depends on the previous hidden state at site $j-1$. Second, the probability of observing a particular genotype pair at site j only depends on the hidden state at site j and not on any of the other states.

Let the hidden state space be denoted by $S = \{ibd, non\}$, and the set of observable states by $G = \{g_{00},\ g_{01},\ g_{02},\ g_{11},\ g_{12},\ g_{22}\}$. The model itself is denoted by

$$\lambda \ = \ \{\Psi, \xi, \pi\} \tag{4.5}$$

where Ψ is a matrix of state *transition* probabilities and ξ corresponds to a set of vectors which store the probability of observing each of the possible genotype pairs; *i.e.* the *emission* probabilities in each state. The model is illustrated in Figure 4.13 (next page), where the probabilities of emission from *ibd* are denoted by $\delta_{k_1 k_2}$ and from *non* by $\eta_{k_1 k_2}$. The *initial* probabilities of being in either state at the start of the sequence is given by π.

The parameters of the model are defined in two ways. First, theoretical expectations for transition, emission, and initial probabilities are derived; see this page, page 142, and page 144, respectively. Then, in Section 4.4.3 (page 144), the model is extended to include genotype error from empirical data as obtained in Section 4.2 (page 114).

4.4.2.1 Transition probabilities

Given the two hidden states, the transition matrix Ψ is defined as a 2×2 matrix which stores the probabilities of moving from one state into another state, as well as the probabilities of remaining in the same state; see below.

$$\Psi_{j,k} \ = \ \begin{bmatrix} \psi_{j,k}(ibd \mid ibd) & \psi_{j,k}(non \mid ibd) \\ \psi_{j,k}(ibd \mid non) & \psi_{j,k}(non \mid non) \end{bmatrix} \tag{4.6}$$

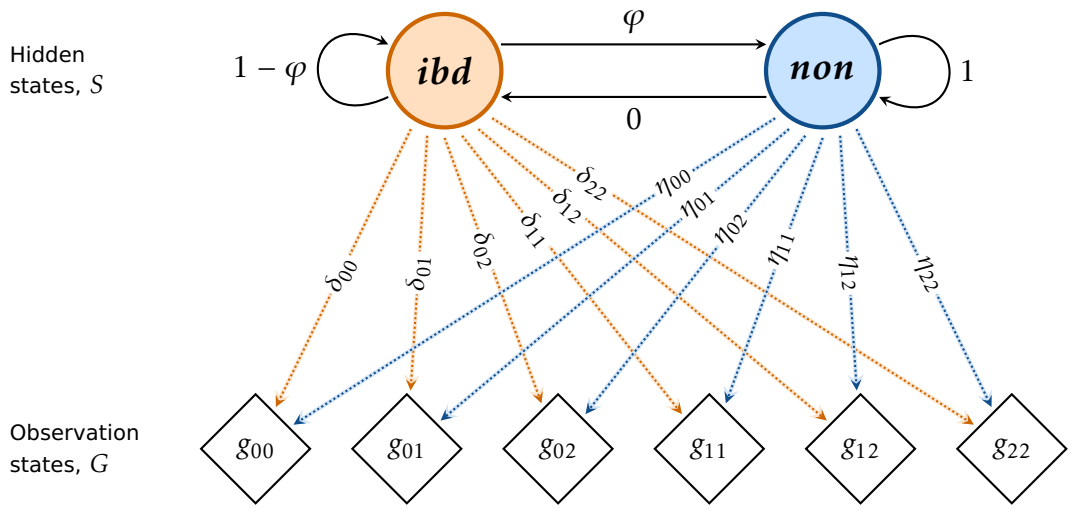


Figure 4.13: Illustration of the Hidden Markov Model for IBD inference. Two hidden states are assumed to generate the observations in a Markov process; *ibd* and *non*. Transitions from each state into any state are indicated by *solid* lines. The probability of transition from *ibd* to *non* is denoted 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*.

In particular, the probability of transition from *ibd* to *non*, denoted by $\varphi = \psi_{j,k}(non \mid ibd)$, is modelled dependent on the rate of recombination between consecutive sites, in order to estimate the probability of the distance to the first recombination breakpoint along the sequence. Two variables are considered; the genetic distance between the current and the previous position, and the expected $\mathrm{T}_{\mathrm{MRCA}}$ of the focal f_k variant.

Let the genetic distance between positions b_j and b_{j-1} be denoted by r_j, measured in *Morgan*, which is the product of the recombination rate per site per generation, ρ, and the physical distance of the sequence interval in basepairs. If the recombination rate varies over the length of the chromosome, that is if a genetic map is available, r_j can be obtained from map distances. Note that the model considers $2r_j$ to account for recombination occurring along either of the two lineages considered. In a population genetics setting, time is scaled in units of $2N_e$ generations for a sample of diploid individuals, where N_e is the diploid effective population size of the population under consideration. Thus, the scaled rate of recombination within the interval between consecutive sites and per time unit is equal to $4N_e r_j$.

The expected age of an allele, measured in scaled time units and denoted by τ_k, can be estimated directly from its frequency; as already presented in Section 1.7 (page 35). Briefly, Kimura and Ota (1973) derived a formulation for the expected age of a selectively neutral allele in a stationary population using diffusion theory; see Equation (1.30). Later, Griffiths and Tavaré (1998) showed that the expected age of an allele in a constant population can be derived in context of the coalescent, given the assumptions of the infinite sites model; see Equation (1.31). Both approaches result in approximately equal distributions for allelic age with negligible differences; *e.g.* for a sample of $n = 1{,}000$ haplotypes, the expected age of an f_2 allele is $\mathbb{E}[t_m] = 0.025$ using Equation (1.30) and $\mathbb{E}[t_m] = 0.024$ using Equation (1.31). Here, Equation (1.30) was used for computation of τ_k due to its simplicity.

It should be noted that the expectation of allele age as used here implies the assumption of a constant population size, which is rarely observed in nature and also not the case for the simulated dataset on which the presented method was evaluated (as presented further below). The value of the expected age is nonetheless useful to arrive at approximate transition probabilities that are assumed to be suitable for the current HMM.

The distance to a recombination event follows the geometrical distribution if measured in discrete generations. However, it can be approximated on a continuous time scale using the exponential distribution in the limit as N_e tends to infinity; that is, generally, if population size is sufficiently large (see Hein *et al.*, 2004). Thus, the probability of transition from *ibd* to *non* can be expressed as follows.

$$\varphi \;=\; \psi_{j,k}(non \mid ibd) \;=\; 1 - \left(1 - \frac{4N_e r_j}{2N_e}\right)^{2N_e \tau_k} \;\approx\; 1 - \mathrm{e}^{-2N_e r_j \tau_k} \tag{4.7}$$

The probability of remaining in *ibd* is therefore $\psi_{j,k}(ibd \mid ibd) = 1 - \varphi$, because the probability distribution over possible states for a given state must sum to 1. For illustration, Figure 4.14 (next page) shows the probability of transition from *ibd* dependent on the genetic distance between consecutive sites along the sequence and the allele frequency of the focal allele.

Note that the model relies on the assumption that the probability of transition from *non* to *ibd* has zero probability; *i.e.*

$$\psi_{j,k}(ibd \mid non) = 0\,, \quad \psi_{j,k}(non \mid non) = 1\,.$$

Therefore, the architecture of the model is not fully connected or *ergodic*. This is typically referred to as a left-to-right or *Bakis* HMM, as transitions can only proceed in one direction. Once the *ibd* state has been left, the chain remains in the *non* state such that only the innermost IBD segment is inferred, relative to the focal site at the start of the sequence.

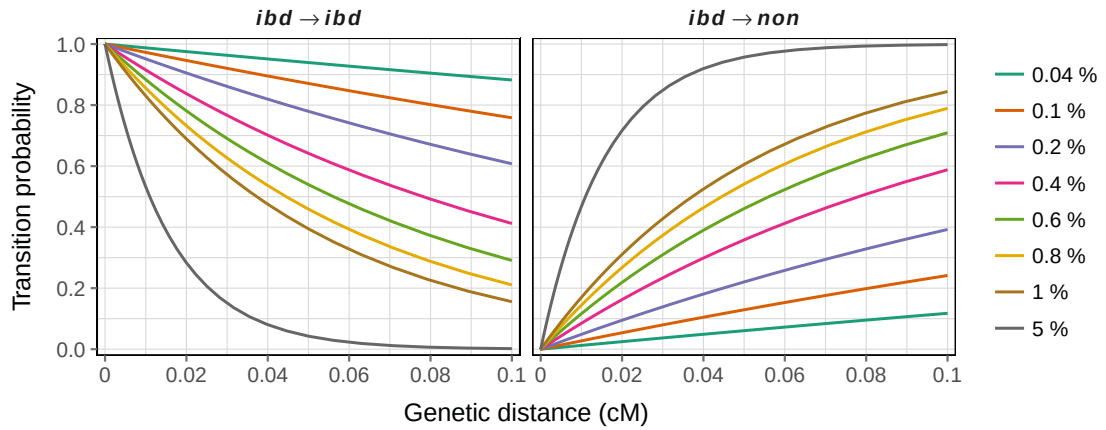


Figure 4.14: Probability distribution of transition dependent on IBD. The probability of transition was modelled dependent on the genetic distance between a particular site and the previous site and the expected age of the focal allele. The frequency of the focal allele determines its expected age, which is shown for different frequency values. An effective population size of N_e = 10,000 was specified. For example, the frequency of a f_2 allele in a sample of 5,000 haplotypes is equal to 0.04% (*green* line).

It is necessary to note that a pair of diploid individuals may share more than one recent haplotype identical by descent; *e.g.* along the same two chromosomes or any pair of the four chromosomes. Here, this possibility was not considered due to the variant-centric approach of the method. As such, inference is dependent on the properties of a given f_k variant. The focal allele serves as an indicator for haplotype sharing and transition probabilities are computed dependent on the expected time of the focal mutation event, given the allele frequency at the focal site. For example, by allowing transitions from the *non* state back to *ibd*, the IBD inference would be biased as the length of distinctly inferred segments (*i.e.* for other genealogies along the chromosome) would be conditioned on the expected age of the focal allele.

4.4.2.2 Emission probabilities

The model parameter ξ stores the emission or *output* probability vectors of the hidden states. Each vector is a probability distribution over the possible observation states with sum 1. There are six possible states in which a pair of genotypes can be observed. A genotype pair is denoted by g_{k_1,k_2}, where $k_1, k_2 \in \{0, 1, 2\}$. In the following, the emission probabilities for the possible genotype pairs are derived for each hidden state; *non* and *ibd*. The probability to observe a given genotype pair is written as $P_{non}(k_1, k_2) = \eta_{k_1 k_2}$ in *non*, and $P_{ibd}(k_1, k_2) = \delta_{k_1 k_2}$ in *ibd*.

Consider a pair of genotypes observed in two diploid individuals at a single locus. Each genotype can be observed in one of three possible states, which are again indexed by $k \in \{0, 1, 2\}$, where k counts the alternate alleles that compose a genotype. Recall that the expected frequency of a single genotype is $f_g(k) = \binom{n}{k}\, p^{n-k}\, q^k$ where $n = 2$ for the two haplotypes per individual, an where p and $q = 1 - p$ correspond to the frequency of the reference and alternate allele, respectively, as given in Equation (4.1) on page 111. In the general case, that is in a randomly mating population, the genotypes in both individuals are assumed to be independent. It follows that the expected frequency of a genotype pair is the joint probability of the expected genotype frequencies involved.

Table 4.3: Punnett squares of genotype pair partitions under non-IBD and IBD. Allele frequency contributions are itemised for each possible pair of genotypes. Rows and columns correspond to alleles in ordered haplotype combinations, (h_{c_1}, h_{c_2}), with $f_h(c = 0) = p$ and $f_h(c = 1) = q$, where $c \in \{0, 1\}$. Expressions in cells are the product of these combinations. Genotype pairs are formed by summing over the cells corresponding to the two genotypes in a given pair (labelled on the right in each row and at the bottom of each column). Panel **(a)** shows the partitions of expected frequencies for genotype pairs that do not share a haplotype (*i.e. non* state). In Panel **(b)**, if a haplotype is identical by descent (*i.e. ibd* state), one of the haplotypes is marked as shared; denoted by an asterisk, h^*_k. Note that a haplotype can only be shared, if contained in both row-by-column combinations, or frequencies are zero otherwise.

(a) *non*

	h_0, h_0	h_0, h_1	h_1, h_0	h_1, h_1	
h_0, h_0	p^4	p^3q	p^3q	p^2q^2	g_0
h_0, h_1	p^3q	p^2q^2	p^2q^2	pq^3	g_1
h_1, h_0	p^3q	p^2q^2	p^2q^2	pq^3	g_1
h_1, h_1	p^2q^2	pq^3	pq^3	q^4	g_2
	g_0	g_1	g_1	g_2	

(b) *ibd*

	h_0, h^*_0	h_0, h^*_1	h_1, h^*_0	h_1, h^*_1	
h_0, h^*_0	p^3	0	p^2q	0	g_0
h_0, h^*_1	0	p^2q	0	pq^2	g_1
h_1, h^*_0	p^2q	0	pq^2	0	g_1
h_1, h^*_1	0	pq^2	0	q^3	g_2
	g_0	g_1	g_1	g_2	

Here, independence of genotype frequencies is assumed for the *non* state, but which does not apply if the two individuals share a haplotype identical by descent as considered in the *ibd* state. For example, under the infinite sites model, it is expected that genotypes $g_{0,2}$ and $g_{2,0}$ cannot be observed if they share a haplotype by descent. For simplicity, Table 4.3 (this page) provides a convenient representation of the composition of haplotypes per genotype pair in *non* and *ibd*, from which the expected genotype pair frequency can be derived.

The probability of observing genotype pair g_{k_1,k_2} is equal to its frequency in the sample. In the *non* state, the expectation is given by

$$\eta_{k_1k_2} = \begin{cases} p^4 & \text{if} \quad k_1 = 0, k_2 = 0 \\ 4p^3q & \text{if} \quad k_1 = 0, k_2 = 1 \quad \text{or} \quad k_1 = 1, k_2 = 0 \\ 2p^2q^2 & \text{if} \quad k_1 = 0, k_2 = 2 \quad \text{or} \quad k_1 = 2, k_2 = 0 \\ 4p^2q^2 & \text{if} \quad k_1 = 1, k_2 = 1 \\ 4pq^3 & \text{if} \quad k_1 = 1, k_2 = 2 \quad \text{or} \quad k_1 = 2, k_2 = 1 \\ q^4 & \text{if} \quad k_1 = 2, k_2 = 2 \end{cases} \tag{4.8}$$

and likewise, for the *ibd* state, by

$$\delta_{k_1k_2} = \begin{cases} p^3 & \text{if} \quad k_1 = 0, k_2 = 0 \\ 2p^2q & \text{if} \quad k_1 = 0, k_2 = 1 \quad \text{or} \quad k_1 = 1, k_2 = 0 \\ 0 & \text{if} \quad k_1 = 0, k_2 = 2 \quad \text{or} \quad k_1 = 2, k_2 = 0 \\ p^2q + pq^2 & \text{if} \quad k_1 = 1, k_2 = 1 \\ 2pq^2 & \text{if} \quad k_1 = 1, k_2 = 2 \quad \text{or} \quad k_1 = 2, k_2 = 1 \\ q^3 & \text{if} \quad k_1 = 2, k_2 = 2. \end{cases} \tag{4.9}$$

However, note that Equation (4.9) above implicitly assumes that no mutations have occurred on either lineage after co-inheritance of the shared haplotype. While this assumption may hold for a haplotype co-inherited only a few generations ago, it is easily violated and therefore unrealistic for the general case.

The expected emission probability distributions for each possible genotype pair in *non* and *ibd* are shown in Figure 4.15 (next page).

4.4.2.3 Initial state probabilities

The model parameter π stores the probabilities of being in either state at the start of the sequence. Since the focal allele is used to identify the shared haplotype in a pair of individuals, the probability of being in *ibd* is assumed to be $\pi_{ibd} = 1$, such that $\pi_{non} = 0$.

4.4.3 Integration of empirically determined error rates

In this section, the data generated in Section 4.2 (page 114) were used to inform the model parameters in the HMM. This was done, first, to validate the expectations formulated in the previous sections, second, to explore variation instigated by genotype error and, third, to obtain empirical parameter values for emission and initial state probabilities in *non* and *ibd*.

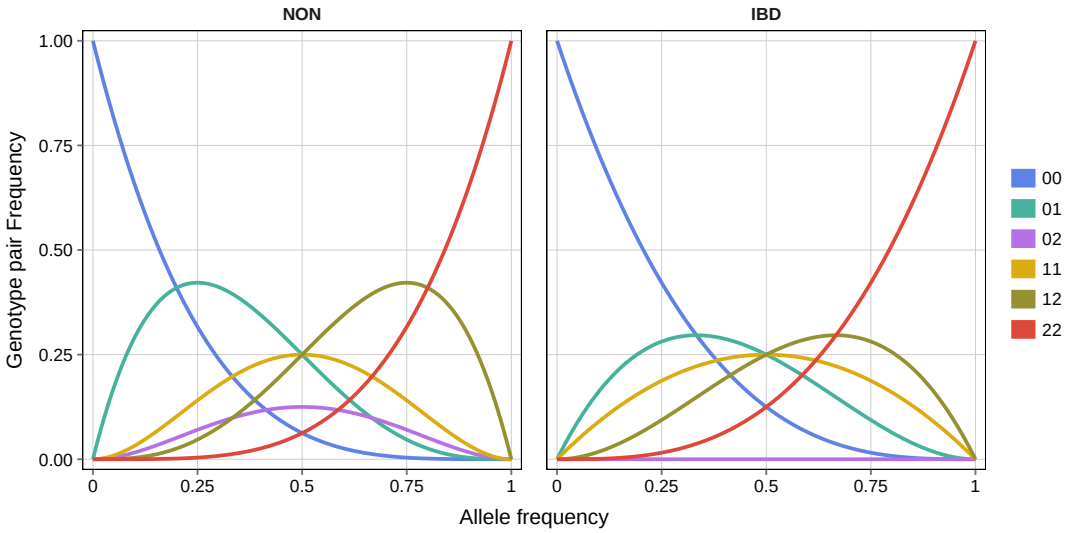


Figure 4.15: Expected frequency distribution of genotype pairs under non-IBD and IBD. Proportions were calculated using Equation (4.8) and Equation (4.9) in both hidden states, *non* and *ibd*, respectively. Colours distinguish the six possible genotype pairs, given by g_{k_1,k_2}, as indicated.

Note that the effect of genotype error on state transition probabilities is not considered. The computation of transition probabilities include the expected age of a focal allele dependent on its frequency, which could be biased in presence of genotype error, but where deviations are expected to be negligibly small if sample size is large. In particular, the expected age represents an approximation to the T_{MRCA} of the focal allele, which is more likely to be affected by unconsidered demographic parameters such as selection, migration, growth, and population structure, as well as sampling bias.

Two datasets were available from previous analyses; the original genotype matrix as produced from simulated haplotypes, denoted by $\mathcal{D}$, and a corresponding, but modified genotype matrix, $\mathcal{D}^*$, in which the empirical, frequency-dependent proportions of genotype error were included in Section 4.3 (page 124). These datasets allowed analysis *before* and *after* error, respectively. Data consisted of $n = 2{,}500$ individuals and $m = 672{,}847$ variant sites.

4.4.3.1 Empirical emission probabilities

Information about IBD status was available through coalescent records obtained in the simulation. By performing scans over all coalescent trees, true IBD intervals were determined for f_k variants at $k \in \{2, \ldots, 25\}$ (allele frequency between 0.04% and 0.5%). In total, a set of 11.598 million true IBD segments was compiled. Each segment was recorded

as a tuple of two breakpoint coordinates (b_L and b_R to the left and right-hand side of a focal variant, respectively) and two individuals, *i.e.* indices for the pair of individuals who share a haplotype identical by descent within the breakpoint interval.

The set of compiled IBD segments was used to determine the empirical probability to observe a given genotype pair in *ibd*. This was done by randomly sampling 500,000 segments with replacement, for which genotype data were extracted in $[b_{L+1}, b_{R-1}]$ for the two individuals. Note that breakpoint sites were excluded to ensure IBD over the entire region. For each segment, extracted genotype sequences were paired and collected by their coordinates along the length of the chromosome. In a similar fashion, the empirical probability of observing genotype pairs in *non* was determined using the same sample of segments, but where the two individuals sharing the IBD segment were ignored. Instead, the two individuals were drawn at random from the subset of samples which did not share a haplotype IBD within $[b_{L+1}, b_{R-1}]$. After sampling was complete, genotype pairs were aggregated by allele frequency per site, such that the frequency-dependent proportion of each genotype pair could be calculated in *ibd* and *non*. In both cases, genotype data were taken separately from $\mathcal{D}$ and $\mathcal{D}^*$ to measure proportions before and after error.

The resulting probability distributions after error were used to define the empirical emission model in *ibd* and *non*, again denoted by $\delta_{k_1 k_2}$ and $\eta_{k_1 k_2}$, respectively. For illustration, deviations from expected genotype pair proportions are shown in Figure 4.16 (next page), both before error (4.16a) and after error (4.16b). Expectations in *ibd* and *non* were calculated according to Equations (4.8) and (4.9) on page 144, respectively. Differences were calculated by subtracting empirical from expected genotype pair proportions, which was done in discrete allele frequency units, but also averaged per frequency bin, in 100 bins of equal size across the allele frequency spectrum.

Before error, empirical and expected proportions in *non* were equal on average, in each of the six possible genotype pairs. The variability along the allele frequency spectrum was negligibly small, where deviations per frequency unit were seen as stochastic noise around the mean and ranged between −1% and +1%. In contrast, the variability across frequency units was overall amplified in *ibd*. The mean proportion of g_{11} was up to 2% higher than expected towards the higher end of the frequency spectrum, whereas g_{22} was up to 2% lower than expected towards higher frequencies. Notably, g_{02} is expected to have a constant zero probability of observation in *ibd*, which was confirmed from the data.

After error, overall variability increased in each comparison. In *non*, the mean proportion of g_{11} showed deviations of up to +1% towards 50% allele frequency, which was also seen for g_{12}, but towards higher frequencies. In *ibd*, mean proportions showed

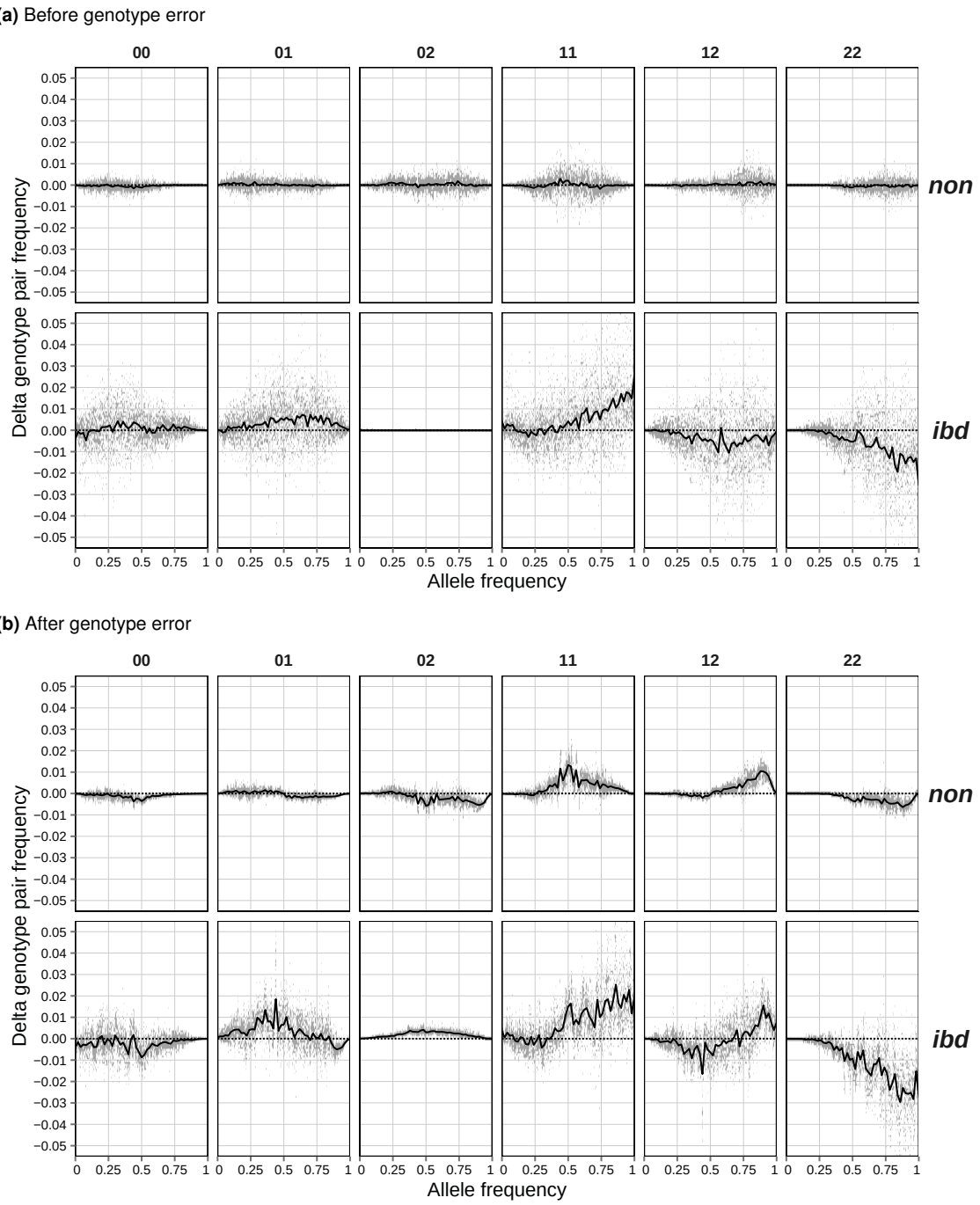


Figure 4.16: Difference between empirical and expected proportions of genotype pairs. In total, 500,000 segments were sampled in *non* and *ibd* as determined from coalescent records. Segments were aggregated by allele frequency to calculate empirical proportions for each of the six possible genotype pairs (g_{k_1,k_2}, indicated above each panel). Delta values were calculated by subtracting empirical from expected proportions; the latter were calculated using Equations (4.8) and (4.9) under *non* and *ibd*, respectively. Each panel is a scatterplot showing the deviation at each discrete step in allele frequency. The mean (±SE) of delta values was calculated in steps of 1% allele frequency; indicated by the *black* line. Results in Panel **(a)** were generated on data before the inclusion of genotype error, D, and Panel **(b)** on data after genotype error was included, D^*.

as similar distribution as in comparisons before error, but where the difference between empirical and expected values was further increased. For example, deviations of g_{11} were increased up to +2.5% towards higher frequencies, which was mirrored by g_{22} but reaching up to −3%. Importantly, on average the empirical proportion of g_{02} was non-zero along the frequency spectrum, but which increased up to +0.5% towards 50% allele frequency.

4.4.3.2 Empirical initial state probabilities

Genotype error can affect the allele frequency distribution and thus bias the identification of individuals which share an allele at a given site. Some of the formed pairs may therefore be wrongly included, whereas some others may be missed. In particular, the following four cases can be distinguished:

(a) **True positives.** The focal allele correctly identifies haplotype sharing in two individuals which are heterozygous for the allele; *i.e.* $g_1 \rightarrow g_1$.

(b) **False positives.** The focal allele is observed in a misclassified genotype, $g_0 \rightarrow g_1$, such that IBD is wrongly assumed for a pair which does not share a haplotype. Note that the change $g_2 \rightarrow g_1$ also leads to the inclusion of an individual which actually is homozygous for the focal allele, but which is not considered in the model.

(c) **False negatives.** The genotype of an individual was misclassified at the focal site, $g_1 \rightarrow g_0$, such that the focal allele is missed and the individual wrongly excluded. Note that this also considers the change $g_1 \rightarrow g_2$, leading to the exclusion of the individual due to the assumptions of the model.

(d) **True negatives.** An individual is correctly excluded due to not being heterozygous for the focal allele, *i.e.* $g_0 \rightarrow g_0$ or $g_2 \rightarrow g_2$, as well as $g_0 \rightarrow g_2$ or $g_2 \rightarrow g_0$

The inference of IBD segments in a pair where at least one individual is a false positive, Case (b), is likely to result in a disproportionately reduced segment length. In principle, such falsely identified individuals, and thereby specific false genotypes, may be exposed if segment lengths are consistently shorter than expected in each pairwise analysis. On the other hand, genotype error leading to false negatives, Case (c), is inadvertently missed, because it is not directly possible to assume that particular individuals carry the focal allele if not observed in the data.

The proportion of genotype pairs identified as true positives, Case (a), is relevant to determine the probability of the initial state at the start of the sequence. The true positive rate was determined by comparison of the data before and after error. All f_k variants at

$k > 1$ were identified in $\mathcal{D}^*$, as well as all the individuals carrying the alternate allele at a particular variant site. This resulted in a set of matrix coordinates (marker by individual) which were pooled into site frequency bins, defined by k. Bins with less than 1,000 markers were removed. Then, for each k, all possible pairs of individuals were formed at each marker and the dataset $\mathcal{D}$ was queried with the joint set of coordinates. This was done to extract the corresponding vector of true genotype pairs, from which the true positive rate was calculated as the proportion of pairs in which both genotypes were heterozygous.

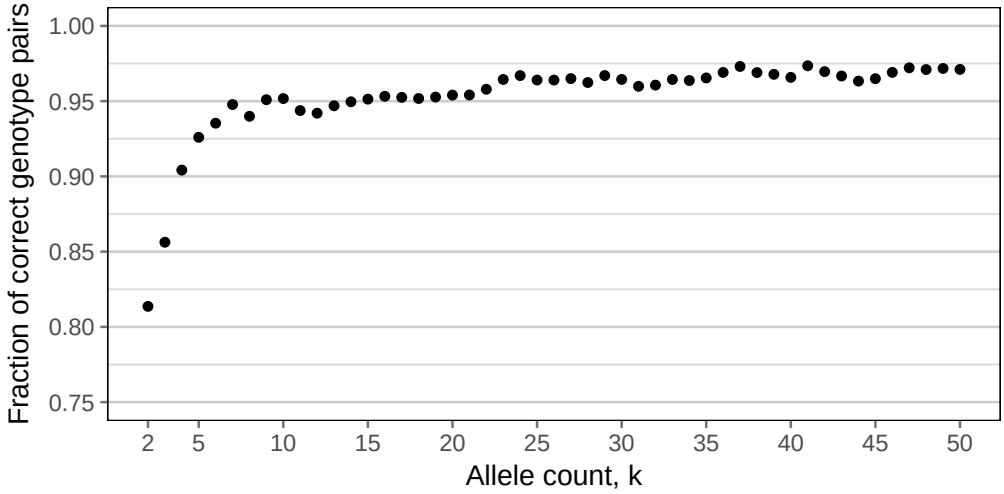


Figure 4.17: True positive rate of identified genotype pairs at focal sites. Pairwise shared genotypes at focal f_k variants with $k > 1$ were compared between datasets before and after error. The true positive rate was determined for each k. Results are shown for k in $[2, 50]$, which corresponds to an allele frequency between 0.04% and 1%.

The empirical distribution of correctly identified genotype pairs was used to define the initial state probability of being in *ibd*, given the frequency of the focal allele expressed in f_k. The resulting distribution is shown in Figure 4.17 (this page), for focal variants with k in $[2, 50]$, corresponding to an allele frequency between 0.04% and 1%. The fraction of correctly observed genotype pairs was lowest for f_2 variants, found at 0.812, but rapidly increased to 0.913 and 0.950 for f_5 and f_{10} variants, respectively. At higher frequencies, the true positive rate stabalised around 0.975 and 0.995. At frequencies near 100%, however, the number of markers observed per k was too low to provide conclusive estimates. These values were stored in an array such that the initial state probability for a given k can be accessed through the functions $\pi_{ibd}(k)$ and $\pi_{non}(k) = 1 - \pi_{ibd}(k)$.

4.4.4 Inference of IBD segments

The aim of the presented IBD-model is to find the most likely position along the sequence at which the *ibd* state changes into the *non* state, which is done independently to the left and right-hand side of the focal f_k variant; *i.e.* the focal site sits at the start of both observation sequences. Recall that the IBD segment around the focal variant is defined by the interval $[b_L, b_R]$, where b_L and b_R denote the breakpoints which delimit the region in which at least one recombination event is likely to have occurred to the left and right-hand side of the focal variant site at position b_i. To infer this interval, the most likely hidden state which generated the observed genotype pair is inferred at each site along the sequence. In general, given H hidden states and an observation sequence of length m, there are H^m possible state sequences. For example, given this two-state HMM and a short region of only 100 genotype pairs, the number of possible state sequences already exceeds the number of seconds the universe has existed.* To circumvent this problem, Rabiner (1989) formally advised the use of the *Viterbi algorithm* for sequence classification in HMMs, which scales quadratically with the number of hidden states and has a time complexity of $O(H^2 m)$.

The Viterbi algorithm is a dynamic programming technique which finds the most likely sequence of hidden states that maximises the probability of observing the data (Viterbi, 1967; Forney, 1973). Let X_j denote the hidden state at site j which generated the observed genotype pair o_j. Following Rabiner (1989), the probability of the most likely sequence of hidden states until site j and ending in state x is given by

$$v_j(x) = \max_{X_0, X_1, \ldots, X_{j-1}} P(X_0, X_1, \ldots, X_{j-1},\ X_j = x,\ o_1, o_2, \ldots, o_j \mid \lambda) \tag{4.10}$$

where λ denotes the model; see Equation (4.5) on page 139. The procedure to retrieve the actual state sequence is summarised as follows.

1. **Initialisation.** The probability that a given state generated the genotype pair observed at the focal site is simply the product of its initialisation and emission probabilities. If genotype error is included, the initialisation probability is defined conditionally on the frequency of the focal f_k variant. Note that emission probabilities were defined as $\delta_{k_1 k_2}$ and $\eta_{k_1 k_2}$ in *ibd* and *non*, respectively. For simplicity, these are now written as $\delta_j(o_j)$ and $\eta_j(o_j)$, where the index j refers to the position in the sequence at which the allele

* Current age of the universe: 42×10^{16} seconds [Date accessed: 2017-02-18]

frequency is taken to retrieve the frequency-dependent probability of the observed genotype pair at site j.

$$\begin{aligned} v_0(ibd) &= \pi_{ibd}(k)\ \delta_0(o_0) \\ v_0(non) &= \pi_{non}(k)\ \eta_0(o_0) \end{aligned} \tag{4.11}$$

The Viterbi algorithm involves the successive multiplication of probabilities during the recursion step (see below), which may result in values too small to be distinguishable from zero using conventional computers. To avoid this problem, a commonly implemented solution is a log-transformation of probabilities. Here, it is more convenient (and computationally less demanding) to define a weighting function to obtain a scaling factor which is stored in an additional array, w.

$$w_0 = \max_{x \in S} \left[v_0(x) \right] \quad \textbf{s.t.} \quad v_0'(x) = \frac{v_0(x)}{w_0} \quad \forall \quad x \in S \tag{4.12}$$

2. **Recursion.** The array u is defined to keep track of the states traversed along the path; that is, $u_j(x)$ stores a back-pointer to the state at site $j-1$ which resulted in the highest probability $v_j(x)$ at site j.

$$u_j(x) = \arg\max_{y \in S} \left[v_{j-1}'(x)\ \psi_{j,k}(y \mid x) \right] \quad \forall \quad x \in S;\ j = 1, 2, \ldots, m \tag{4.13}$$

Recall that $\psi_{j,k}$ refers to the transition probability from a given state to another or the same state, and is dependent on the frequency of the focal allele (k), as defined in Equation (4.7), page 141. The chain proceeds through the most likely path at each site along the sequence by following the transitions that maximise the probability of observing a given state. Given Equation (4.10), by induction on j it follows that

$$\begin{aligned} v_j(ibd) &= \delta_j(o_j)\ \max_{y \in S} \left[v_{j-1}'(ibd)\ \ \psi_{j,k}(y \mid ibd) \right], \quad j = 1, 2, \ldots, m \\ v_j(non) &= \eta_j(o_j)\ \max_{y \in S} \left[v_{j-1}'(non)\ \psi_{j,k}(y \mid non) \right], \quad j = 1, 2, \ldots, m. \end{aligned} \tag{4.14}$$

Note that the current state probability is computed conditionally on the weighted probability value at the immediate previous site, but which does not affect the outcome of the maximisation.

$$w_j = \max_{x \in S} \left[v_j(x) \right] \quad \textbf{s.t.} \quad v_j'(x) = \frac{v_j(x)}{w_j} \quad \forall \quad x \in S;\ j = 1, 2, \ldots, m \tag{4.15}$$

3. **Termination.** At the last site in the sequence, m, the state with the highest probability is picked to mark the final state of the most likely sequence of hidden states (*i.e.* the "Viterbi path"), denoted by X^*.

$$X_m^* = \arg\max_{x \in S} \left[v'_m(x) \right] \tag{4.16}$$

4. **Path backtracking.** Given the array of back-pointers, u, the most likely path is found by tracing back from the final state until the initial site in the sequence.

$$X_j^* = u_{j+1}\left(X_{j+1}^*\right), \quad j = m-1, m-2, \ldots, 0 \tag{4.17}$$

The IBD segment is determined from the two resulting state sequences, which were obtained independently from the observation sequence to the left and right-hand side of the focal position. The Viterbi paths on the left and right-hand side are denoted by L^* and R^*, respectively. The breakpoint interval defining the segment, $[b_L, b_R]$, is found by scanning each path from its start (*i.e.* from a given target site) to the first position at which the *non* state was inferred. Note that this includes the site of the first *non* state, which is defined as a breakpoint. In boundary cases, when each site until the end of the chromosome was inferred as being in the *ibd* state, the last site in the sequence is taken as a breakpoint.

4.4.5 Results

The HMM-based method for IBD inference was evaluated on the same error-treated dataset as used in Section 4.3.2 (page 126), as well as the corresponding dataset before the integration of (realistic) error rates. As was done in the previous analyses, if multiple identical breakpoint intervals were inferred for a given pair, only the one detected around the allele of the lowest frequency within that interval was retained (or one was sampled if multiple shared alleles occurred at the same frequency). The retained (*unique*) segments were thereby tagged by the presumably youngest shared allele within a given interval, such that breakpoint accuracy could be measured conditional on the frequency of the target allele.

4.4.5.1 Shared haplotype inference in simulated data before and after error

The number of unique segments inferred was 3.179 million and 3.236 million before and after error, which corresponds to 32.250 % and 32.827 % of the reported set of IBD segments, respectively. In comparison, the number of unique segments in the set of true IBD segments, those determined from simulation records on the same targets, was 2.721 million (27.599 %). The proportion of breakpoints overestimated (relative to the corresponding true breakpoint position) was similar for both datasets; 55.716 % before error and 54.094 % after error.

Table 4.4: Accuracy comparison per f_k category after error. The accuracy of detected IBD breakpoints was measured using the squared Pearson correlation coefficient, r^2, and the RMSLE in relation to the true IBD segments determined from simulation records; measured in terms of the distance between breakpoint site and the corresponding focal position per segment. Results were obtained on the error-treated dataset, D^* (which included the true haplotypes, phased haplotypes, and genotype data), using the FGT, DGT, and the HMM-based method for targeted IBD detection. Accuracy was measured after duplicate segments had been removed, and the same set of target sites was assessed in each approach; separately computed per f_k category. The approach with the highest accuracy (highest r^2 and lowest RMSLE) per f_k is indicated (**bold**).

f_k	Freq. (%)	r^2				RMSLE			
		FGT*	FGT**	DGT	HMM	FGT*	FGT**	DGT	HMM
2	0.04	0.011	0.015	0.018	**0.982**	1.208	1.362	1.029	**0.390**
3	0.06	0.031	0.039	0.050	**0.908**	1.013	1.112	0.855	**0.452**
4	0.08	0.049	0.060	0.079	**0.832**	0.924	0.993	0.792	**0.490**
5	0.10	0.064	0.075	0.097	**0.755**	0.869	0.913	0.756	**0.534**
6	0.12	0.084	0.093	0.120	**0.682**	0.832	0.863	0.730	**0.555**
7	0.14	0.089	0.097	0.123	**0.602**	0.791	0.810	0.711	**0.582**
8	0.16	0.094	0.106	0.135	**0.575**	0.767	0.784	0.692	**0.588**
9	0.18	0.105	0.112	0.138	**0.523**	0.754	0.766	0.690	**0.613**
10	0.20	0.113	0.123	0.147	**0.492**	0.733	0.740	0.682	**0.635**
11	0.22	0.122	0.128	0.149	**0.440**	0.713	0.716	0.677	**0.659**
12	0.24	0.139	0.144	0.173	**0.424**	0.718	0.721	0.691	**0.654**
13	0.26	0.117	0.120	0.154	**0.424**	0.708	0.711	0.686	**0.675**
14	0.28	0.149	0.157	0.178	**0.386**	0.699	0.696	0.683	**0.678**
15	0.30	0.126	0.129	0.151	**0.408**	0.691	0.690	**0.676**	0.681
16	0.32	0.146	0.150	0.175	**0.379**	0.676	0.675	**0.669**	0.689
17	0.34	0.132	0.140	0.158	**0.312**	0.683	**0.682**	0.690	0.712
18	0.36	0.143	0.158	0.175	**0.334**	0.669	**0.667**	0.669	0.702
19	0.38	0.149	0.153	0.170	**0.303**	0.675	**0.669**	0.673	0.705
20	0.40	0.173	0.179	0.192	**0.327**	0.664	**0.665**	0.681	0.716
21	0.42	0.154	0.165	0.172	**0.309**	0.667	**0.660**	0.682	0.720
22	0.44	0.151	0.154	0.160	**0.257**	0.659	**0.657**	0.684	0.725
23	0.46	0.139	0.144	0.160	**0.265**	0.653	**0.649**	0.675	0.723
24	0.48	0.153	0.153	0.168	**0.247**	0.663	**0.655**	0.690	0.746
25	0.50	0.098	0.102	0.102	**0.239**	0.664	**0.656**	0.702	0.740

* True haplotypes
** Phased haplotypes

Overall accuracy, measured as the physical distance between breakpoint and target site, was $r^2 = 0.634$ (RMSLE = 0.781) before error and $r^2 = 0.638$ (RMSLE = 0.791) after error. While it appears from these results that accuracy was relatively low, consider accuracy measured per focal allele frequency. Table 4.4 (page 153) shows r^2 and RMSLE measured per f_k category, where the HMM is compared to corresponding results obtained on the same set of target sites using the FGT on true (simulated) haplotypes as well as phased haplotypes, and the DGT on genotype data; shown for analyses conducted after the integration of error. The HMM-based IBD detection approach outperformed each of the rule-based methods (in terms of r^2). However, for each method, accuracy decreases rapidly towards higher frequencies.

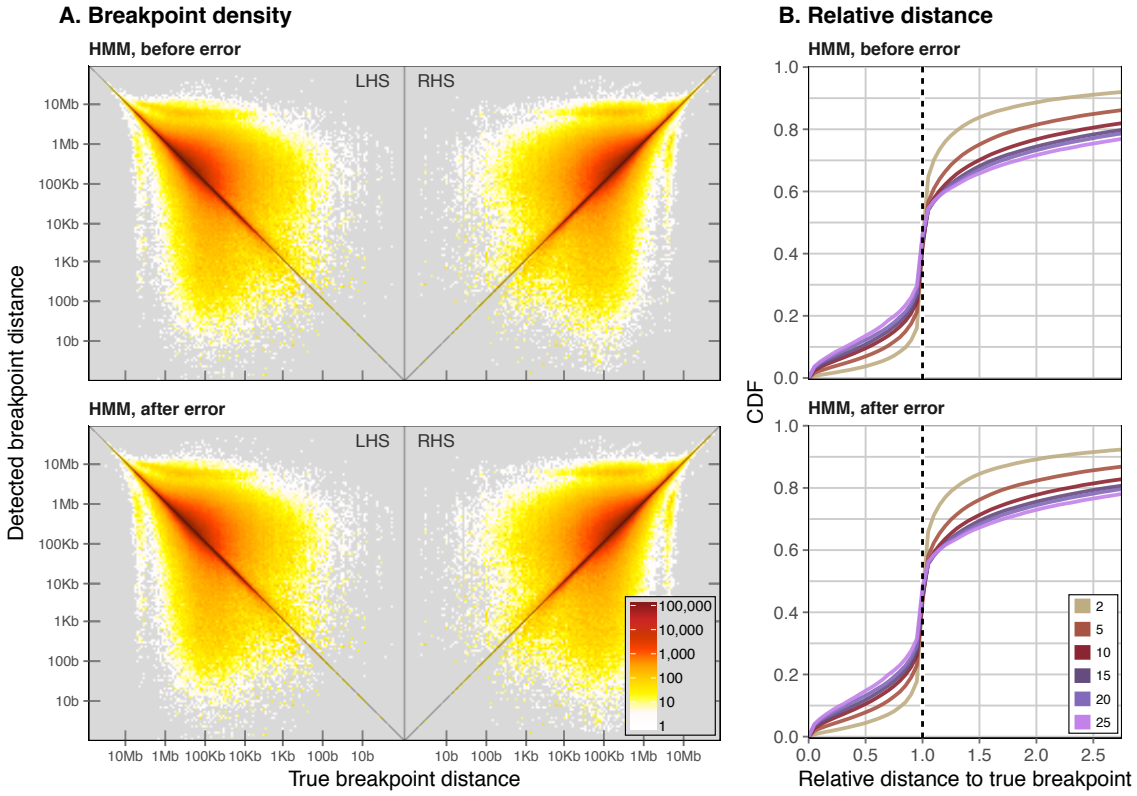


Figure 4.18: Breakpoint accuracy using the HMM before and after error. Panel **(A)** shows the density in terms of the physical distance between true and detected breakpoints and the position of a given focal allele. Panel **(B)** shows the physical length in terms of the relative distance between a focal site and the detected breakpoint. See Figure 4.7 (page 129) for a detailed description.

Median length was assessed after removal of boundary cases; 1.241 % of segments were removed in the analysis conducted on data before the integration of error and 1.216 % after error. This proportion was similar for the set of true IBD segments (1.377 %). Before

error, overall median length was 0.526 Mb (0.884 cM), and 0.504 Mb (0.845 cM) after error; this is compared to the shorter median length found for the true dataset; 0.343 Mb (0.590 cM). At f_2 alleles, median length was 2.209 Mb (3.690 cM) for the set of true shared haplotype segments, which is shorter compared to the inferred length of 2.499 Mb (4.207 cM) before error and 2.458 Mb (4.129 cM) after error. Lengths decreased towards higher focal allele frequencies, but where inferred lengths still remained overestimated; *e.g.* at f_{25} alleles, median length for true segments was 0.231 Mb (0.407 cM), compared to 0.379 Mb (0.626 cM) before error and 0.360 Mb (0.593 cM) after error. The distribution of IBD length by focal allele frequency is shown in Figure 4.19 (this page).

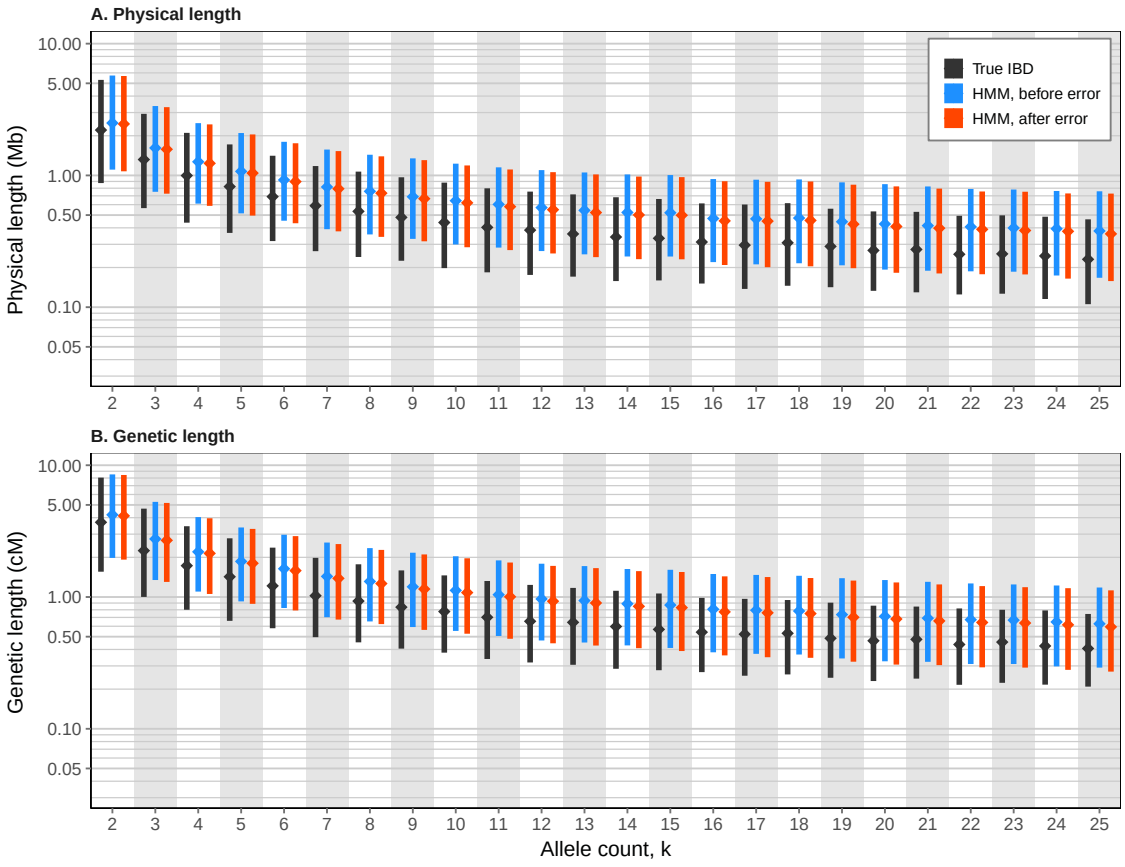


Figure 4.19: IBD lengths using the HMM before and after error. The HMM-based approach for targeted IBD detection was applied to simulated data before and after the integration of error; *i.e.* datasets D and D^*, respectively. Lengths were compared to the corresponding set of true breakpoint segments as determined from simulation records. Bottom and top of each bar indicate 1st and 3rd quartiles, respectively, between which the median (2nd quartile) is marked (*diamonds*).

4.4.5.2 Analysis of 1000 Genomes data, chromosome 20

In the previous analysis of 1000G data in Section 3.5.0.2 (page 103), I showed that neither the FGT nor the DGT were likely to detect shared haplotype segments with sufficient accuracy. I therefore attempted to demonstrate in this chapter (Section 4.3) that presence of error in the data can have a dramatic impact on such rule-based detection methods. It was suggested that even relatively small error rates (as expected in real data) may lead to the detection of false positive breakpoints and, thus, the observation of truncated breakpoint intervals in the majority of scans. In the section above, I demonstrated that the HMM-based approach for targeted shared haplotype inference is robust towards error, but where overall length of detected segments is likely to be overestimated. Nonetheless, it should be possible to re-produce similar patterns when applying this method to real data.

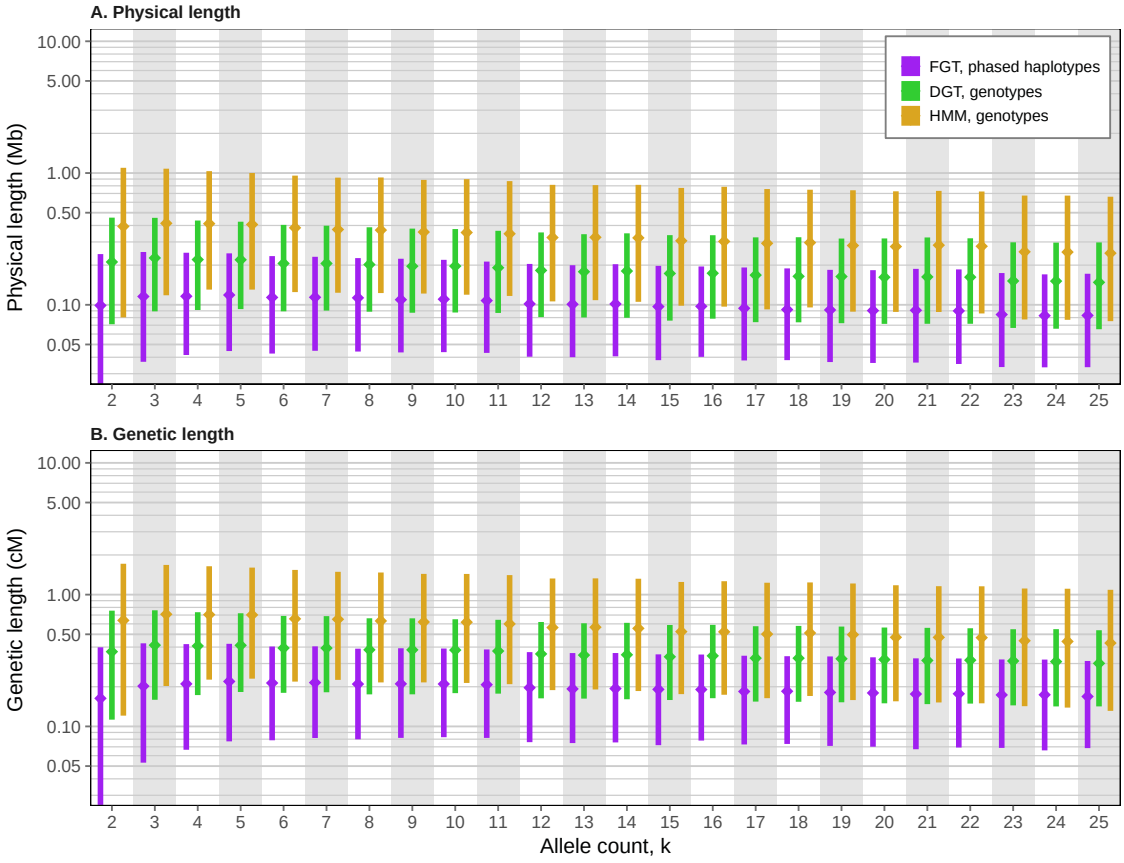


Figure 4.20: IBD inference using the HMM on 1000 Genomes data, chr. 20. IBD detection using the HMM-based method was performed under the empirical error model defined on genotype data from the 1000 Genomes Project. The resulting length distribution is compared to previous results obtained on the same set of target sites using the FGT and DGT. Bottom and top of each bar indicate 1st and 3rd quartiles, respectively, between which the median (2nd quartile) is marked (*diamonds*).

I applied the HMM-based method to data from chromosome 20 in 1000 Genomes Project Phase III and inferred 18.040 million shared haplotype segments around all variants observed at $f_{[2,25]}$ (*i.e.* frequency below 0.5%) of which 39.3 % were unique. For direct comparison to corresponding results obtained using the FGT and DGT (see Section 3.5.0.2), I retained a random subset of 1 million unique segments that were inferred around the same target site and haplotype pair in each method, after also removing boundary cases.

Overall median physical (genetic) length was 0.312 Mb (0.538 cM) for the HMM-based detection method, which was longer compared to the FGT and DGT where median length was 0.098 Mb (0.190 cM) and 0.176 Mb (0.344 cM), respectively. The distribution of IBD length by focal allele frequency is shown in Figure 4.20 (page 156). While it is shown that only the HMM is able to infer longer IBD segments that are more consistent with expectations, such a comparison is limited due to population structure in the 1000G sample. For example, the simulated dataset was generated as a sample of "European" haplotypes only (as defined in the demographic model; see Section 3.4.1.1, page 90.

Since IBD is delimited by past recombination events that occurred independently along each lineage back in time, the length of a shared haplotype region is indicative for the time since a haplotype was co-inherited from a common ancestor. It is therefore expected that IBD segments are longer within a given population than segments shared between individuals from different populations. I further subdivided the set of IBD segments based on the focal allele being shared between individuals from the same or different populations, as recorded in 1000G. Figure 4.21 on next page shows the distribution of median physical and genetic IBD length by focal allele frequency for each case, which also includes the results obtained using the FGT and DGT, where segments were detected on the same target site and haplotype pair. Table 4.5 on page 159 gives the overall median lengths for each comparison (sharing within and between populations).

The length of segments inferred around alleles shared within the same population decreased towards higher allele frequencies and were generally longer compared to haplotypes shared between a given population and any of the others. Such expected differences were more pronounced for results obtained using the HMM. While the FGT and DGT overall resulted in shorter detected haplotype segments, the HMM was able to infer even shorter segments; for example, see haplotype sharing between African (AFR) and East Asian (EAS) individuals in Figure 4.21. Further, note that the effect of error on the FGT and DGT would be reduced at shorter IBD segments, as it becomes less likely to encounter false positive breakpoints.

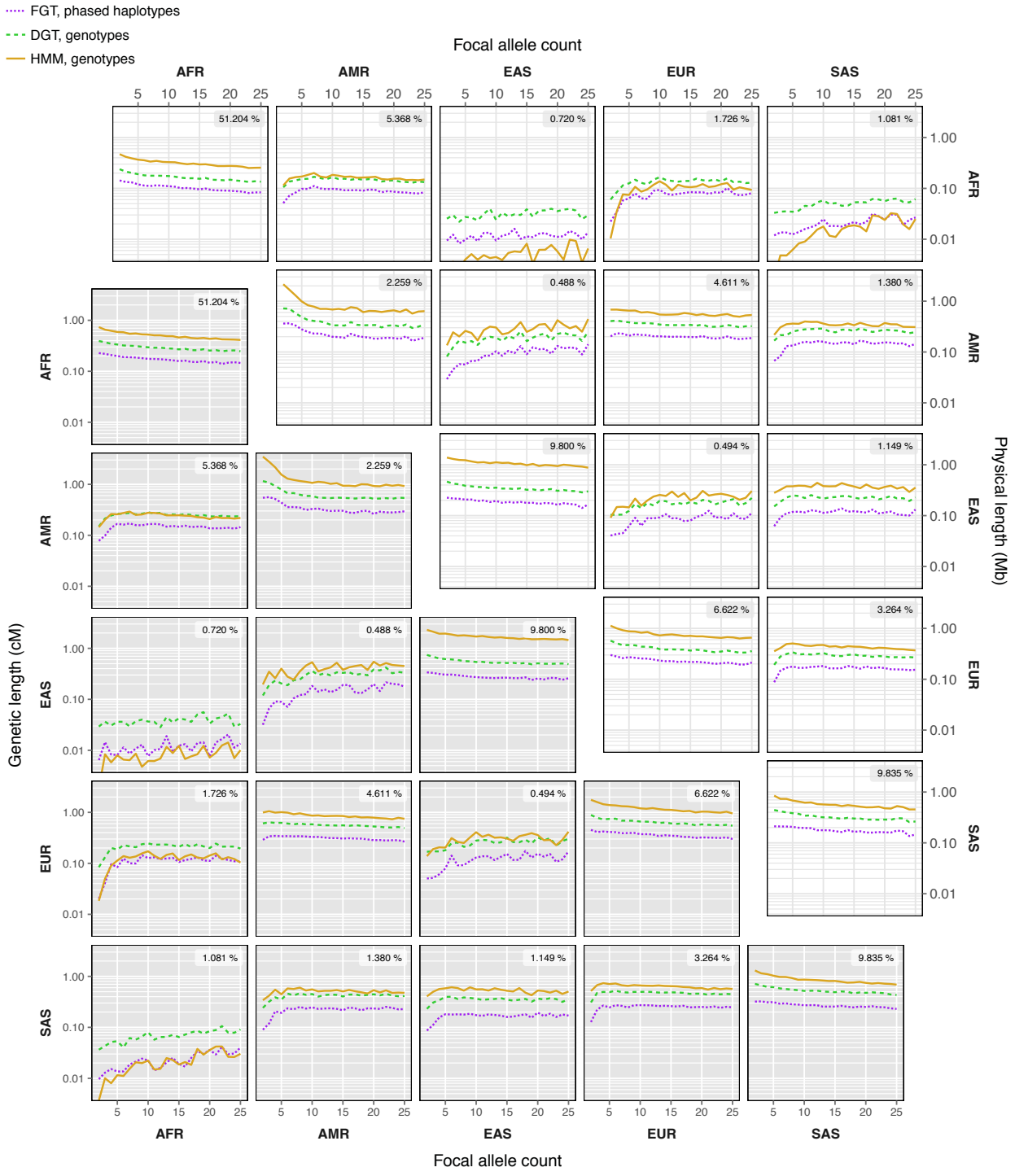


Figure 4.21: Inferred shared haplotype lengths by population in 1000 Genomes, chr. 20. The distributions of physical (upper triangle) and genetic (lower triangle) length by frequency of the focal allele are shown for alleles shared between pairs of individuals from the same population (*diagonal* panels) and any of the other populations sampled in 1000G. The 1000G Phase III dataset comprises samples from five continental populations (or *super-populations*); African (AFR), Ad-Mixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). Results are shown for shared haplotype lengths inferred using the FGT, DGT, and HMM, on the same set of segments inferred at target sites found at $f_{[2,25]}$ (*i.e.* allele frequency below 0.5%). The proportion of haplotypes shared within or between each population is given in each panel (upper right corner).

Table 4.5: Median shared haplotype length per population in 1000 Genomes, chr. 20. Shared haplotype segments found around alleles shared within and between populations, as contained within 1000G. The 1000G Phase III dataset comprises samples from five continental populations (or *super-populations*); African (AFR), Ad-Mixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). The results shown compare median physical and genetic lengths of segments detected using FGT, DGT, and HMM. Shared haplotypes inferred around alleles shared within the same population are marked (∗).

Populations			Physical length (Mb)			Genetic length (cM)		
			FGT	DGT	HMM	FGT	DGT	HMM
AFR	∗	AFR	0.095	0.154	0.294	0.161	0.274	0.471
AFR		AMR	0.088	0.144	0.160	0.147	0.248	0.237
AFR		EAS	0.012	0.032	0.005	0.012	0.038	0.008
AFR		EUR	0.077	0.137	0.102	0.114	0.217	0.128
AFR		SAS	0.021	0.052	0.017	0.024	0.071	0.023
AMR	∗	AMR	0.204	0.356	0.713	0.311	0.572	1.089
AMR		EAS	0.103	0.197	0.294	0.156	0.315	0.429
AMR		EUR	0.197	0.337	0.553	0.304	0.548	0.831
AMR		SAS	0.147	0.261	0.344	0.230	0.425	0.510
EAS	∗	EAS	0.181	0.337	1.047	0.269	0.534	1.659
EAS		EUR	0.085	0.171	0.227	0.120	0.260	0.295
EAS		SAS	0.115	0.224	0.367	0.168	0.353	0.524
EUR	∗	EUR	0.221	0.380	0.722	0.346	0.619	1.132
EUR		SAS	0.164	0.288	0.420	0.254	0.468	0.619
SAS	∗	SAS	0.172	0.312	0.551	0.265	0.513	0.834

It should be noted that different values of N_e would apply to the different populations. The results shown here were obtained from the analysis on the full 1000G sample, where $N_e = 10{,}000$ was used as model parameter in the HMM. The effect of using different values of N_e, which is one of the parameters in the computation of transition probabilities, would need further evaluation. Also, recall that the emission probabilities used here were derived from data simulated using $N_e = 7{,}300$, which may further impact the accuracy of inference. Such possible variations, however, were treated as negligible here, as it appeared more relevant to first demonstrate the general feasibility of the proposed method.

4.4.6 Discussion

The analysis on simulated data has shown that the HMM-based approach to infer IBD around target sites is able to operate equally well in both absence and presence of error. In particular, I showed that IBD segments detected using the HMM maintained high levels of accuracy when genotype error was present. However, a notable caveat is seen in the decreasing accuracy towards higher frequencies of the focal allele.

A possible explanation could be that the emission model was generated under the assumption of recent IBD, where the empirical distributions closely followed the expected genotype pair frequencies where it was assumed that no further mutations occur on a

co-inherited haplotype. For example, as given in Equation (4.9) on page 144, the expected probability to observe genotype pair g_{02} or g_{20} in *ibd* would be equal to zero. While this assumption may hold true for very recent co-inheritance, and perhaps under the convenient conditions of simulation, it is easily violated in reality. The empirical model provided a more realistic approximation to observing genotype pairs in real data (as well as simulated data), but was still limited to observations made under recent co-inheritance.

To have a more complete picture of the differences in the probability of observing each possible genotype pair, I again used the simulation records to sample IBD segments that were co-inherited at varying points in time. The resulting empirical probability distributions are dependent on the allele frequency at the site of a given genotype pair and the T_{MRCA} of the underlying shared haplotype, which is shown in Figure 4.22 (next page), both before and after error. For example, at $T_{MRCA} \leq 1$, the rate at which genotype pair g_{02} or g_{20} was observed was zero throughout before error, but non-zero after error. But, notably, differences due to error were subdued at older relationships.

One caveat of the HMM-based method is its reliance on genotype data. It is possible that there may be several, similar recently co-inherited shared haplotypes involved, whose combined effects on the observed genotype frequency distribution is not straightforward to distinguish based on observations from genotype data alone. The analysis of the simulated dataset and its tree structure has suggested that there was little overlap of recently co-inherited IBD segments on average for a pair of diploid individuals. However, this observation cannot be generalised as it would be expected that the underlying shared haplotype structure is affected by population stratification and other demographic parameters such as inbreeding and geographic isolation of a population.

Another consideration is that focal allele frequency may not be an ideal indicator for allele age or, in particular, T_{MRCA}. The HMM-based method uses the expected age of the target allele observed at a given frequency to modulate the transition probability from the focal *ibd* state to the *non* state. This approximation could be regarded as being unsuitable, because the expected length of an IBD segment is dependent in T_{MRCA} and where the actual age of an allele within a given segment may only be informative if that allele derived from a mutation event around the time of the MRCA of the two haplotypes involved. For example, it would be expected that the majority of shared alleles within the interval of a recent and relatively long segment are (much) older than the time since co-inheritance of that segment. However, recall that I attempted to minimise such confounders in the current analysis by removing "duplicate" segments, where only the segment found around the target allele with the lowest frequency was retained per pair.

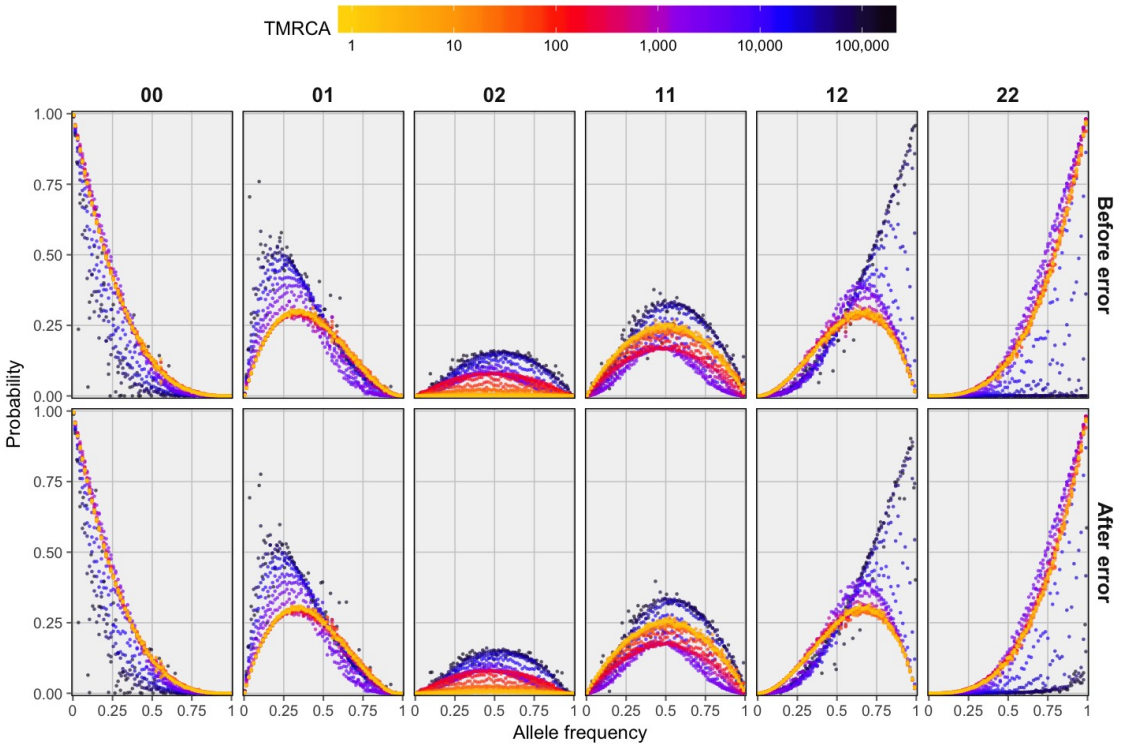


Figure 4.22: Empirical emission probabilities of genotype pairs observed at different T_{MRCA}. The relative proportions of genotype pairs observed in IBD segments are distinguished by time to the most recent common ancestor (T_{MRCA}) for a given pair of haplotypes, which was derived from simulation records. The datasets used to obtain the empirical distributions before and after error were the original, error-free and the error-treated dataset, respectively; see Section 4.3.1 (page 124). IBD segments were sampled at 50 time intervals equally spaced on log-scale. Observation rates of genotype pairs were calculated per allele frequency bin, defined on equal steps of 1%.

Lastly, it must be noted that the results obtained from the analysis of 1000G data are likely to be confounded due to false allele sharing. In presence of data error, not all observed shared alleles correctly identify a recently co-inherited shared haplotype, which is particularly problematic towards lower allele frequencies; for example, see Figure 4.6 (page 127). Future analyses may therefore consider to only analyse focal alleles that were called or typed with high confidence.

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

In contrast to existing methods for allele age estimation (*e.g.*, see review by Slatkin and Rannala, 2000), the method I present in this chapter does not require prior knowledge about past demographic processes or events. Although an assumption of certain population parameters is required, such as effective population size (N_e), as well as mutation and recombination rates, these are expected to affect the scaling of time, such that differences between age estimates for different alleles are expected to be proportionally constant.

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

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

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

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

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 all copies of the allele were co-inherited from that one chromosome in which the mutation occurred at some point in the past.

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

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

There are two main sources of information available from sample data which relate to the T_{MRCA}. First, mutation events occur independently in each lineage and mutations accumulate along the sequence as the haplotype is passed on over generations. Second, recombination events break down the length of the haplotype in each generation independently in each lineage. Thus, the T_{MRCA} between a given pair of chromosomes can

be estimated from the number of mutations which segregate in two haplotypes, as well as the genetic length of the haplotype region that is shared between two chromosomes in the sample. In the following (Section 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. 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 find a common ancestor at exactly t generations in the past is

$$f(t) \approx \frac{1}{2N_e} \mathrm{e}^{-\frac{t}{2N_e}} \tag{5.1}$$

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

$$\pi(\tau) \propto \mathrm{e}^{-\tau}. \tag{5.2}$$

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

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

$$f_S(s) \;=\; P(S = s \mid \theta, h, \tau) \;=\; \frac{(\theta h \tau)^s}{s!}\, \mathrm{e}^{-\theta h \tau}. \tag{5.3}$$

The likelihood function for the time parameter τ is proportional to Equation (5.3), but requires only those terms that involve τ and where constant terms can be dropped, such that

$$\mathcal{L}(\tau \mid \theta, h, s) \;\propto\; \tau^s\, \mathrm{e}^{-\theta h \tau}\,. \tag{5.4}$$

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

$$\begin{aligned} p(\tau \mid \theta, h, s) \;&\propto\; \mathcal{L}(\tau \mid \theta, h, s) \;\times\; \pi(\tau) \\ &\propto\; \tau^s\, \mathrm{e}^{-\tau(\theta h + 1)} \end{aligned} \tag{5.5}$$

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

In the above, the density of the posterior probability is specified up to a missing normalising constant. Note that Equation (5.5) is proportional to (has the form of) the Gamma probability density function (PDF), namely

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

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

$$p(\tau \mid \theta, h, s) \;=\; g(\tau \mid s + 1, \theta h + 1). \tag{5.6}$$

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

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

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

$$P(D = d \mid \psi, \tau) \;=\; \psi\tau \; \mathrm{e}^{-\psi\tau d}. \tag{5.7}$$

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

$$P(D > d \mid \psi, \tau) \;=\; \mathrm{e}^{-\psi\tau d}. \tag{5.8}$$

Equations (5.7) and (5.8) above can be simplified to

$$f_D(d) \;=\; (\psi\tau)^b \; \mathrm{e}^{-\psi\tau d} \tag{5.9}$$

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

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

Considering Equation (5.9), the likelihood function for τ can now be written as

$$\mathcal{L}(\tau \mid \psi, d, b) \;\propto\; \tau^b \; \mathrm{e}^{-\psi d\tau} \tag{5.10}$$

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

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

Hence, the likelihood function in support of τ is

$$\mathcal{L}(\tau \mid \psi, h, b_L, b_R) \propto \tau^{b_L+b_R} \, \mathrm{e}^{-\psi h \tau} \tag{5.11}$$

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

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

The posterior probability is obtained as

$$\begin{aligned} p(\tau \mid \psi, h, b_L, b_R) &\propto \mathcal{L}(\tau \mid \psi, h, b_L, b_R) \times \pi(\tau) \\ &\propto \tau^{b_L+b_R} \, \mathrm{e}^{-\tau(\psi h+1)}. \end{aligned} \tag{5.12}$$

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

$$p(\tau \mid \psi, h, b_L, b_R) = g(\tau \mid b_L + b_R + 1, \psi h + 1). \tag{5.13}$$

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

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 (5.3) and (5.9) on page 167 and page 168 and is given as

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

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

$$\mathcal{L}(\tau \mid h_p, h_g, s, b_L, b_R) \propto \tau^{s+b_L+b_R} \, \mathrm{e}^{-\tau(h_p+h_g)} \tag{5.14}$$

from which the posterior probability is obtained as

$$\begin{aligned} p(\tau \mid h_p, h_g, s, b_L, b_R) &\propto \mathcal{L}(\tau \mid h_p, h_g, s, b_L, b_R) \times \pi(\tau) \\ &\propto \tau^{s+b_L+b_R} \, \mathrm{e}^{-\tau(h_p+h_g+1)}. \end{aligned} \tag{5.15}$$

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

$$p(\tau \mid h_p, h_g, s, b_L, b_R) \;=\; g(\tau \mid s + b_L + b_R + 1, h_p + h_g + 1). \tag{5.16}$$

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 inherited the allele (*carriers*) are distinguished from the haplotypes that do not carry the allele (*non-carriers*). Thus, the sample is divided into two disjoint subsamples; let X_c denote the set of chromosomes which share a given allele, and X_d the set of chromosomes which do not carry that allele.

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

In the following, I describe how the posterior density of the $\mathrm{T_{MRCA}}$ is obtained for concordant and discordant pairs to eventually arrive at an estimate of allele age. To distinguish the population-scaled time τ, as defined for the $\mathrm{T_{MRCA}}$, from the time of the mutation event, let the latter be denoted by the likewise population-scaled time t. Informally, the actual time of a mutation event is found at the "sweet spot" in between the earlier coalescent event at time t_c and the later coalescent event at time t_d; see Figure 5.1.

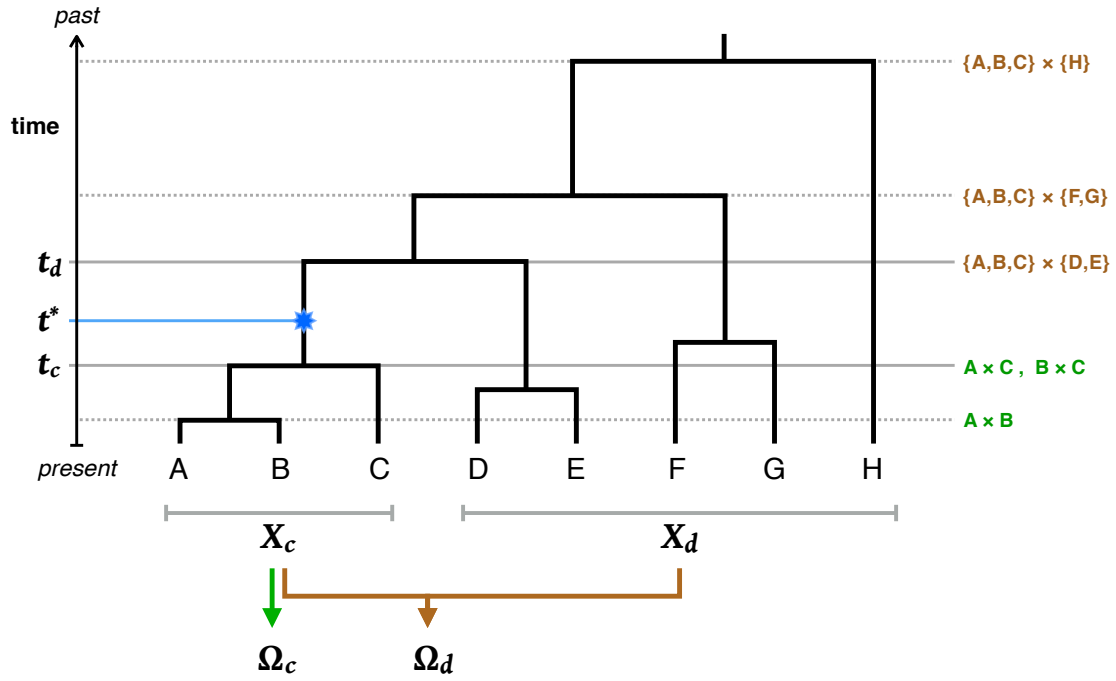


Figure 5.1: Allele age in relation to concordant and discordant pairs. The genealogy of a sample of eight haplotypes is shown of which A, B, and C share a focal allele that derived from a mutation event as indicated in the tree (*star*). These chromosomes constitute the set of *carriers*, denoted by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

5.2.2.1 Cumulative coalescent function (CCF)

At a given focal site at which the possible concordant and discordant pairs in the sample have been sorted into the sets Ω_c and Ω_d, respectively, each pair is analysed in turn to obtain a posterior on their T_{MRCA}. Importantly, to 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

$$\Phi_{ij}(t) = \begin{cases} P(\tau \leq t) & \text{if } \{i,j\} \subseteq \Omega_c \quad (\textit{i.e.} \text{ concordant pairs}) \\ P(\tau > t) = 1 - P(\tau \leq t) & \text{if } \{i,j\} \subseteq \Omega_d \quad (\textit{i.e.} \text{ discordant pairs}). \end{cases} \tag{5.17}$$

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

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

$$G(t) \;=\; P(\tau \leq t \mid \alpha, \beta) \;=\; \int_0^t g(u \mid \alpha, \beta)\, du \tag{5.18}$$

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

$$F(t) \;=\; P(\tau \leq t \mid \alpha, \beta) \;=\; 1 - \mathrm{e}^{-\beta t} \sum_{i=0}^{\alpha-1} \frac{(\beta t)^i}{i!}. \tag{5.19}$$

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

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.

$$\Lambda_k(t) \;\propto \prod_{i,j \in A_k} \Phi_{ij}(t \mid \alpha, \beta) \tag{5.20}$$

for focal site k at which haplotype pairs have been sorted into the collection $A_k = \{\Omega_c, \Omega_d\}$, according to allele sharing at that site. Again, α, β are defined by the clock model used and obtained from parameter values observed for pair i, j. In the following, the term *composite posterior* is used to refer to the result obtained using Equation (5.20).

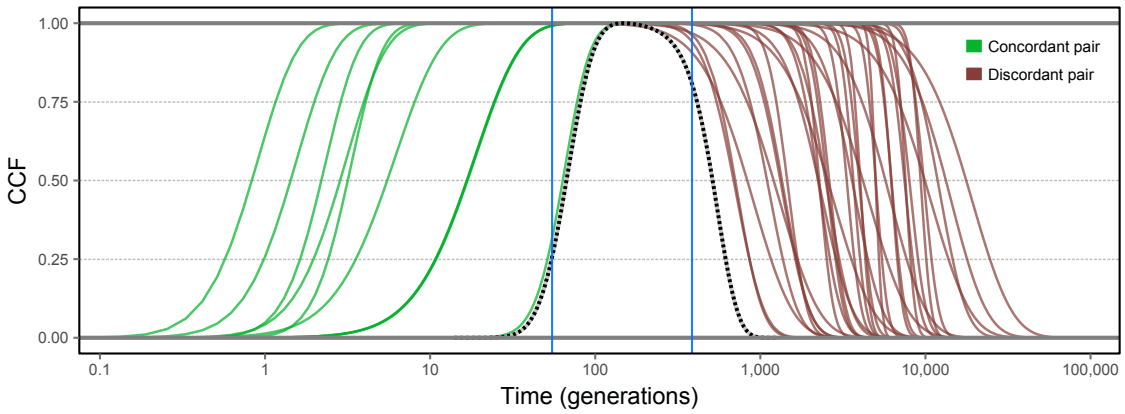


Figure 5.2: Example of concordant and discordant posterior distributions and the resulting composite posterior. A target variant was randomly selected from simulated data. The CCF was obtained for the set of possible concordant pairs and a random subset of discordant pairs. The thicker *dotted* line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

The composite posterior distribution can now be obtained over $t \in (0, \infty)$. However, in practise, it is unlikely that the relationship of i, j can be traced back further than a small multiple of N_e (*e.g.* ~ 10). An example is given in Figure 5.2 (this page), showing the CCF for concordant and discordant pairs, as well as the maximised composite posterior distribution. Notably, the "width" of the distribution is expected to be determined by the underlying branch length. In the following, the mode of the composite posterior distribution is taken as a point estimate for the age of an allele, denoted by $\hat{t}$.

5.2.2.3 Note on composite likelihood methods

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)

$$\mathcal{CL}(\vartheta \mid y) = \prod_{z \in Z} \mathcal{L}_z(\vartheta \mid y)^{w_z} \tag{5.21}$$

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

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

$$p_{\mathcal{CL}}(\vartheta \mid y) \propto \pi(\vartheta) \times \mathcal{CL}(\vartheta \mid y) \tag{5.22}$$

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

The "composite posterior" given in Equation (5.20) 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, yet without specifying an appropriate weighting function.

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

$$\max[n_c] = \binom{k}{2} = \frac{k(k-1)}{2} \tag{5.23}$$

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

$$\max[n_d] = k(2N - k) \tag{5.24}$$

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

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

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 (both "true" and phased haplotype data). Here, data *before* and *after* the inclusion of error are distinguished by referring to dataset $\mathcal{D}_B$ and dataset $\mathcal{D}_B^*$, respectively. Note that in the following the term *genotype error* is used, even in analyses that operate on haplotype data, as error proportions were estimated from misclassified genotypes in 1000G data ("1000G.A" in Chapter 4, Section 4.2.2, page 116).

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

$$t_m = \sqrt{t_c \, t_d} \, . \tag{5.25}$$

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

Accuracy was measured using Spearman's rank correlation coefficient, r_S, which is a robust measure for the strength of the monotonic relationship between two variables; *i.e.* the inferred allele age ($\hat{t}$) and true time proxies (t_c, t_m, or t_d). Note that the squared Pearson correlation coefficient, r^2, was used in previous chapters but was regarded as

being less suitable here, as both the inferred and true age are expected to vary on log-scale, and the Pearson coefficient measures the linear relationship between variables. However, for example, r^2 of log-transformed age could have been used, but which was not additionally considered here, in order to keep the analysis brief. Also, to indicate bias, the root mean squared logarithmic error (RMSLE) was calculated as a descriptive score for the magnitude of error (here defined on $\log_{10}$).

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

$$\hat{t}_{rel} = \frac{\log\left[\frac{\hat{t}}{t_c}\right]}{\log\left[\frac{t_d}{t_c}\right]} \tag{5.26}$$

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

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

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. Model parameters in `rvage` were specified according to parameters used for the simulation of $\mathcal{D}_A$ ($N_e = 10{,}000$; $\mu = 1 \times 10^{-8}$; $\rho = 1 \times 10^{-8}$).

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.

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 %).

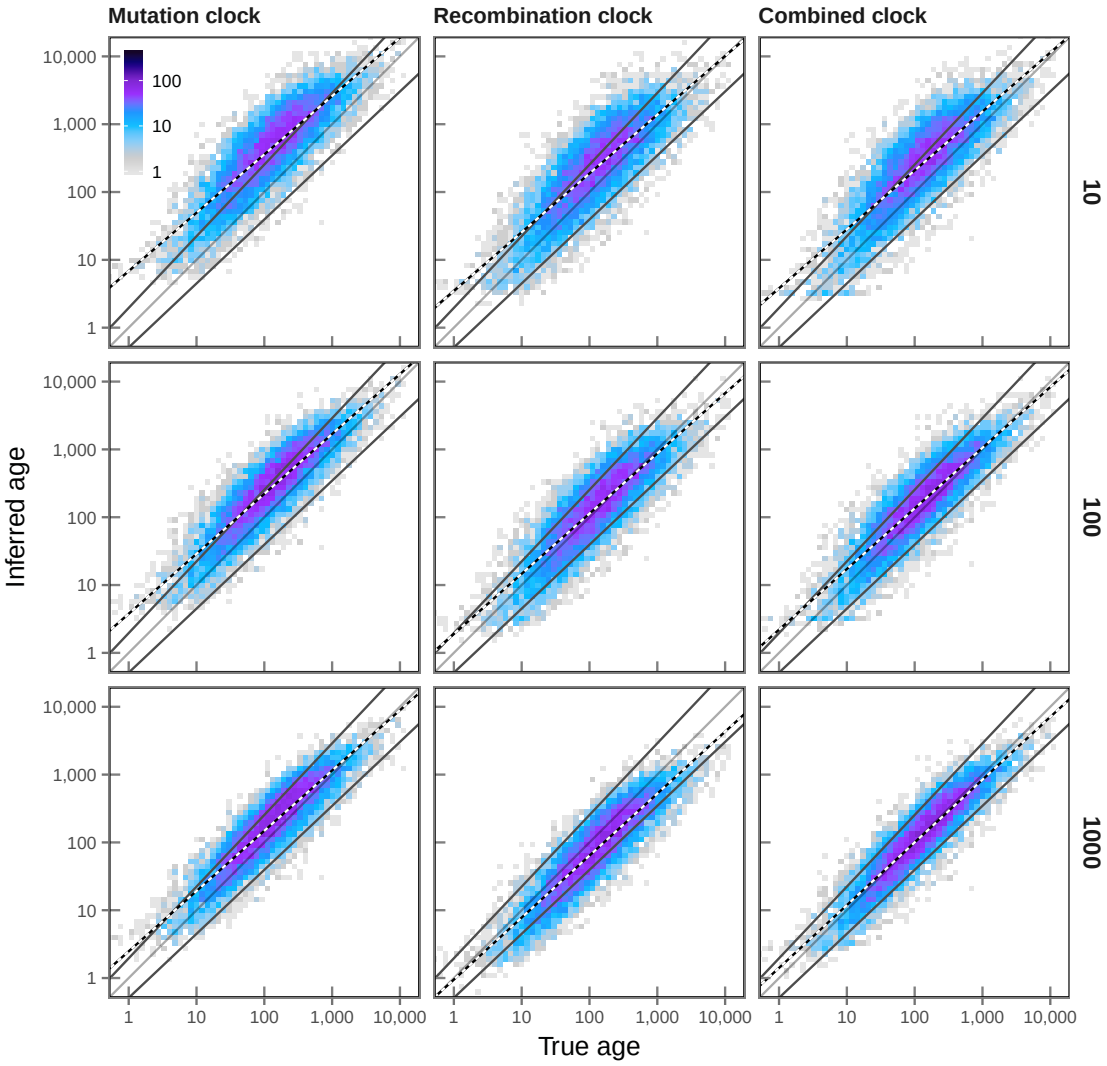


Figure 5.3: True and inferred age under varying numbers of discordant pairs. A set of 10,000 target sites was randomly drawn in $f_{[2,20]}$ (shared allele frequency ≤ 1%) in a simulated sample of 2,000 haplotypes. Different numbers of sampled discordant pairs were analysed on the same set of target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

Table 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 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.

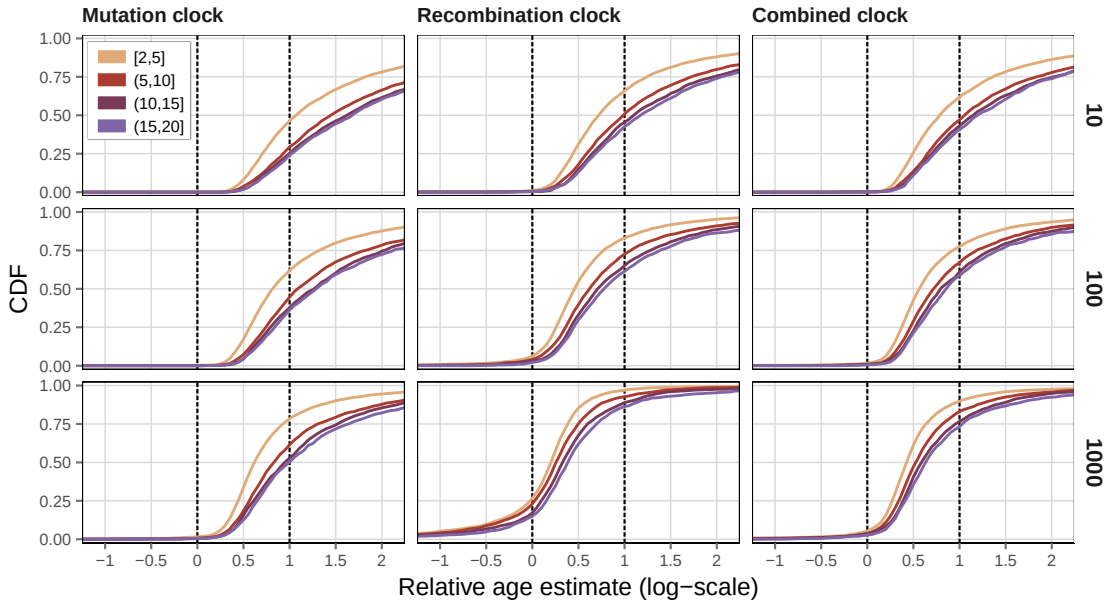


Figure 5.4: Relative age under varying numbers of discordant pairs. A randomly drawn set of 10,000 target sites at allele frequency ≤ 1%, *i.e.* $f_{[2,20]}$, was analysed under each of the three clock models (indicated at the *top* of each column) and with different numbers of sampled discordant pairs; $n_d = 10$, $n_d = 100$, 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 (5.26), such that the true times of concordant and discordant coalescent events, t_c and t_d, sit at 0 and 1, respectively (*dashed* lines). Note that $\hat{t}_{rel}$ is defined on log-scale. The CDF of relative age estimates is shown per f_k group, where target variants were pooled by their allele count in the data, in ranges of $f_{[2,5]}$, $f_{(5,10]}$, $f_{(10,15]}$, and $f_{(15,20]}$.

A comparison of the inferred age distributions at distinct f_k ranges is presented in Figure 5.4 (this page), again shown for $n_d = 10$, $n_d = 100$, and $n_d = 1{,}000$. Notably, the accuracy of target alleles at lower frequencies was overall higher compared to alleles observed at higher frequencies. This difference was consistent across n_d thresholds under the mutation clock model, $\mathcal{T}_{\mathcal{M}}$. For example, at $n_d = 10$, the proportion of correctly dated alleles was higher in the $f_{[2,5]}$ range (48.356 %) compared to alleles at $f_{(5,10]}$ (29.445 %). At $n_d = 1{,}000$, overall accuracy was increased but the difference for alleles at lower and higher frequencies remained; *i.e.* 77.819 % and 60.834 % at $f_{[2,5]}$ and $f_{(5,10]}$, respectively. Under the recombination clock model, $\mathcal{T}_{\mathcal{R}}$, these differences were reduced at higher n_d thresholds. At $n_d = 10$, 66.608 % and 50.344 % of alleles were correctly dated at $f_{[2,5]}$ and $f_{(5,10]}$, respectively, whereas at $n_d = 1{,}000$ these proportions were 72.258 % and 69.826 % at the same frequency ranges, respectively.

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 random sampling would strike 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, 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 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 $\mathrm{f}_c < \mathrm{f}_k$ and $\mathrm{f}_c < \mathrm{f}_s$; additionally $\mathrm{f}_s > 1$ to exclude singletons and $(\mathrm{f}_s - \mathrm{f}_c) > (2N - \mathrm{f}_k)$ to exclude sites that are monomorphic within X_d, where $2N$ refers to the number of haplotypes in the sample. The condition implies

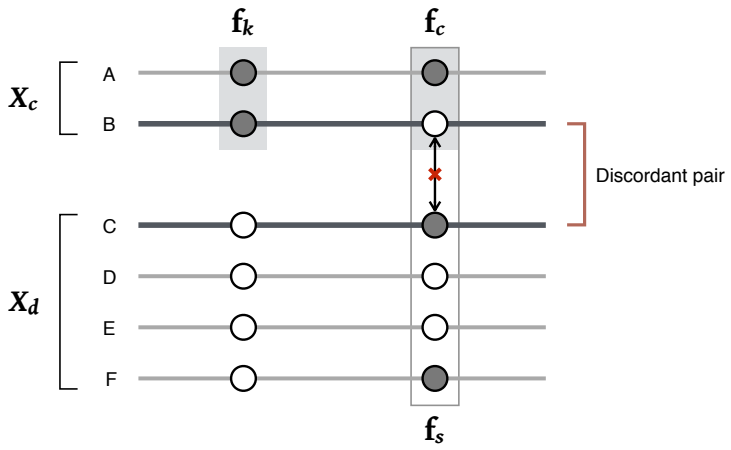


Figure 5.5: Breakpoint detection in discordant pairs. A discordant pair is formed by one haplotype from X_c (which share the focal allele) and one haplotype from X_d (which do not share the focal allele). The lines indicate the chromosomal sequence where the alleles at two sites are indicated; allelic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

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.

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

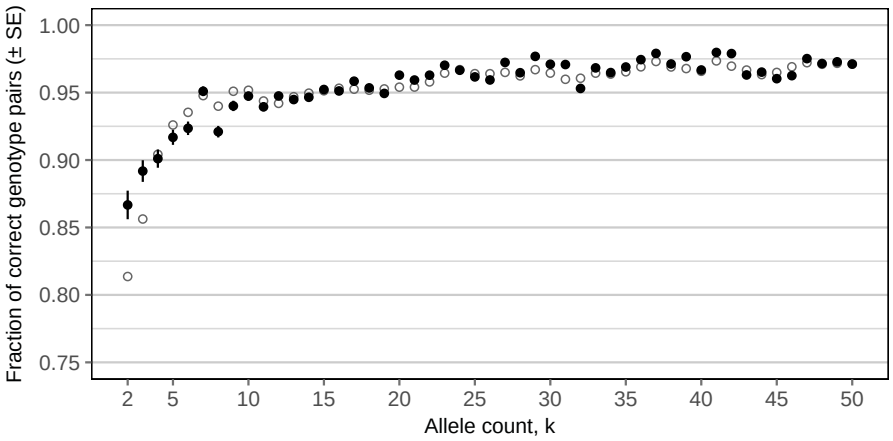


Figure 5.6: Initial state probability of discordant pairs in the genotype-based HMM. The proportion of discordant pairs that were correctly identified by their genotypes was empirically determined from data before and after the inclusion of realistic genotype error rates. The mean per f_k was used as 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).

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

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

Note that both the DGT and the HMM-based approach may operate on genotype data alone. Importantly, if haplotype information is not available, the sets X_c and X_d are formed by assigning all individuals that are heterozygous to X_c while all others are assigned to X_d, but excluding individuals that are homozygous for the focal allele. Since haplotype data

are required to determine pairwise differences along haplotype sequences, $\mathcal{T}_{\mathcal{M}}$ and $\mathcal{T}_{\mathcal{MR}}$ cannot be used with genotype data. Here, analyses using the DGT and the HMM-based approach were performed on haplotypes although genotype data alone would suffice.

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, 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 (page 188). 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 188) 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 $\mathrm{f}_{[2,5]}$ 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 $\mathrm{f}_{[2,5]}$ 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

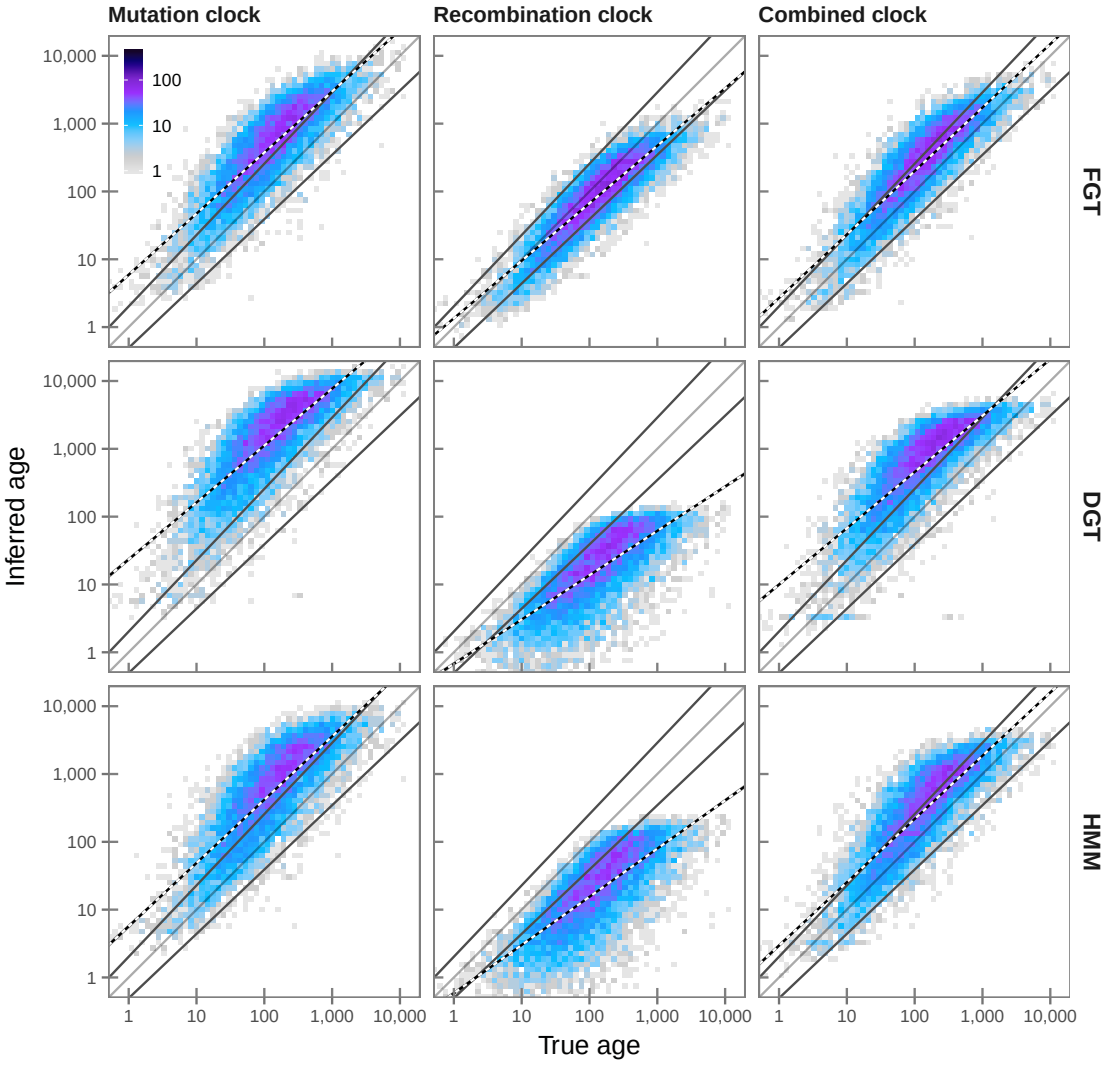


Figure 5.7: Distribution of true and inferred age using different IBD detection methods. The three IBD detection methods FGT, DGT, and HMM were compared under each clock model, on the same set of target sites that were drawn from $f_{[2,20]}$ variants (allele frequency $\leq 1\%$) in $\mathcal{D}_A$. 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.

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

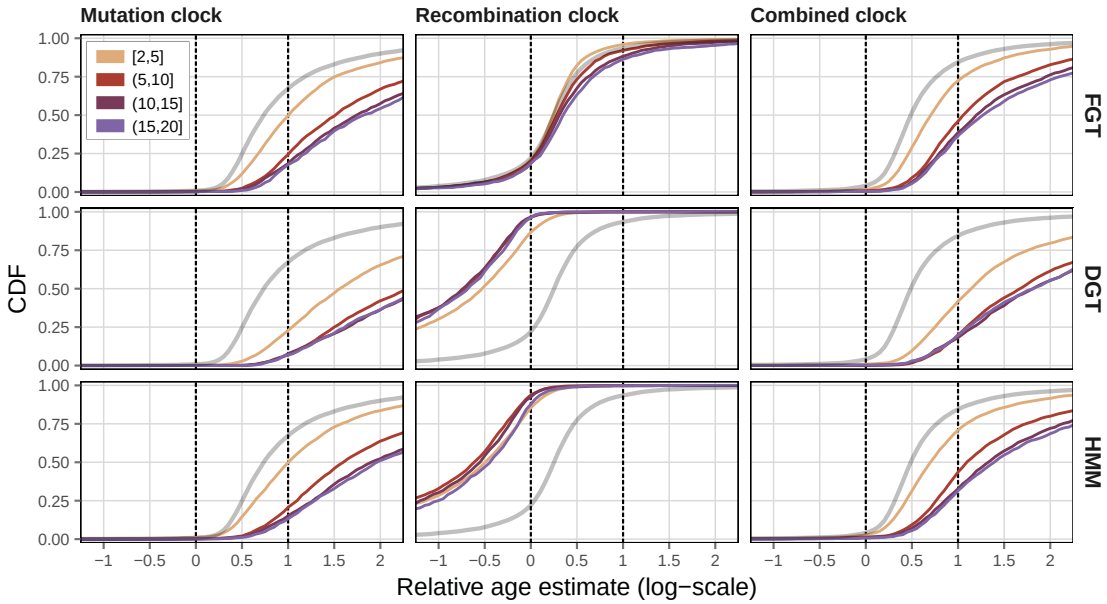


Figure 5.8: Relative age using different IBD detection methods. The three IBD detection methods implemented in `rvage` were compared, *i.e.* FGT, DGT, and HMM (indicated at the *right* of each row), under each clock model (indicated at the *top* of each column). The relative age, $\hat{t}_{rel}$, was calculated as given in Equation (5.26), such that t_c and t_d sit at 0 and 1 (*dashed* lines). The CDF of relative age estimates is shown for different frequency ranges; namely $f_{[2,5]}$, $f_{(5,10]}$, $f_{(10,15]}$, and $f_{(15,20]}$. The *grey* line provides a comparison to age estimated using true IBD information as shown in Figure 5.4 (page 181), but for $f_{[2,20]}$.

$\mathcal{T}_{\mathcal{M}}$ 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.

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

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 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.

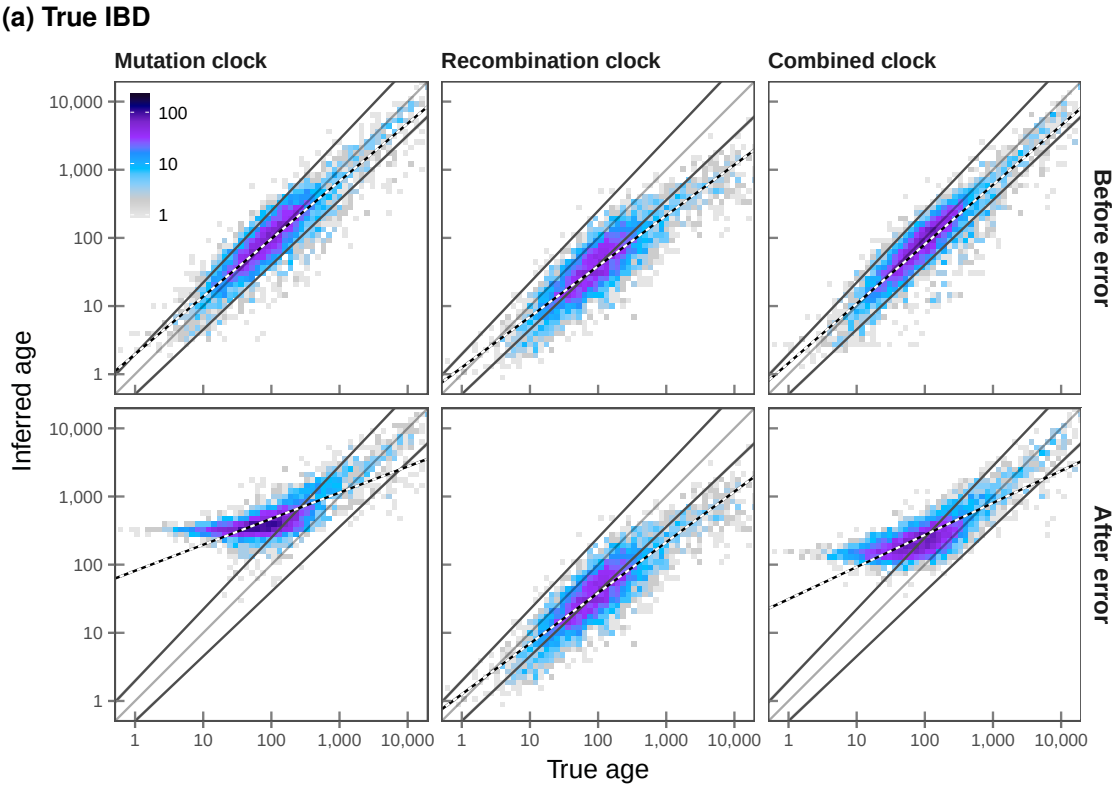


Figure 5.9: Density of allele age before and after error in simulated data. The effects on the estimation process *before* and *after* error are compared. Note that the "true age" was set to t_m, which is the geometric mean of t_c and t_d. Lines *below* and *above* the dividing line correspond to the 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 $f_{[2,25]}$.

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 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 (page 194).

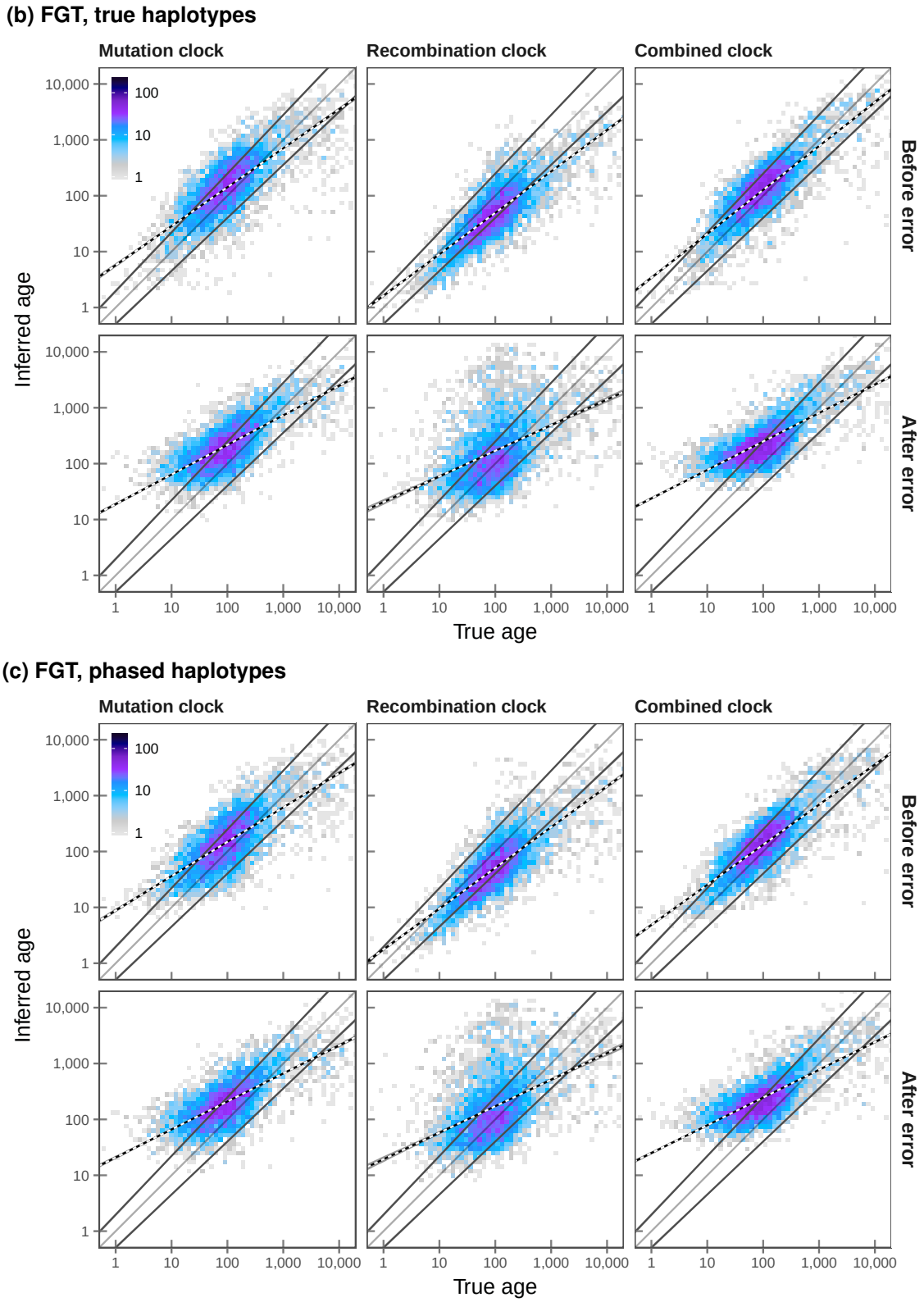


Figure 5.9: Continued.

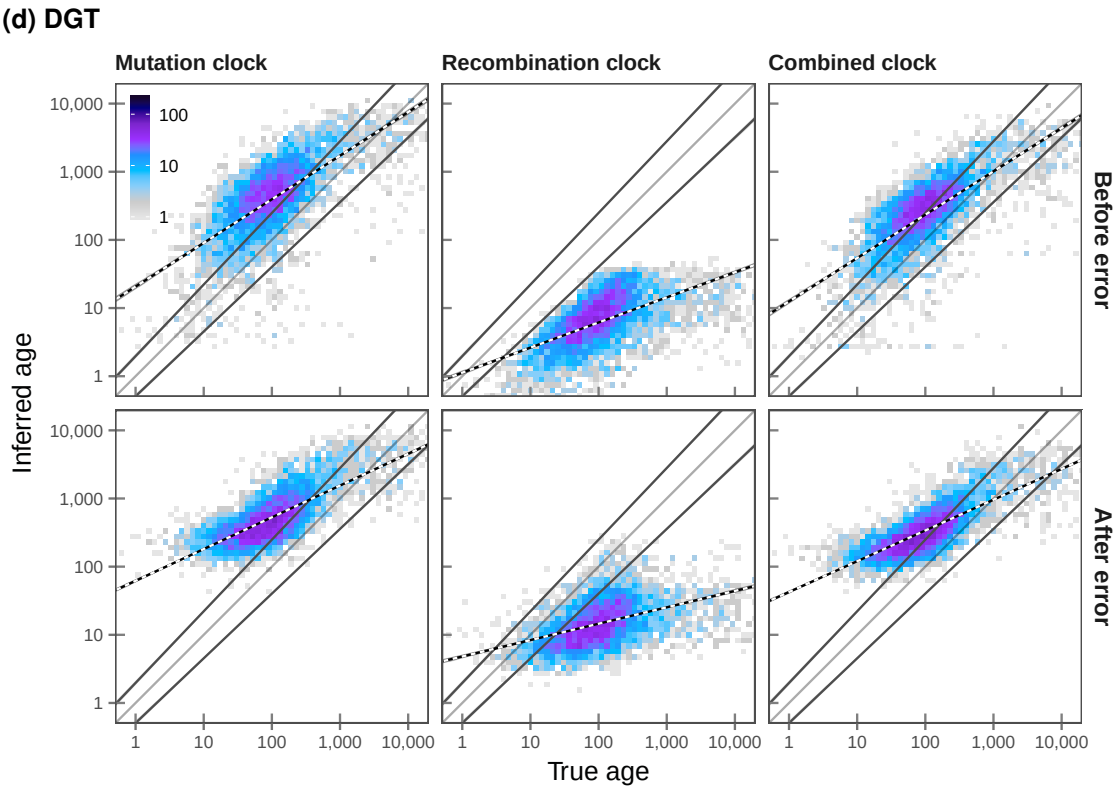


Figure 5.9: Continued.

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 $\mathcal{T}_{\mathcal{MR}}$. Although rank correlation for the HMM was high in $\mathcal{T}_{\mathcal{R}}$, *e.g.* r_S at t_c was 0.751 before and 0.737 after error, estimation bias was substantial both before and after error.

Table 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.

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

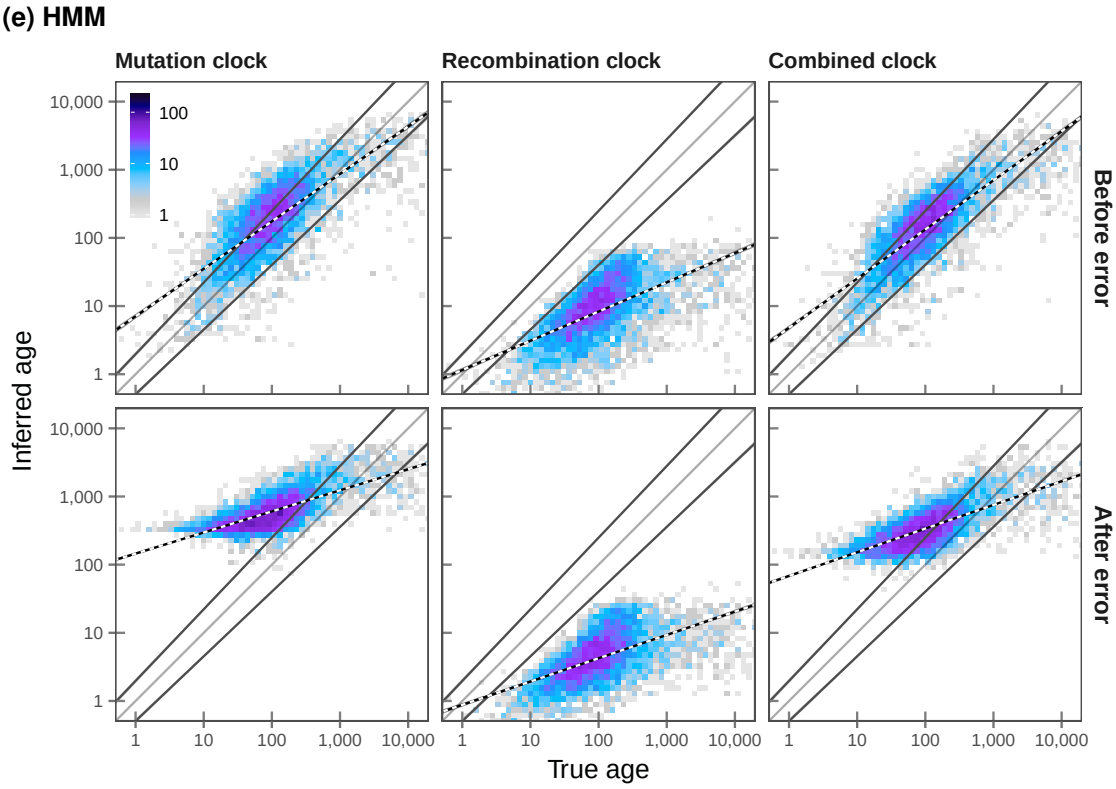


Figure 5.9: Continued.

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

5.6 A haplotype-based HMM for shared haplotype inference

As shown in the previous sections, the allele age estimation method was able to infer the time of mutation events with high accuracy, but where it was suggested that this is largely dependent on (ideally) correct inference of the underlying shared haplotype structure around a given target site. The inference methods presented so far were either rule-based (FGT and DGT) and thus highly susceptible to data error, or probabilistic but genotype-based (HMM) where accuracy may easily be influenced by the shared haplotype structure at the "unshared" haplotypes in the individuals considered. Although the latter was implicitly impervious to data error due to phasing, it was shown that such errors were less problematic on average, but where error due to misclassification of alleles may be regarded as the main source of estimation bias.

In this section, I present a novel haplotype-based HMM for the targeted and pairwise detection of shared haplotypes. This method can be seen as the conclusion of the previous genotype-based approach and is therefore algorithmically defined in a similar way. I describe the model in the section below. Note that I implemented additional modifications in the age estimation method, which I describe in Section 5.6.2 (page 200).

As done previously, I evaluated the haplotype-based HMM in terms of the resulting age estimation accuracy in data before and after error (Section 5.6.3, page 203). The method was further compared to the Pairwise Sequentially Markovian Coalescent (PSMC) in terms of the inferred T_{MRCA} and subsequent estimation of allele age (Section 5.6.4, page 209). Lastly, I briefly present age estimation results from an analysis using 1000 Genomes Project (1000G) Phase III data (Section 5.6.5, page 215).

5.6.1 Description of the model

The general algorithm and model specifications of the HMM presented here follow the previously described genotype-based HMM; see Section 4.4.1 (page 137). Briefly, at a given target position in the genome, the allelic sequence of two selected haplotypes is

independently traversed to the left and right-hand side from that position, until the end of the chromosome is reached. As before, two hidden states are assumed; *ibd* and *non*. The observation states are defined as the possible allelic pairs $(0,0)$, $(0,1)$, and $(1,1)$, where the order of alleles is not relevant; *i.e.* observing $(0,1)$ in haplotypes A and B is equivalent to observing $(1,0)$. The model is illustrated in Figure 5.10 (this page), where φ denotes a transition parameter. The probabilities of emission from *ibd* are denoted by δ_{ab} and from *non* by η_{ab}, where $a, b \in \{0,1\}$ refer to the allelic states. Model transitions, emissions, and initial state probabilities are described below.

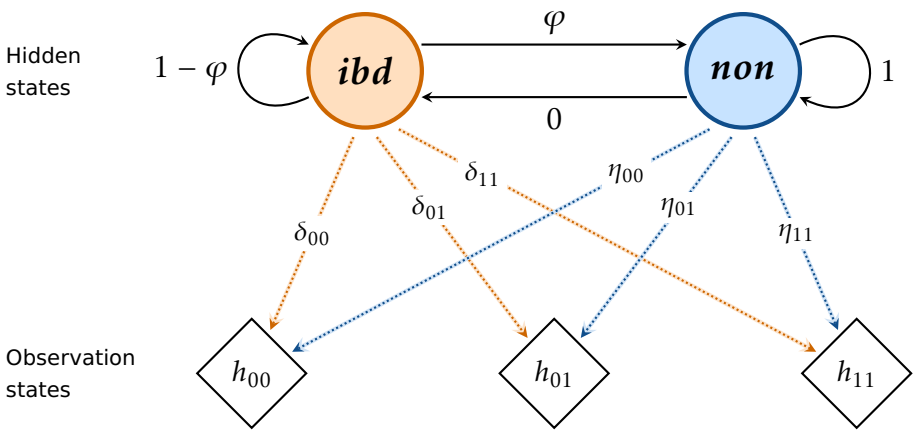


Figure 5.10: Illustration of the haplotype-based HMM for shared haplotype inference. Two hidden states are assumed to generate the observations in a Markov process; *ibd* and *non*. Transitions from each state into any state are indicated by *solid* lines. The probability of transition from *ibd* to *non* is denoted by φ, and from *non* to *ibd* is set to zero; hence, once the chain proceeds into the *non* state it cannot go back into *ibd*. This is because the process is modelled such that only the innermost segment is inferred, relative to the focal position which sits at the start of the sequence. The input sequence consists of sequence data from a pair of haplotypes, resulting in three possible observation states; denoted by h_{ab}, where $a, b \in \{0,1\}$. The probabilities of emitting each possible haplotype pair given each hidden state are denoted by δ_{ab} and η_{ab} for *ibd* and *non*, respectively; indicated by the *dotted* lines. The direction of arrows indicates conditional dependence; *i.e.* the transition from one hidden state into another state, or emission of a genotype pair while being in *ibd* or *non*.

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

$$\varphi \;=\; 1 - \mathrm{e}^{-2N_e r_j \xi_k} \tag{5.27}$$

where j indicates the position at the current site in the sequence and k indicates the focal position. The value of r_j is the genetic distance between the current and the immediately previous site observed in the sequence. The value of ξ_k is the expectation of the age of the

focal allele after Kimura and Ota (1973), calculated using Equation (1.30) on page 35. The probability of remaining in *ibd* is $1 - \varphi$. The model employs a "left-to-right" architecture, where transition from *non* to *ibd* have zero probability; *i.e.* once *ibd* has been left the sequence stays in *non* with probability equal to one. The intuition and shortcomings of such a transition model were discussed in Section 4.4.2.1 (page 139).

Emission probabilities. An empirical emission model was generated from simulated haplotype data, both before and after the integration of empirically determined error rates. In the previous genotype-based model, the empirical rate of observing possible genotype pairs was determined from haplotype segments shared between individuals who carried any allele observed at low frequency ($f_{[2,25]}$). The same was done here, but such that the empirical probabilities were measured based on the T_{MRCA} between haplotype pairs. A full representation of observation rates found by T_{MRCA} and allele frequency is given in Figure 5.11 (next page). Note that, again, datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$ were used to obtain emission models before and after error.

However, the generation of the emission model was not straightforward as any pair of haplotypes is "identical by descent" at any position in the genome; according to the coalescent and the assumptions of the infinite sites model. The definition of the hidden states as being "in IBD" and "not in IBD" may therefore be seen as arbitrary. Nonetheless, to generate empirical probabilities, I defined a nominal "cutoff" for the T_{MRCA} at 100 generations to calculate the mean empirical observation rates for each state. The average rate seen ≤ 100 generations was taken for *ibd*, and > 100 generations for *non*, at each allele frequency bin (rates were averaged over 100 equally sized bins). Linear interpolation was used to obtain rates at sites with frequencies not captured by the model. The resulting model is illustrated in Figure 5.12 (page 200).

One caveat of applying such a cutoff is that the model is implicitly conditioned on observations deriving from relatively recent coalescent events. But, notably, when comparing cutoffs at 100 and 1,000 generations, differences between averaged distributions were relatively small; as shown in Figure 5.12 (page 200). In the following, the emission models generated from dataset $\mathcal{D}_B$ were used in all analyses of error-free data; otherwise, the model generated from $\mathcal{D}_B^*$ was used.

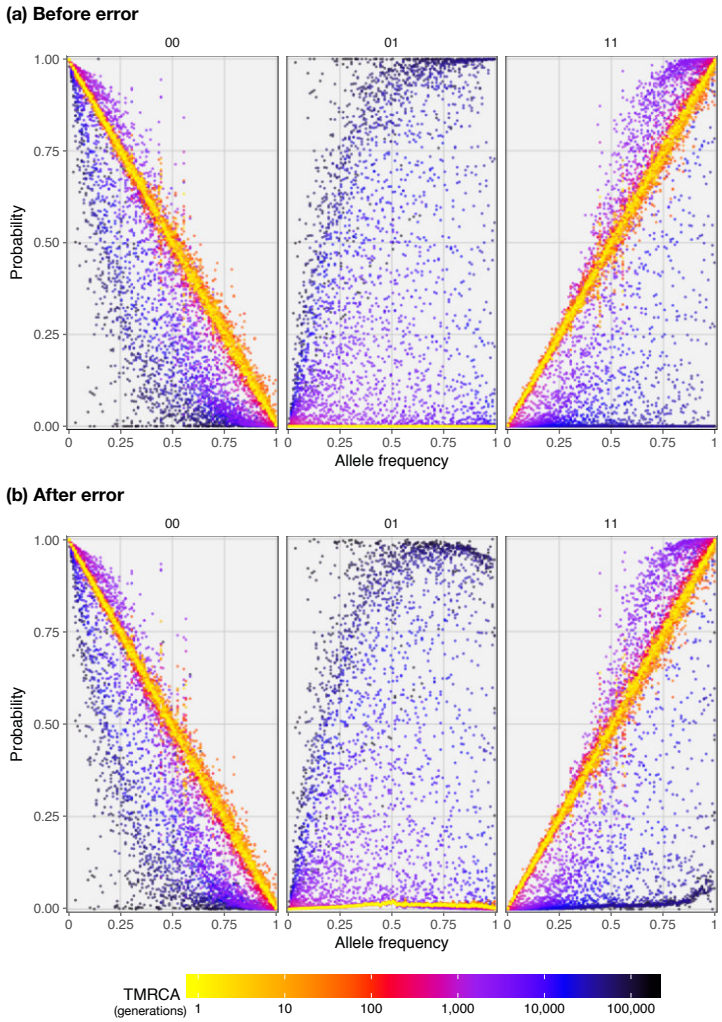


Figure 5.11: Empirical probability to observe allelic pairs dependent on T_{MRCA}. Using data before and after error (datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$), the rate at which the possible allelic pairs $(0,0)$, $(0,1)$, and $(1,1)$ were observed is shown by allele frequency, distinguished by the T_{MRCA} between two haplotypes at a given site. Empirical observation rates were measured at 100,000 randomly selected haplotype pairs. The resulting rates were averaged over 100 equally sized allele frequency bins and 100 time intervals of equal size on log-scale. Panels **(a)** and **(b)** show the empirical distribution before and after error, respectively.

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

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

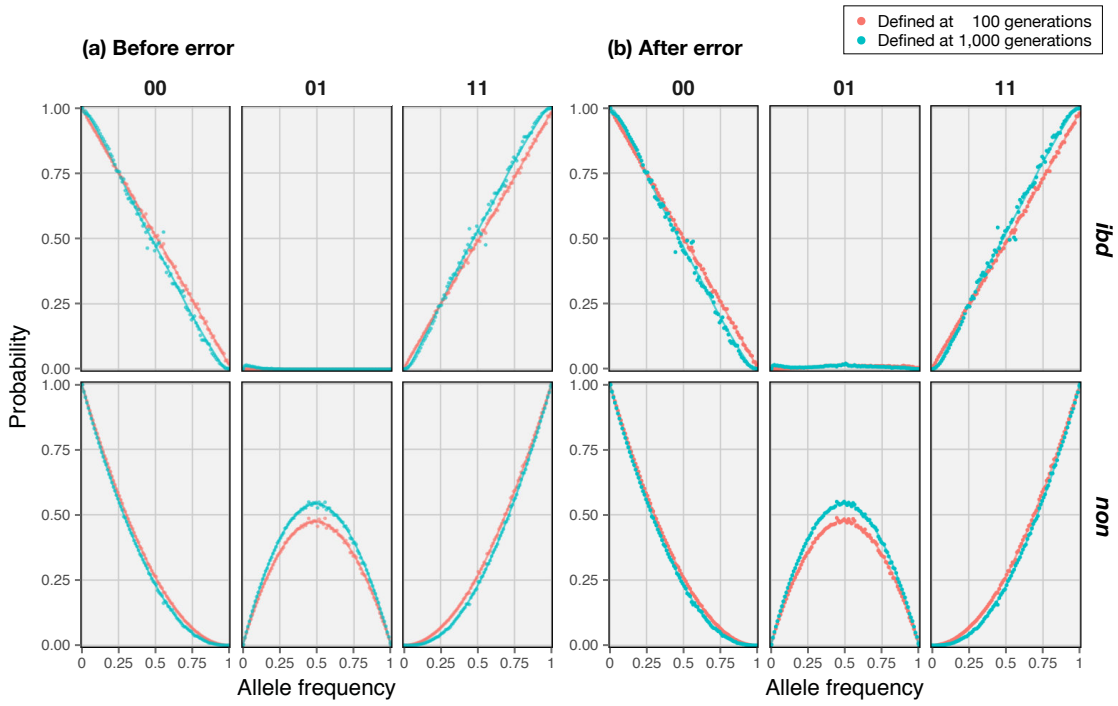


Figure 5.12: Empirical emission model used in the haplotype-based HMM. The emission probabilities used in the haplotype-based HMM were determined empirically for each possible allelic pair, using datasets $\mathcal{D}_B$ (before error) and $\mathcal{D}_B^*$ (after error). To distinguish *ibd* from *non*, a nominal cutoff was applied to the result shown in Figure 5.11 (page 199). For comparison, the resulting distributions for a T_{MRCA} cutoff at 100 generations (*red*) are compared to a cutoff at 1,000 generations (*blue*).

observed at that site and that pair. That is, allelic combinations $(1, 1)$ in concordant pairs and $(0, 1)$ or $(1, 0)$ in discordant pairs. This was done at each value of the allele count in the simulated dataset (N = 5,000 haplotypes), but where a maximum of 1,000 randomly selected sites was analysed from the sites found at the same allele count. Measured rates were then averaged per bin; the result is shown in Figure 5.13 (next page). When applied to data of different size N, rates were linearly interpolated based on allele frequency.

5.6.2 Modifications of the age estimation method

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

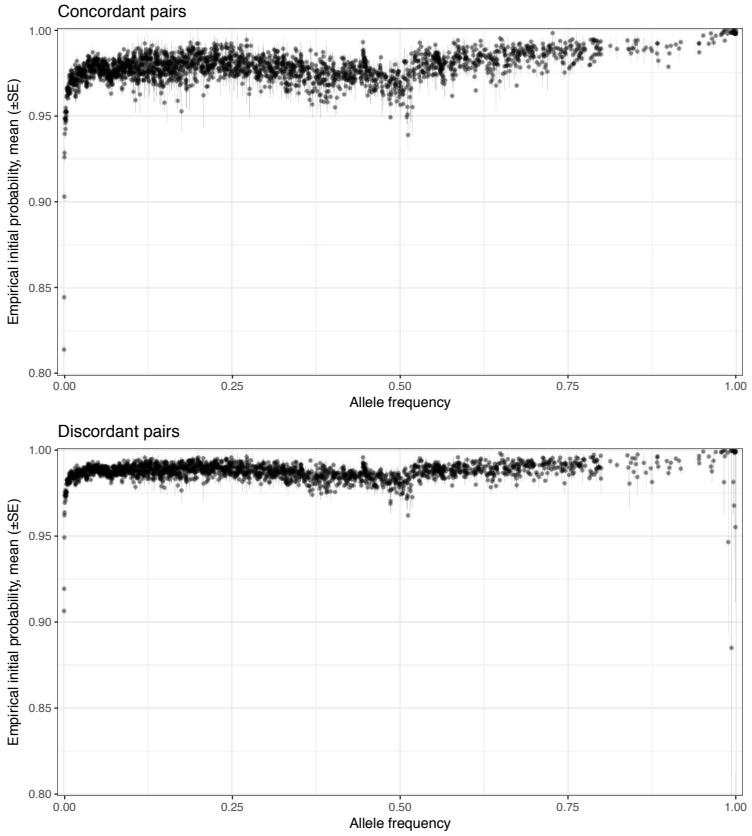


Figure 5.13: Empirical initial state probabilities used in the haplotype-based HMM. The rate of correctly observing allelic combinations in pairs of haplotypes was measured by comparing data points before and after error, using simulated datasets $\mathcal{D}_B$ and $\mathcal{D}_B^*$. At a given focal site, haplotypes were sorted into concordant and discordant pairs as determined from allele sharing in the sample before error. The allelic state at each pair was then compared to the same location and pair in data after error, to measure the true positive rate of observing the $(1, 1)$ allelic combination in concordant pairs and $(0, 1)$ in discordant pairs. Rates are shown as the mean for observations at a given focal allele count (±SE). The number of focal sites observed at a specific frequency in the sample differed along the site frequency spectrum, but was capped at a maximum of 1,000 randomly sampled sites found at a given allele count.

5.6.2.1 Nearest neighbour selection of discordant pairs

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

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

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

The implementation of this algorithm substantially improved computation time due to retaining a smaller number of pairs. The results presented below were obtained using only $n_d = 100$ discordant pairs (as opposed to several hundreds or thousands of randomly selected pairs). Note that a nearest neighbour approach could also be applied to select concordant pairs (prioritising higher distances), but which was not done here, as it is assumed that random sampling is sufficient to retain pairs that indicate coalescent events close to the time of the focal mutation event.

5.6.2.2 Restriction of counting pairwise differences in concordant pairs

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

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

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

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

5.6.3 Impact of data error

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

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

5.6.3.1 Shared haplotype inference

In Figure 5.14 (next page), the physical distance between inferred breakpoints and the corresponding focal site is shown relative to the true distance as determined from simulation records. In the following, the genetic lengths of inferred breakpoint intervals around alleles at a given frequency was used to evaluate the accuracy of the HMM; note that boundary cases were removed. Results are shown in Figure 5.15 (page 205), for the analysis before and after data error, for analyses on true and phased haplotypes, and for discordant pairs selected as the nearest neighbours or at random. The summary statistics used to measure accuracy are also given in Figure 5.15.

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

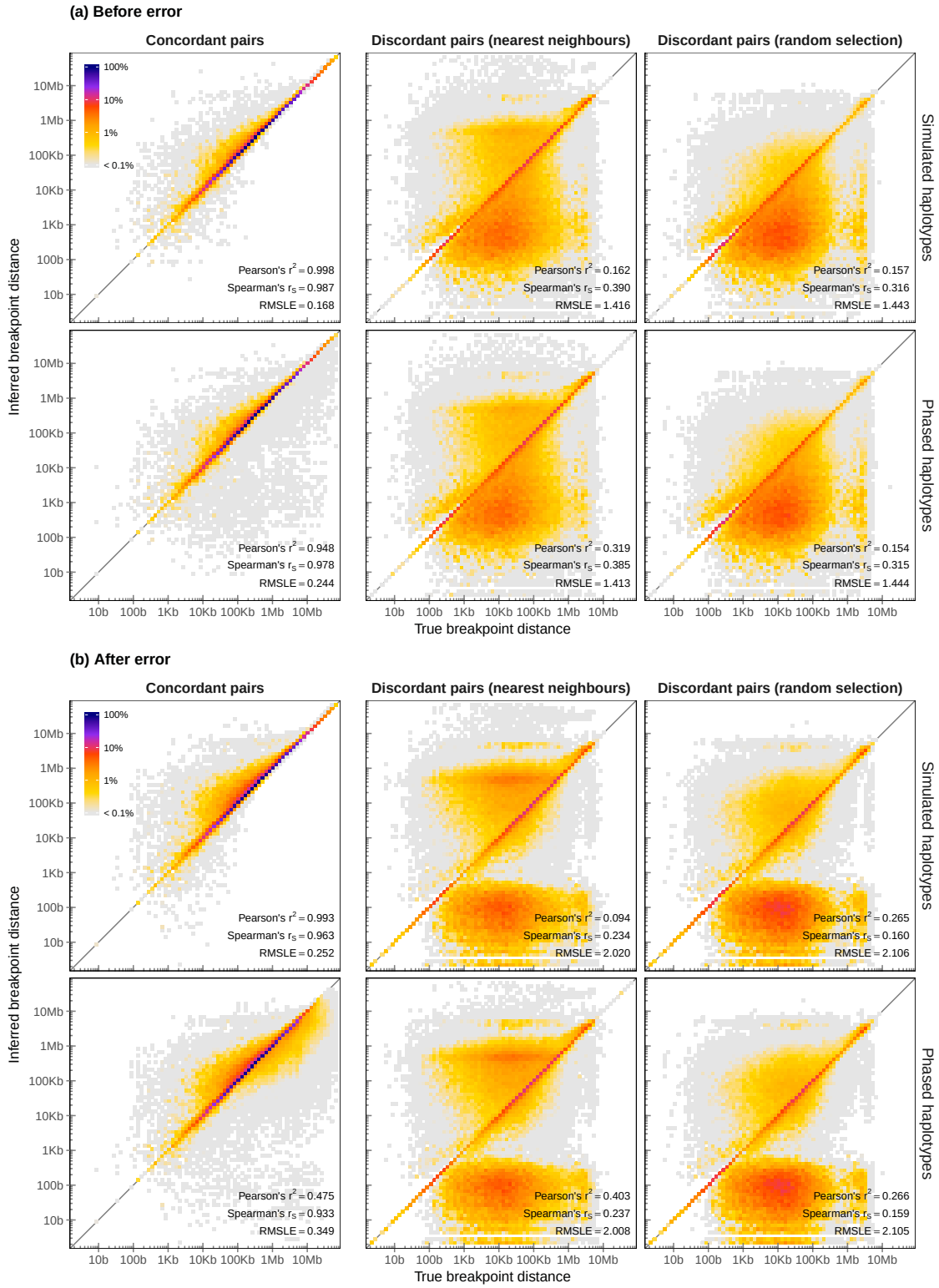


Figure 5.14: Density of breakpoint positions inferred using the haplotype-based HMM. The physical distance between true and inferred breakpoints (either left or right-hand side) is shown, where colours indicate the maximised density of breakpoints at the relative distance to the focal position, around which a shared haplotype segment was detected. Pair selection was done using the relaxed nearest neighbour approach and at random. Note that concordant pairs were selected at random throughout. Results were obtained on data before **(a)** and after error **(b)**, separately on simulated and phased haplotypes.

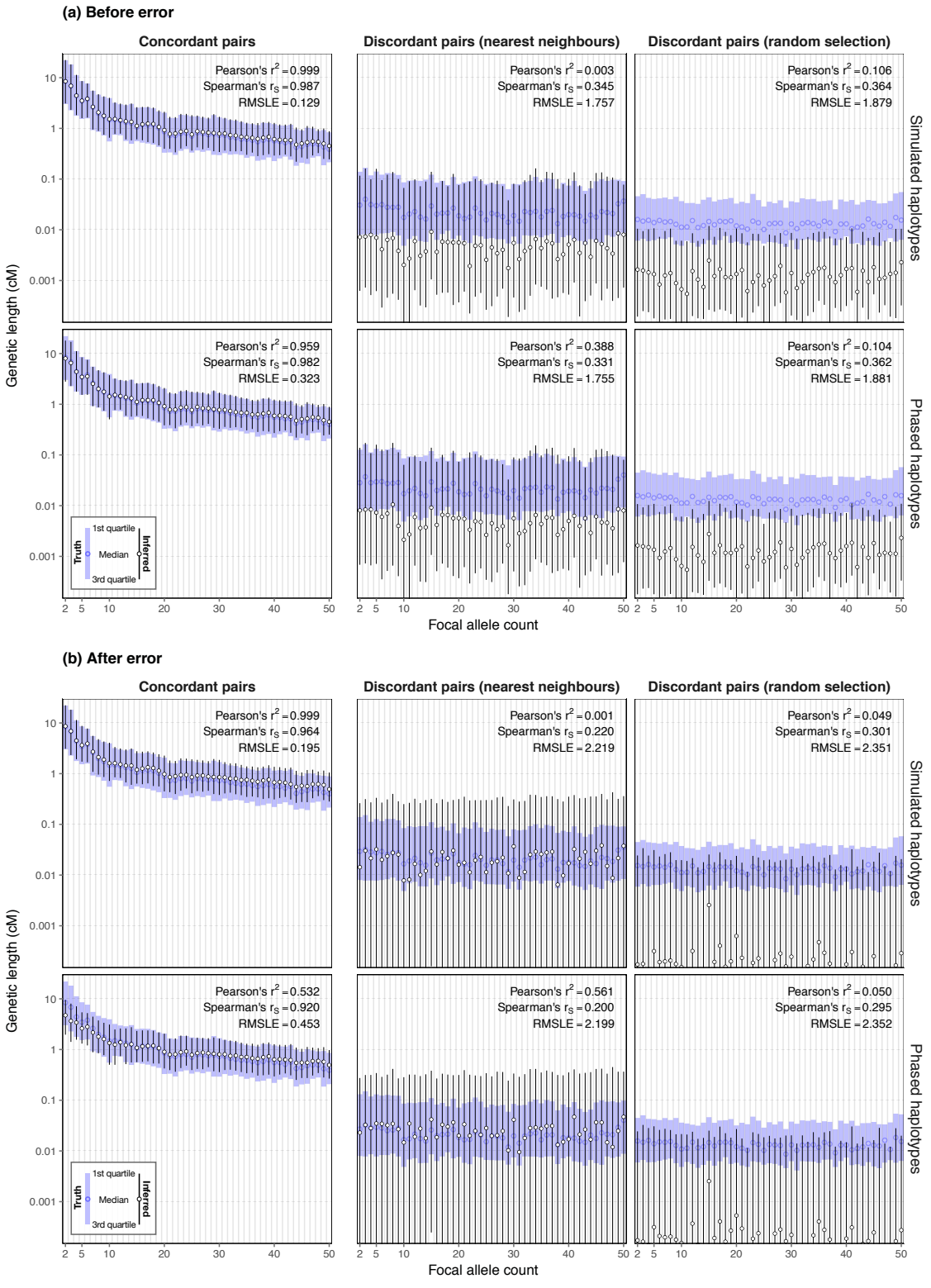


Figure 5.15: Genetic length of shared haplotype segments using the haplotype-based HMM. Inferred genetic length is shown by allele count of the focal variant in the simulated sample of $N = 5{,}000$ haplotypes, in direct comparison to the corresponding true segment lengths (*blue*). Pair selection was done using the relaxed nearest neighbour approach and at random. Note that concordant pairs were selected at random throughout. Results were obtained on data before **(a)** and after error **(b)**, separately on simulated and phased haplotypes.

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

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

5.6.3.2 Allele age estimation

The shared haplotypes inferred using the haplotype-based HMM were subsequently used to estimate allele age at the selected target sites. Figure 5.16 (next page) shows the results obtained when the nearest neighbour approach was used. Age estimated using the mutation clock ($\mathcal{T}_{\mathcal{M}}$), recombination clock ($\mathcal{T}_{\mathcal{R}}$), and combined clock ($\mathcal{T}_{\mathcal{MR}}$) were compared before and after error, and for simulated and phased data. Summary statistics for both the nearest neighbour and random selection approaches are given in Table 5.4 (page 208). Note that "true" age was set at t_m; *i.e.* the geometric mean between the time of coalescent events below and above the focal mutation event (t_c and t_d), according to which accuracy was measured.

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

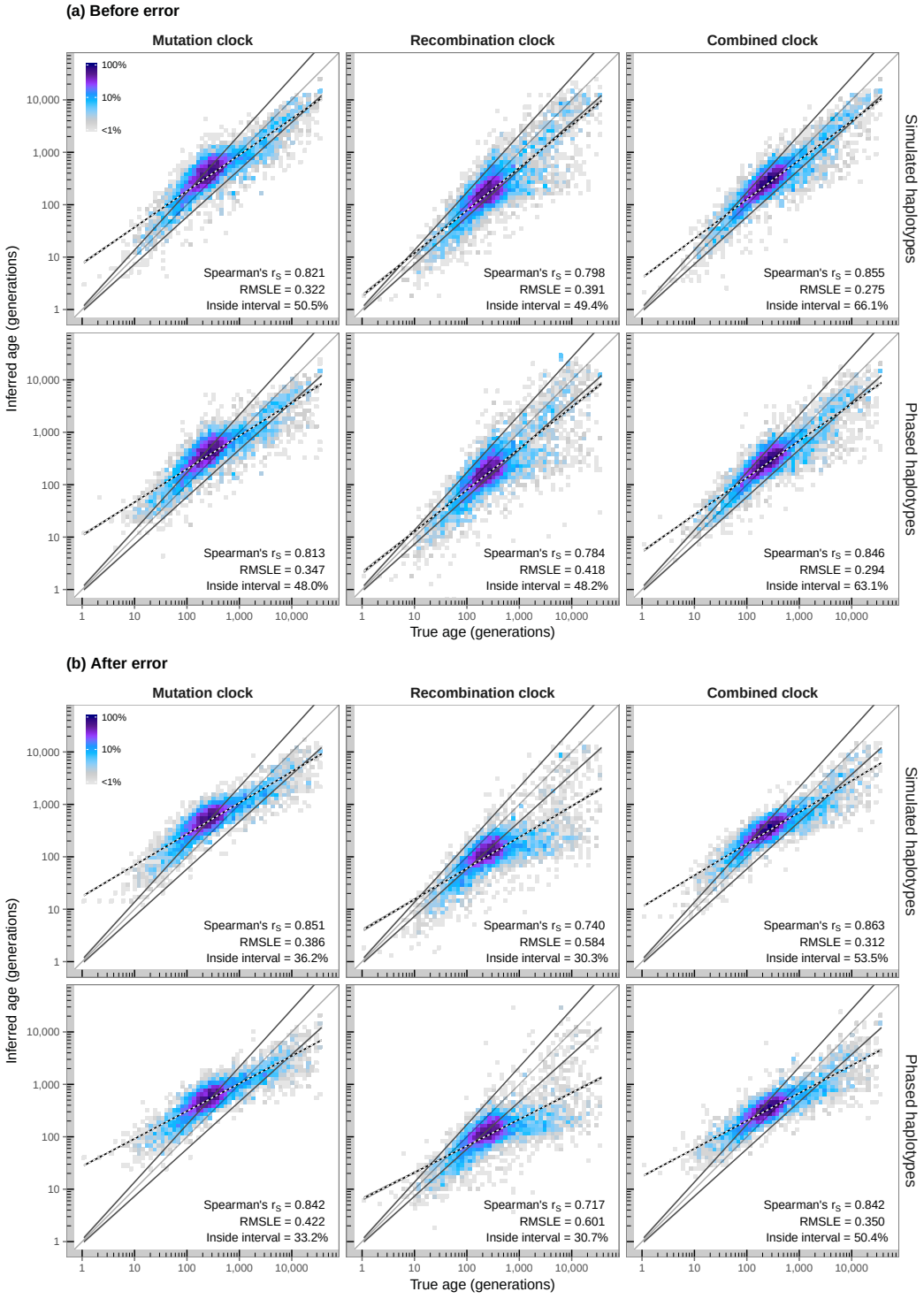


Figure 5.16: Allele age inferred using the haplotype-based HMM. The haplotype-based HMM was used to infer breakpoint intervals in data before **(a)** and after error **(b)**; analyses were conducted on "true" (simulated) haplotypes and phased haplotypes. Age was estimated using each of the three clock models, for a random set of 5,000 target sites at $f_{2,50}$, for which 100 concordant pairs were randomly selected and 100 discordant pairs were selected using the relaxed nearest neighbour approach. Note that "true" age was set at t_m, according to which the summary metrics shown were calculated. Regression lines above and below the dividing line indicate t_c and t_d. The *black-white* line shows the regression over age estimates. Colours indicate the density of true and estimated age; scaled as the maximised density per panel.

Table 5.4: Accuracy of inferred age using the haplotype-based HMM. Age was estimated using the mutation clock ($\mathcal{T}_{\mathcal{M}}$), recombination clock ($\mathcal{T}_{\mathcal{R}}$), and combined clock ($\mathcal{T}_{\mathcal{MR}}$) based on inference of breakpoint intervals using the haplotype-based HMM. Analyses were conducted on dataset $\mathcal{D}_B$ and $\mathcal{D}_B^*$ (*i.e.* before and after error), as well as simulated and phased haplotype data in both. Note that these metrics were calculated with respect to t_m; *i.e.* the geometric mean between t_c and t_d.

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

5.6.3.3 Discussion

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

However, it would be useful to compare this method to other, established approaches. For this purpose, in the following section, I used the Pairwise Sequentially Markovian Coalescent (PSMC) to infer the T_{MRCA} at concordant and discordant pairs, from which I derived a posterior distribution that was implemented as described for the CCF in Section 5.2.2.2 (page 172) to estimate allele age.

5.6.4 Comparison to the Pairwise Sequentially Markovian Coalescent (PSMC)

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

I used the PSMC-HMM as a method to infer the T_{MRCA} of concordant and discordant pairs in the estimation of allele age. For a given pair, I extracted the computed coalescent time posteriors at a focal position, which I then used in the same way as the posteriors obtained through a "clock" model in the cumulative coalescent function (CCF), so as to compute the composite posterior distribution at discrete time intervals for subsequent age estimation.

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

5.6.4.1 Implementation to estimate allele age

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

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

$$t_i = 0.1 \times e^{\frac{i}{n}\log(1+10T_{\max})} - 0.1 \tag{5.28}$$

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

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

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

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

5.6.4.2 Results for the T_{MRCA}

Simulation records were scanned to obtain the true T_{MRCA} for each haplotype pair at a given target site. The true time of coalescent events was compared to point estimates taken at the median of posterior distributions. The median was chosen because the Gamma distribution used to compute posteriors in the clock models may equal zero at the mode, such that the median was seen as a more reliable estimate. Results for concordant and discordant pairs are shown in Figure 5.17 (next page).

The combined clock, $\mathcal{T}_{\mathcal{MR}}$, showed the highest accuracy overall among the clock models when concordant pairs were considered; measured using rank correlation (r_S) and root mean squared logarithmic error (RMSLE). The mutation clock, $\mathcal{T}_{\mathcal{M}}$, was slightly more

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

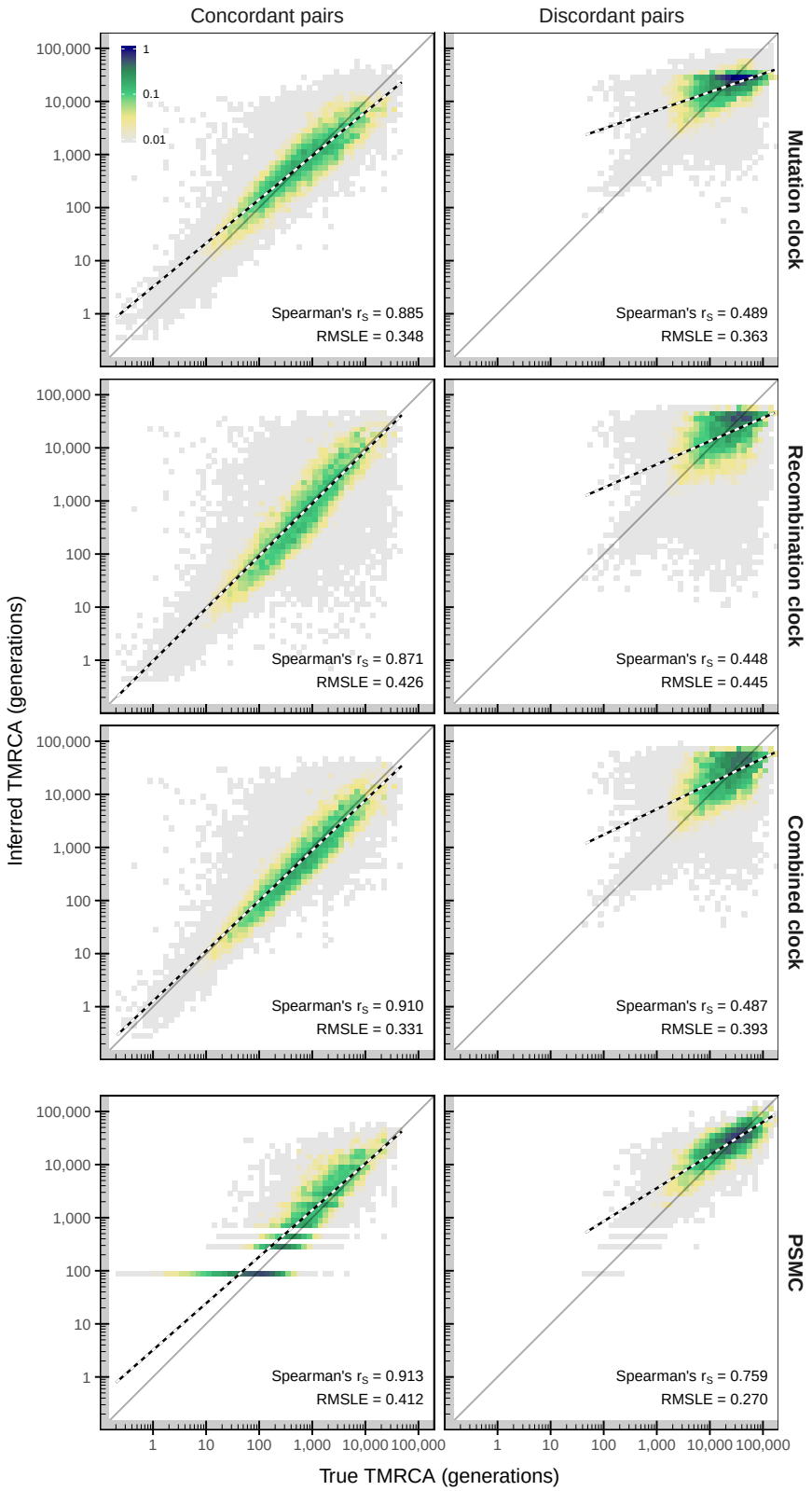


Figure 5.17: True and estimated T_{MRCA} using PSMC. The clock models developed in this chapter were compared to the PSMC-based method for inference of the T_{MRCA}. The posterior distribution on the coalescent time was obtained for the same sets of concordant and discordant pairs at 1,000 randomly selected target positions in simulated data ($\mathcal{D}_A$). Point estimated were taken at the mode of posterior distributions. Colours indicate the density of true and estimated coalescent time; scaled as the maximised density.

accurate for discordant pairs. The PSMC-based approach, however, was highest in terms of r_S in both concordant and discordant pairs, but where RMSLE indicated a higher magnitude of error compared to $\mathcal{T}_{\mathcal{MR}}$ and $\mathcal{T}_{\mathcal{M}}$. Notably, as seen in Figure 5.17, because of the discrete time intervals in PSMC it was problematic to infer recent T_{MRCA}. While it would be possible to increase the number of hidden states to include intervals at more recent points in time, a consequence would be that computations would become substantially slower.

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

An additional analysis was conducted using the results obtained above by sorting pairs by their true T_{MRCA} into broader time intervals at which accuracy was measured (r_S and RMSLE). The results are given in Table 5.5 (this page). The PSMC-HMM was notably more accurate compared to each clock model when discordant pairs where considered. But for concordant pairs, $\mathcal{T}_{\mathcal{MR}}$ outperformed PSMC when true coalescent time was more recent that 10,000 generations.

5.6.4.3 Results for allele age

Next, allele age was estimated by calculating the composite posterior distribution from the T_{MRCA} posteriors obtained in pairwise analyses per approach. The mode of the resulting composite posterior was taken as a point estimate for allele age. Results are shown in Figure 5.18 (this page), in which the relevant summary statistics to quantify accuracy are indicated; that is, r_S, RMSLE, and the proportion of "correct" alleles (estimated to sit between t_c and t_d). As before, "true" age was set at t_m (geometric mean between t_c and t_d).

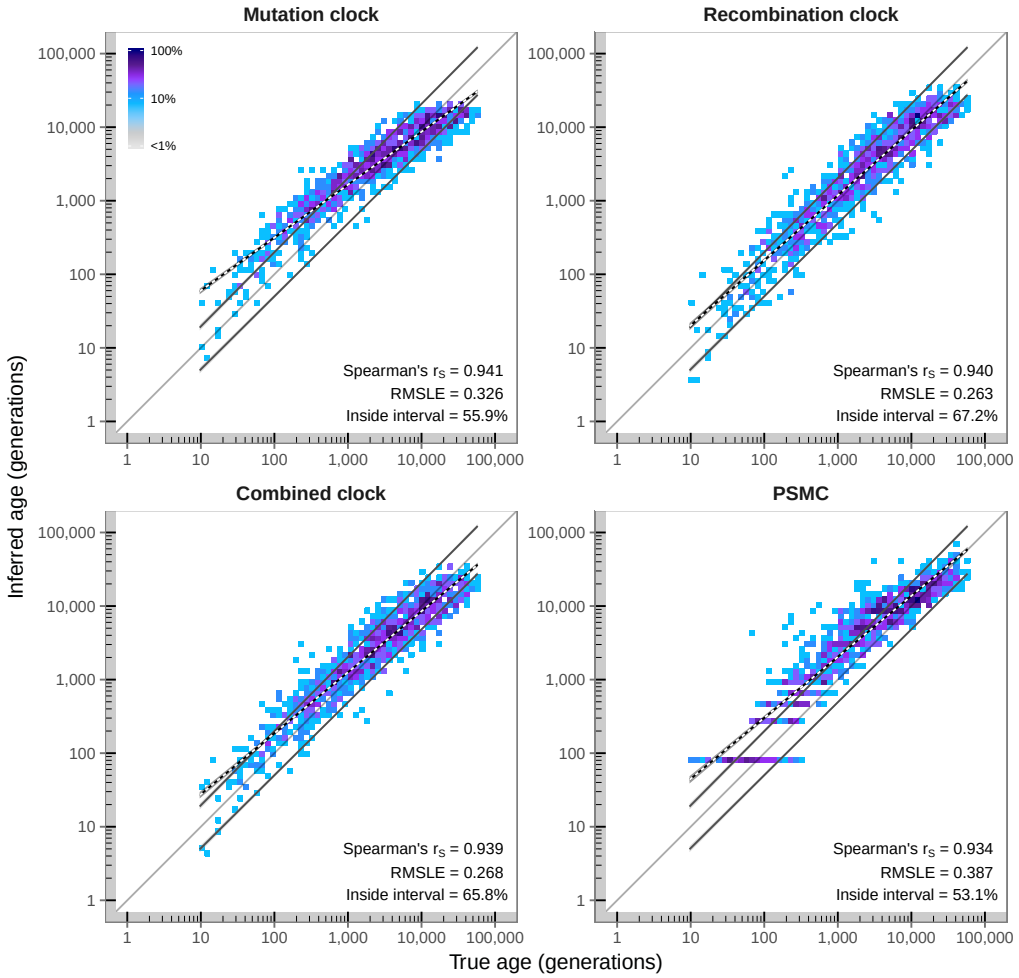


Figure 5.18: Allele age inferred using PSMC. The three clock models ($\mathcal{T}_M$, $\mathcal{T}_R$, and $\mathcal{T}_{MR}$) were compared to PSMC. Posterior distributions obtained on the same set of concordant and discordant pairs were used to compute the composite posterior distribution, from which point estimates were taken at the mode in each approach. The results shown compare the "true" age of an allele (set at t_m) to the estimated age at 1,000 target sites in dataset $\mathcal{D}_A$, which were randomly selected at allele frequency ≤ 50%. Regression lines above and below the dividing line indicate the regression over t_c and t_d. The *black-white* line shows the regression over age estimates. Colours indicate the density of true and estimated age; scaled as the maximised density.

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

Additionally, these results were again sorted into broader time intervals. Three intervals were distinguished at a nominal value of 1,000 generations, so as to distinguish relatively "young" alleles from "old" ones, considering that the distribution of true allele age was defined at overlapping age intervals at t_c and t_d. Exact definitions and accuracy results are given in Table 5.6 (this page). Notably, rank correlation measured at alleles estimated based on PSMC was highest for older alleles, but where r_S was higher for $\mathcal{T}_{\mathcal{M}}$ when younger alleles were considered. The proportion of correct alleles and the magnitude of error, however, favoured $\mathcal{T}_{\mathcal{R}}$ and $\mathcal{T}_{\mathcal{MR}}$ overall.

Table 5.6: Accuracy of allele age inferred using PSMC. The true times of the delimiting coalescent events t_c and t_d are used to distinguish alleles of young, intermediate, and old age; see definitions provided at the bottom of the table.

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

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

5.6.4.4 Discussion

Allele age estimated using $\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, or $\mathcal{T}_{\mathcal{MR}}$ was overall comparable to the estimates obtained using the PSMC-based approach. However, it must be noted that the integration of PSMC as a method to subsequently arrive at the composite posterior distribution may not have been an ideal baseline for comparisons to the clock models developed in this chapter. In particular, the current procedure considered 64 coalescent time intervals, which may not directly compare to the continuous scale used by the estimation method.

Nonetheless, the results presented in this section suggested that the haplotype-based HMM used for targeted shared haplotype inference achieved comparable estimates of T_{MRCA}, despite incorporating a (biased) empirical emission model. I further showed that

the haplotype-based HMM was able to estimate age of alleles that occurred at relatively high frequencies in the data. Also, note that the decreased accuracy in inferences at discordant pairs had only minor effects on subsequent estimation of allele age.

5.6.5 Allele age estimation in 1000 Genomes

The haplotype-based HMM was used for allele age estimation in an extensive analysis of data from the 1000 Genomes Project (1000G) Phase III. I selected 50,000 sites at random in chromosomes 20, but only at positions within high confidence regions as defined in the strict accessibility mask available for the Phase III dataset. The diploid sample size was $N = 2{,}504$. Note that all sites at $f_{\geq 2}$ were considered. Model parameters in `rvage` were $N_e = 10{,}000$, $\mu = 1.2 \times 10^{-8}$, and recombination rates according to genetic maps available from HapMap Phase II, Build 37.*

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

5.6.5.1 Shared haplotypes inferred by population

In this section, I focused on concordant pairs of which 2.357 million were retained after removing boundary cases. Specifically, I compared haplotype length distributions for pairs at which the focal allele was shared among the individuals within the same population and those at which it was shared between different groups. Median physical and genetic lengths are shown in Figure 5.19 (next page); the proportion of pairs at which alleles were shared within and between populations is indicated.

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

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

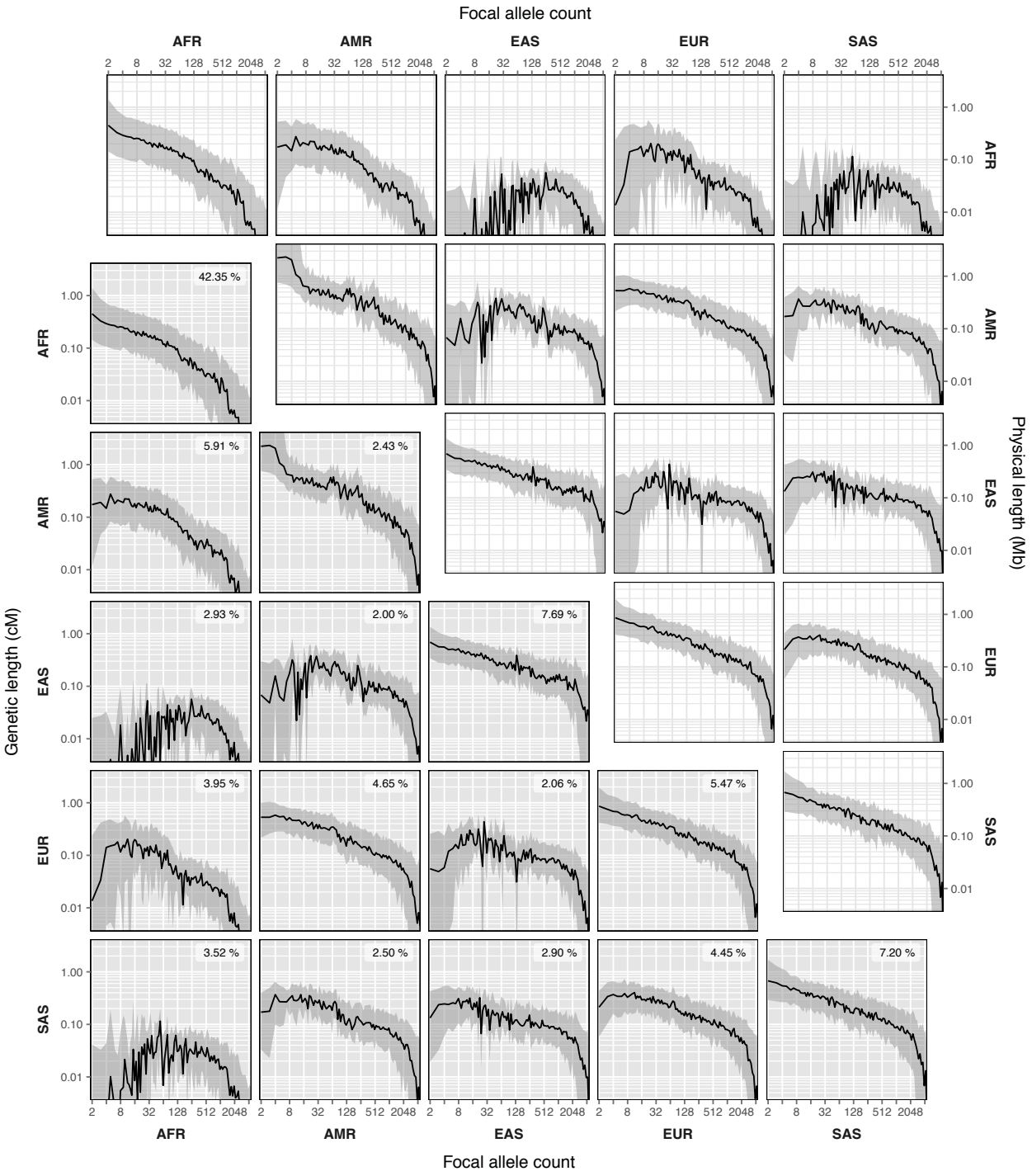


Figure 5.19: Shared haplotype length by population in 1000 Genomes, chr. 20. Median physical and genetic lengths are shown for haplotype pairs where individuals share a focal allele within the same or between different populations. Note that the condition of seeing an allele shared between individuals implied that only concordant pairs were considered. Physical lengths are given in the upper triangle (*white* panels) and genetic lengths in the lower triangle (*grey* panels). The median (*black* lines) is drawn between the 1st and 3rd quartiles, where lengths were binned by focal allele count (log-scale). The lower triangle also shown the proportions of pairs at which the focal allele was shared within or between populations.

5.6.5.2 Allele age estimated by population

Age was estimated using the relaxed nearest neighbour approach to select discordant pairs, and using the same model parameters per population group. Note that N_e is expected to differ across populations, but which may only lead to inappropriate scaling of the population-scaled age estimates per group. Here, I focused on alleles that were shared only within populations, which retained 31,906 target sites. The results for each clock model are shown in Figure 5.20 (this page).

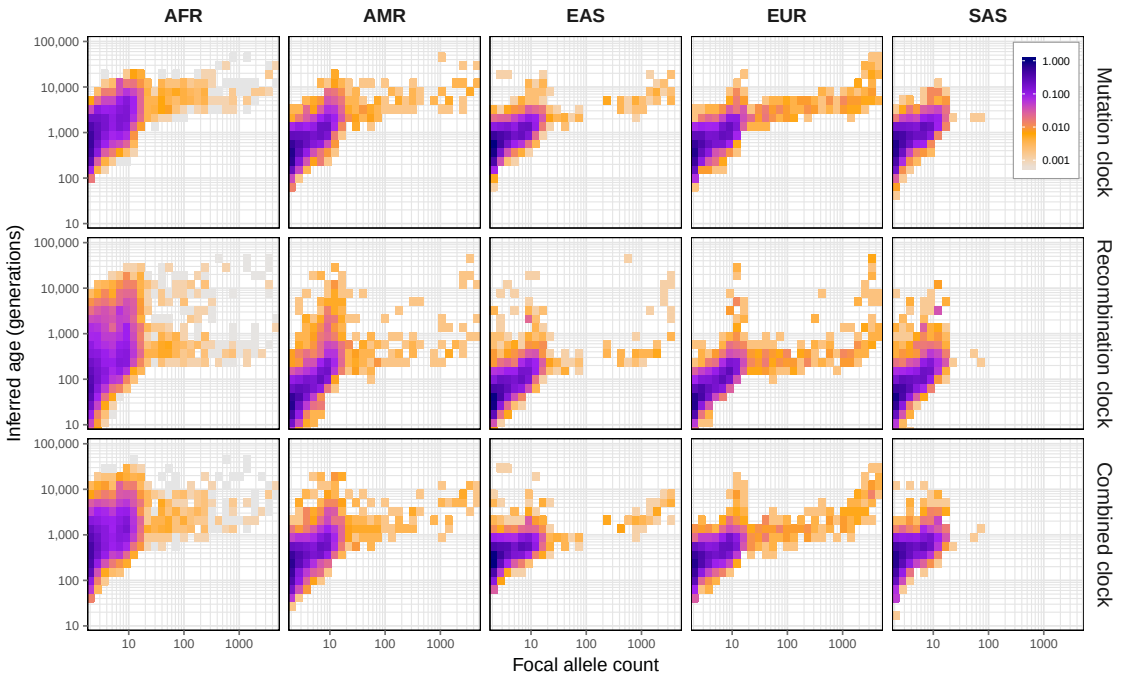


Figure 5.20: Allele age estimated by population in 1000 Genomes, chr. 20. Age was estimated for alleles shared within the same population group contained in the 1000G sample, using the relaxed nearest neighbour approach to select discordant pairs. Shared haplotype detection was performed using the haplotype-based HMM and age was estimated under each of the three clock models ($\mathcal{T}_{\mathcal{M}}$, $\mathcal{T}_{\mathcal{R}}$, and $\mathcal{T}_{\mathcal{MR}}$). Colours indicate the maximised density of allele age by focal allele frequency. Note that results are shown on log-log-scale.

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

5.6.5.3 Selected allele age profiles

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

Figure 5.21 (next page) shows the *profiles* of the inferred shared haplotype distribution and allele age at two selected target sites in 1000G. Pairs were sorted by T_{MRCA} to indicate the shape of the underlying genealogy around a given focal site. Estimated allele age is shown in between the sorted concordant and discordant pairs.

The example shown in Figure 5.21a is an intron variant the MCM6 locus on chromosome 2, which sits approximately 22 kb upstream of the LCT gene (encoding the lactase enzyme) and has been associated with lactase persistence in Europeans. For example, Enattah *et al.* (2002) found that the variant occurs in distantly related populations and therefore concluded that it is relatively old. Research on the LCT gene has indicated signatures of recent positive selection around 5,000 – 10,000 years ago (Bersaglieri *et al.*, 2004). Here, age estimation suggested that the variant originated at around 800 generations ago (according to $\mathcal{T}_{MR}$).

A missense variant at the PRDM9 locus in chromosome 5 is shown in Figure 5.21b. Interestingly, the variant is low in frequency in the sample, but its age was estimated to be surprisingly old, and far older in comparison to Figure 5.21a. Note that I also analysed other SNPs at PRDM9 which also indicated an old age and a similar haplotype structure. However, here, I only attempted to demonstrate one possible application of the methodology, but further research would be needed to arrive at conclusive results for either of the examples provided.

5.6.6 Discussion

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

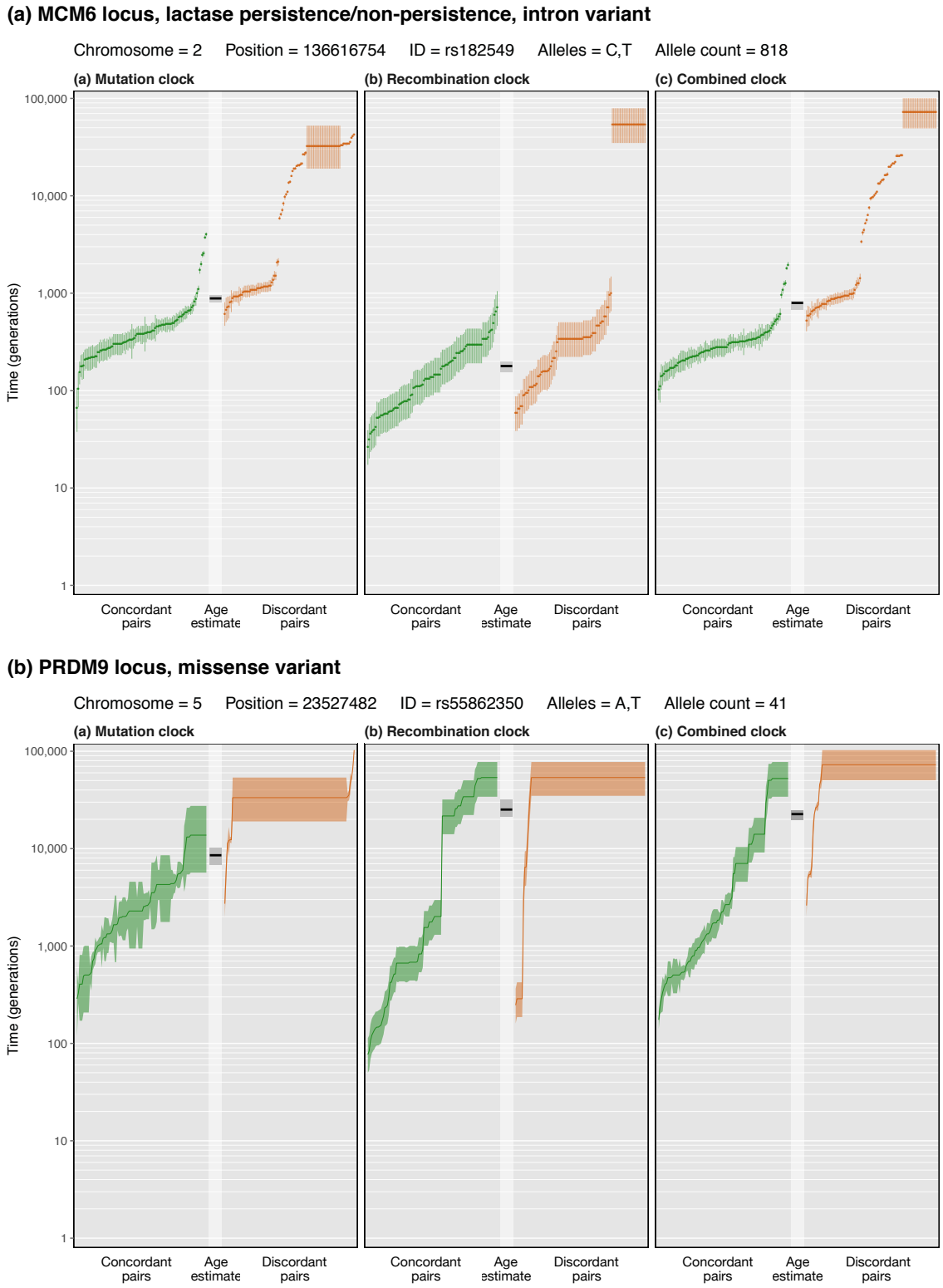


Figure 5.21: Example profiles of estimated allele age in 1000 Genomes, chr. 20. Allele age profiles are shown for two selected loci in 1000G data. Each profile is composed of the T_{MRCA} posterior distributions inferred for concordant pairs (*left*) and discordant pairs (*right*), which are sorted by their inferred T_{MRCA}. The estimated age obtained from the resulting composite posterior distribution is shown in the *middle*. Each T_{MRCA} distribution is shown as the median, around which the 1st and 3rd quartiles are drawn. The mode of the composite posterior distribution is shown within a 95% pseudo-confidence interval. Time was scaled at $2N_e$, where $N_e = 10{,}000$.

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

Write a wise saying and your name will live forever.
— Anonymous

6 Discussion

Contents

This thesis had the aim of developing novel strategies for the statistical analysis of rare genetic variants in context of complex disease research. In the time since I began working towards this broadly defined goal, many major milestones have been achieved in human genetic research, in particular due to (still ongoing) advancements in high-throughput sequencing technologies. This facilitated research on a previously unseen scale and offered a wide range of opportunities to be explored.

One major achievement of human genetic research is seen in the success of genome-wide association (GWA) studies which, within less than a decade, have identified thousands of loci associated to a large number of disease phenotypes. At that time, not much was known about the role of rare variants in complex disease, wherein I saw an opportunity to contribute to current (and highly dynamic) research. I became interested in the development of novel approaches that would allow researchers to harness the growing amount of genomic data to find rare allele associations to disease risk.

The recent past has also seen a major change in the division between areas of applied medical research and the traditionally more theoretical field of population genetics. The line between those fields became more blurred, such that it could be argued that it was never there to begin with. The discoveries made in the human genome, in particular through the Human Genome Project (International Human Genome Sequencing Consortium, 2001, 2004), have necessitated a deeper understanding of how genomic

variation as a consequence of demographic and evolutionary forces contributes to variation of phenotypic traits. I became fascinated by the possibility to use genomic data to draw inferences about past historic events in populations, and to further use such insights to make predictions about the effects on disease traits. In that regard, I intended to focus on rare variants which, due to their presumed recent origin through mutation, could be seen as a source of information about the recent past. My interest grew into what now covers the major parts of this thesis.

6.1 Summary and main conclusions

The following main directions can be distinguished in this thesis. First, I attempted to develop a computational method to integrate different sources of data, so as to increase the number of rare variants that can be interrogated in association analysis. Second, a large part of this thesis is concerned with methods development to detect shared haplotype segments that are inherited from a common ancestor. Lastly, my primary goal was to develop the methodology and statistical theory required to computationally estimate the age of an allele observed at a single locus. In the following, I briefly review the methodology I developed and give a summary of the main ideas, findings, and conclusions.

6.1.1 Imputation and association analysis or low-frequency alleles

In Chapter 2, I attempted to develop a method to combine independent genomic datasets through genotype imputation as a way to improve statistical power to implicate rare variants in GWA studies. Rare variants are less likely to exert high risk effects on complex disease phenotypes, or otherwise could be picked up through conventional linkage analysis. Variants at lower frequencies may nonetheless play a role in disease aetiology, but are often highly population or cohort specific. Imputation from a single reference panel may therefore not produce genotype data where low-frequency alleles would be likely to be implicated in disease risk through association tests.

I developed a computational tool (*meta-imputation*) to combine genotype data obtained in separate imputations from independent reference panels. The intuition was that the information contained in different studies could be leveraged by bringing these datasets together through imputation into the same study sample. I thereby attempted to capture variants at lower frequencies or those specific to one or few source panels. The method produces a single, canonical genotype dataset that can be used in subsequent association analyses.

The method was evaluated, first, in terms of the achieved genotype accuracy, which was compared to each of the separately imputed datasets that were previously combined through meta-imputation. A second evaluation was performed through a series of simulated GWA studies. This, however, turned out to be particularly challenging due to substantial computational demands. Thousands of simulated datasets had to be generated, which were used as sample for imputations from multiple reference panels, such that each dataset was treated as an independent GWA study. The main findings are summarised as follows.

- I showed that the method was able to improve overall genotype accuracy when data imputed from multiple studies were combined.

- The accuracy of meta-imputed genotypes improved notably in cases where data from more diverse population groups were available.

- Statistical power in association analysis was increased at low-frequency variants with intermediate or high penetrance, but where differences seen for common risk variants were negligibly small.

- Rare variants with low penetrance were unlikely to reach statistical significance.

Note that I used Phase I data from the 1000 Genomes Project (1000G). Shortly after my work on this project was completed, 1000G Phase III was released. I decided not to repeat the analysis conducted for the evaluation of meta-imputation with the newly released dataset, because the larger sample size and higher marker density would entail that the time and computational load needed to repeat the analysis would be too prohibitive.

In light of the imputation reference panel that became available through the Haplotype Reference Consortium (HRC), see McCarthy *et al.* (2016), potential applications of the meta-imputation method may be limited to specific niche problems. For example, for populations that are currently underrepresented in the HRC panel, or when novel datasets become available that are more suited for imputations into a given study cohort than HRC alone.

6.1.2 Shared haplotype inference around rare variants

Recent advancements in high-throughput sequencing technologies have enabled the study of rare variation present in the human genome. It is assumed that frequency is indicative for the time since an allele was created through a mutation event. Rare alleles are therefore likely to indicate recent relationships among haplotypes, which is informative about recent demographic history and population structure, within and among populations. Given deeper insight into the sharing structure of alleles, it would be possible, for example, to infer the timing of demographic events. Such inferences could be useful to make further statements with regard to selection or other forces that shaped the genetic variation observed in a population.

In Chapter 3, an initial attempt was made to develop deterministic (rule-based) ways to utilise rare variants as "bookmarks" in the underlying genealogy, to identify haplotypes that recently derived from a common ancestor. I presented two approaches that are based on the four-gamete test (FGT) after Hudson and Kaplan (1985), through which the "breakpoint" of a recombination event along the sequence can be detected. While the FGT is based on observing specific allelic configurations in haplotype data, its simplified form, the discordant genotype test (DGT), requires only genotype data. I developed a computational tool (`tidy`) to detect breakpoint intervals around a given focal site in pairs of diploid individuals that carry the focal allele. The resulting interval is assumed to indicate the underlying shared haplotype segment that was inherited "identical by descent" (IBD) in each pair sharing the focal allele.

In Chapter 4, this idea was extended using a Hidden Markov Model (HMM) that operates on genotype data; referred to as HMM_g below. I determined genotype error rates in different sequencing and genotyping datasets, which I used to modify simulated data to perform subsequent evaluations of the developed detection methods in realistic settings. Empirically measured error rates were also used to construct an emission model to make the HMM robust towards error in applications to real data. I employed a similar approach in a novel haplotype-based HMM, presented in Chapter 5 and referred to as HMM_h below.

I assessed the performance of the detection methodology developed in this chapter in comparison to an existing, probabilistic method for IBD detection (*Refined IBD*), in simulated data with and without empirically measured distributions of data error, as well as in real data. In summary, the main findings are as follows.

- The FGT generally achieved high levels of accuracy when used to detect breakpoints around rare variants (allele frequency $\leq 0.5\%$) in simulated data. The DGT performed similarly well in simulated settings, but resulted in an overall reduced precision of the detected breakpoints. Importantly, both showed a tendency to overestimate the actual shared haplotype interval.
- I showed that the accuracy of the FGT decreased when haplotypes were computationally phased (inferred from genotype data), but where the impact of phasing error was low on average. Note that genotype-based approaches (DGT and HMM_g) are insensitive towards phasing error.
- When data error was considered in simulations, segment detection using either the FGT or DGT resulted in vastly underestimated breakpoint intervals. I showed that segment truncation was due to the detection of false positive breakpoints, caused by a relatively small proportion of misclassified alleles or genotypes at single loci. I further showed that existing (probabilistic) methods may also be adversely affected by data error. Similar patterns were seen in applications to real data. Thus, I concluded that the deterministic approaches developed in this thesis (FGT and DGT) were unreliable to indicate the extent of actual shared haplotype regions.
- In simulated settings without considering data error, I showed that HMM_g outperformed the DGT, but performed less well compared to the FGT in terms of the precision to detect breakpoint locations. But both HMM_g and HMM_h were robust in presence of realistic distributions of data error, where segment length obtained in analyses of real data followed expected patterns.
- A main finding was that HMM_h was able to outperform all previous methods for segment breakpoint detection developed in this thesis. This was seen in simulated settings with and without consideration of data or phasing error, and was likewise suggested in applications to real data.

In conclusion, my work on shared haplotype inference has led to the development of HMM_h as the result of a steady progression to improve the methodology as a consequence of the faults and caveats seen through extensive evaluation of different approaches. However, the method employs an empirical emission model that was constructed on the basis of genotype error measured in the 1000 Genomes Phase III dataset, and was further constrained to consider relatively recent relationships. It is therefore implicitly assumed that the actual shared haplotype around a given focal site is "younger" than the immediately next segment along the sequence. While this may apply to the majority of rare variants, this assumption is less likely to be satisfied for "older" relationships, *e.g.* variants observed at higher frequencies.

6.1.3 Allele age estimation

Knowing the age of an allele would allow us to observe human evolutionary history back in time, to infer past demographic events and processes that resulted in the observed genetic variability in a population. The method for allele age estimation was presented in Chapter 5, which concluded this thesis.

Age is estimated for a given allele at a single locus as observed in the sample. This is done in a Bayesian setting in which a composite posterior distribution is computed from posterior probabilities of the time of coalescent events between pairs of haplotypes. According to observed allele sharing in the sample, a number of "concordant" and "discordant" haplotype pairs are formed and analysed in turn to obtain pairwise posteriors of the time to the most recent common ancestor (T_{MRCA}). These are used to indicate the time of coalescent events that delimit the branch on which the mutation event occurred, from which the focal allele derived. The age of an allele can be estimated directly from the resulting composite posterior distribution.

I provided the definitions for several models to obtain posterior densities from pairwise observations of mutational differences or recombinational length, where prior expectations were based on coalescent theory. As a result, the following "clock" models were defined; mutation clock ($\mathcal{T}_{\mathcal{M}}$), recombination clock ($\mathcal{T}_{\mathcal{R}}$), and combined clock ($\mathcal{T}_{\mathcal{MR}}$).

The methodology to detect shared haplotype segments around a given focal site, as developed in this thesis, was essential for the application of the age estimation method. This is because the observed values of the parameters specified by the clock models are determined from the data with respect to the shared haplotype structure inferred at a focal site in the sample. Note that this thesis had a strong emphasis on identity by descent (IBD) and shared haplotype inference for this reason. I implemented the age estimation method with the different segment detection methods in a computational tool (`rvage`) written in `C++`. In summary, the main findings are listed below.

- In a simulated setting when the actual shared haplotype structure at a focal site is known, I showed that allele age can be estimated with high accuracy, confirming the theoretical validity of the method.

- When shared haplotypes were inferred to subsequently estimate allele age based on observed parameters, haplotype-based detection methods generally outperformed genotype-based approaches in ideal simulated settings. The impact of phasing error was again seen as being low on average.

- In simulations where data error was considered to facilitate realistic evaluations, I showed that using the deterministic detection approaches (FGT and DGT) resulted in very high estimation bias, such that allele age could not be estimated reliably, *e.g.* when applied to real data.

- Although both HMM_g and HMM_h were previously seen as being robust when applied in realistic simulated settings, I showed that HMM_g resulted in an increased bias to underestimate allele age. In contrast, I found that estimation bias was low when HMM_h was used, and that correlations with measures of the actual time of underlying mutation events were high overall. This suggested that allele age can be estimated reliably when HMM_h is applied to real data.

- Using HMM_h, I found that the detection of shared haplotype segments was far less biased when concordant pairs (carrier haplotypes) were analysed compared to discordant pairs (carrier and non-carrier haplotypes), where the inference of T_{MRCA} was similarly affected. However, I found that the reduced accuracy for discordant pairs had a relatively low impact on the subsequent estimation of allele age.

- I additionally assessed the accuracy of the inference of pairwise T_{MRCA} using HMM_h in comparison to posterior probabilities obtained with the Pairwise Sequentially Markovian Coalescent (PSMC) model. I showed that HMM_h performed similarly well to infer T_{MRCA}, and that estimation bias of allele age was low for both. However, I found that PSCM achieved higher levels of accuracy when analysing discordant pairs compared to HMM_h under each clock model, but where bias was nonetheless increased compared to inferences at concordant pairs. Note that this was seen in a simulated setting with known demographic parameters, and without considering data error.

- Although HMM_h is implicitly biased due to its use of an empirical emission model constrained to relatively recent relationships between haplotype pairs, I found that it nonetheless performed well in age estimation of alleles that derived from "old" mutation events or that occurred at higher frequency in the sample.

- Lastly, I showed that each clock model performed well to estimate allele age, but where $\mathcal{T}_{\mathcal{MR}}$ generally achieved lower estimation bias and showed higher correlation with measures of the actual mutation time.

While the initial focus of this work was on rare variants, it would be useful to better assess the possibility of applying HMM_h to age estimation of alleles deriving from relatively old mutation events. The inference of shared haplotype intervals and T_{MRCA} was shown to be problematic for discordant haplotype pairs. Although this was found to have less impact on the estimation of allele age, compared to bias due to incorrect inferences at concordant pairs, it is a main caveat of the current implementation. I discuss future directions arising from the allele age estimation method in the section below.

6.2 Future directions

The allele age estimation method presented in Chapter 5 can be regarded as the main result of this thesis. As the application of this method is dependent on approximations to the underlying genealogy at a given site in the genome, the presented methodology for shared haplotype inference is of equal importance. Further development of such methods could therefore be seen as a goal for future work. For example, I used the PSMC model to achieve similar results, but where HMM_h has the advantage of being relatively fast such that an analysis of hundreds of thousands of haplotype pairs remains computationally tractable.

A possible enhancement would be to substitute the empirical emission model with a theoretical one, in which the probability of observing allelic combinations along the sequence are modelled by their expectations, given an (approximate) expectation of the T_{MRCA}. However, the division between *ibd* and *non* as hidden states remains arbitrary and leads to the implicit assumption that genealogies outside the genealogy at the focal site indicate a far older relationship between the haplotype pair. More complex models could be devised, for example based on the PSMC model, but where computation time should be considered to maintain practical scalability.

In the introduction to this thesis, I advocated a variant-centric view of genetic variation. By looking at the genealogy at a particular site in the genome, it would be possible to arrive at conclusions about the evolutionary and demographic influences that have resulted in the observed allele distribution and shape of the genealogy. Since the genealogy can only be known through statistical inference, the allele age estimation method along with the methods to infer the shared haplotype structure at a particular site may provide the necessary toolset to make significant discoveries. Future research will show to what extent the methodology developed in this thesis is useful.

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

— Elon Musk

Abbreviations

1000G	1000 Genomes Project
ARG	Ancestral recombination graph
CCF	Cumulative coalescent function
CDF	Cumulative distribution function
cM	CentiMorgan
DGT	Discordant genotype test
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
FGT	Four-gamete test
FNR	False negative rate
FPR	False positive rate
GoT2D	Genetics of Type 2 Diabetes Project
GWA	Genome-wide association
HapMap	International HapMap Project
HGP	Human Genome Project
HMM	Hidden Markov Model
HRC	Haplotype Reference Consortium
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IBS	Identity by state
IPG	Illumina Platinum Genomes Project
LD	Linkage disequilibrium
LOD	Logarithm of odds
LR	Likelihood ratio
LRP	Long range phasing
MAF	Minor allele frequency
Mb	Megabase
MRCA	Most recent common ancestor
MSMC	Multiple Sequentially Markovian Coalescent
NGS	Next-generation sequencing
NHGRI	National Human Genome Research Institute
OR	Odds ratio
PCR	Polymerase chain reaction
PDF	Probability density function
PMF	Probability mass function
PSMC	Pairwise Sequentially Markovian Coalescent
QC	Quality control
RFLP	Restriction fragment length polymorphism
RMSLE	Root mean squared logarithmic error
RNA	Ribonucleic acid
SFS	Site frequency spectrum
SMC	Sequentially Markov Coalescent
SNP	Single-nucleotide polymorphism
$\mathbf{T_{MRCA}}$	Time to the most recent common ancestor
VCF	Variant Call Format
WGS	Whole-genome sequencing

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.

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

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.

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

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

Fisher, R. A. (1930). *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. (2011). 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, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357), 493–U84.

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.

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

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

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

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.

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

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., *et al.* (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.

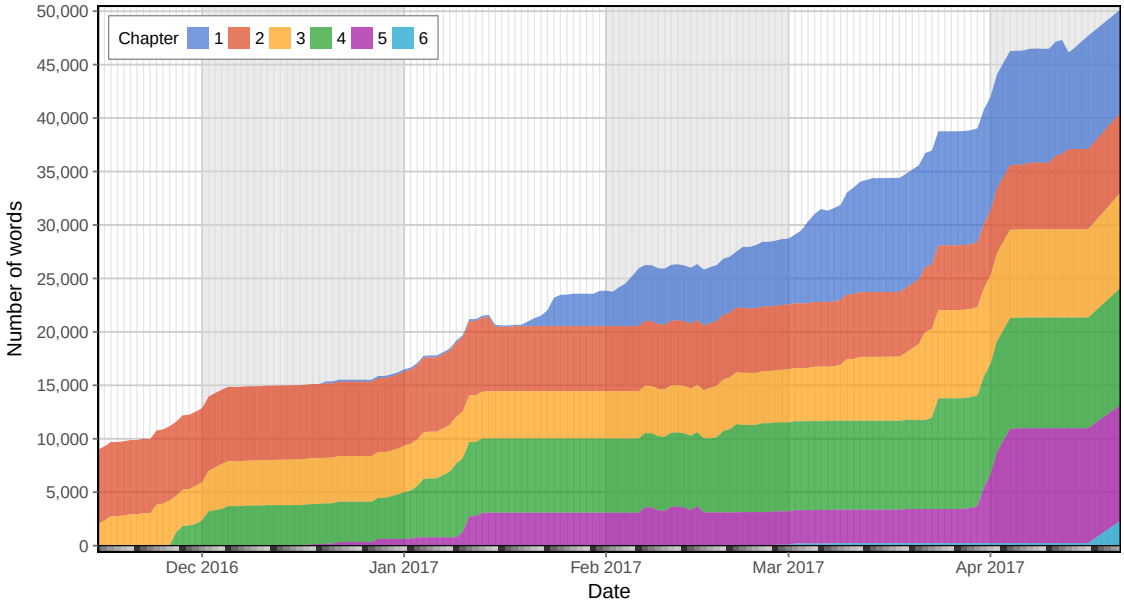
Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, **16**(2), 97–159.

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.

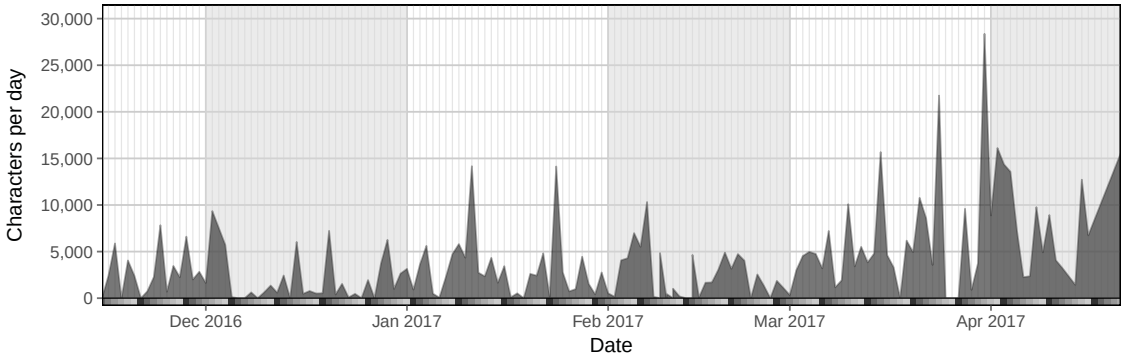
Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, **111**(4), E455–64.

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

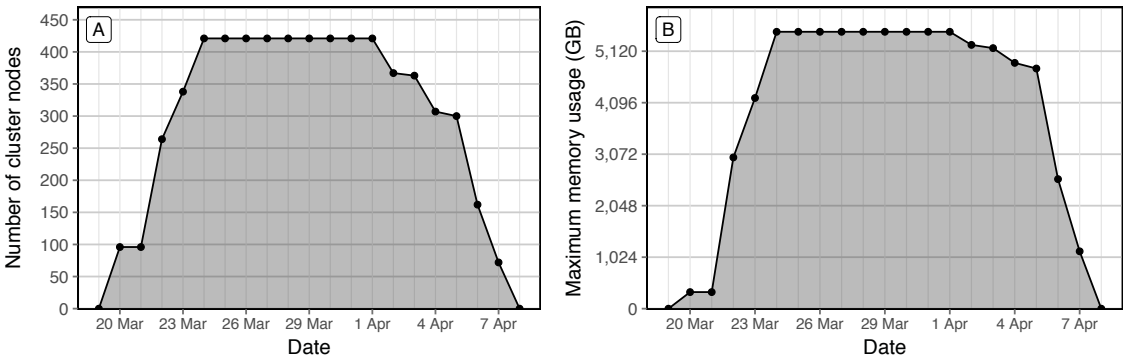
— Adam Savage



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



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



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