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.

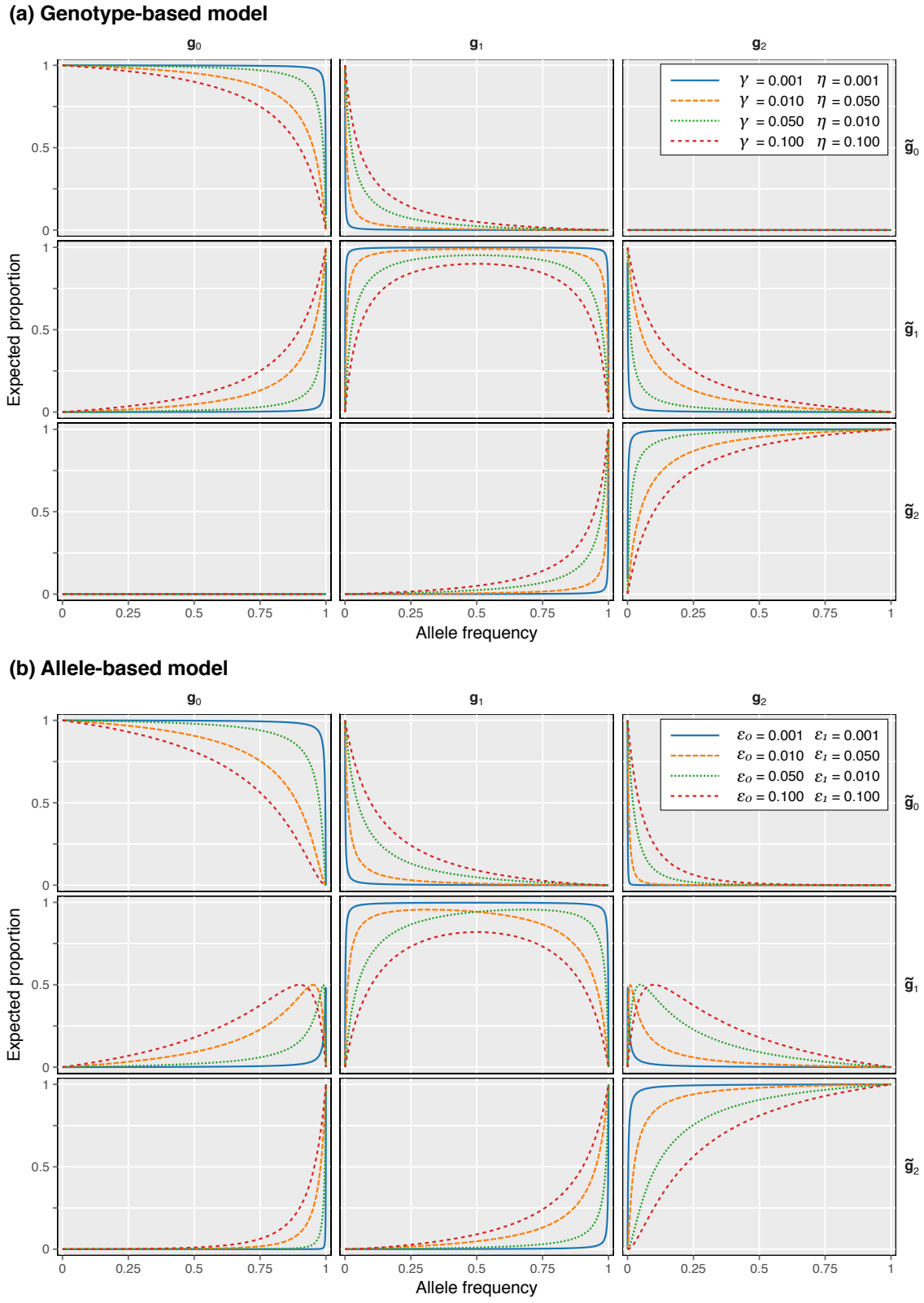


Figure 4.1: Expected proportions of genotype error for the genotype and allele-based models. The graphs show the expected proportion of true genotype g_j observed genotype as ($\tilde{g}_i$) given the population allele frequency; calculated at different, nominal error rates (see legend). The error functions 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.

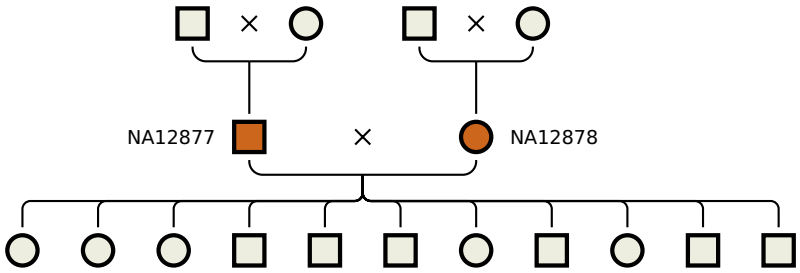


Figure 4.2: CEPH pedigree 1463. The pedigree of the family sequenced in the Illumina Platinum Genomes Project. Genotype data of individuals NA12877 and NA12878 (indicated) were used as reference against which data obtained on other genotyping or sequencing platforms were compared. Figure modified from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

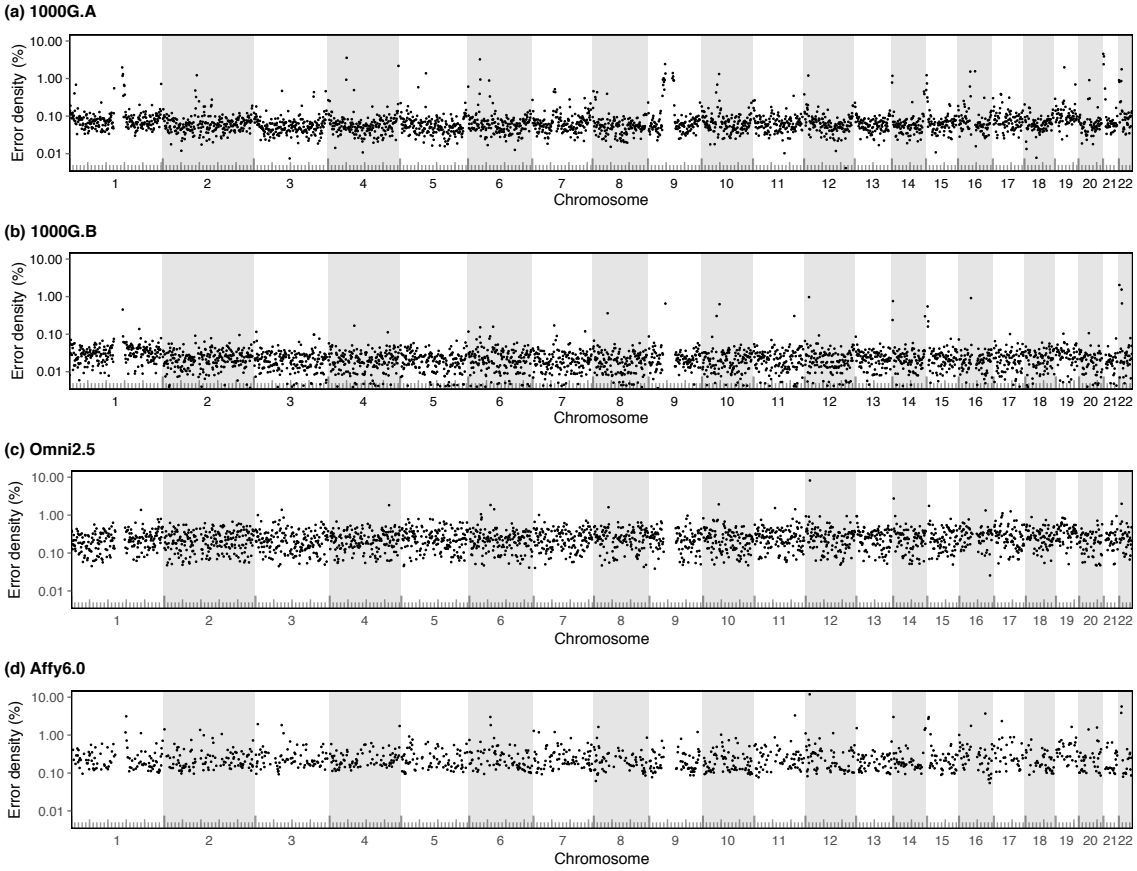


Figure 4.4: Positional genotype error density in sequencing and genotyping datasets. The density of misclassified genotypes was calculated along the length of each chromosome, which were divided into equally sized chunks of 1 Mb size. Error was calculated as the number of misclassified genotypes divided 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

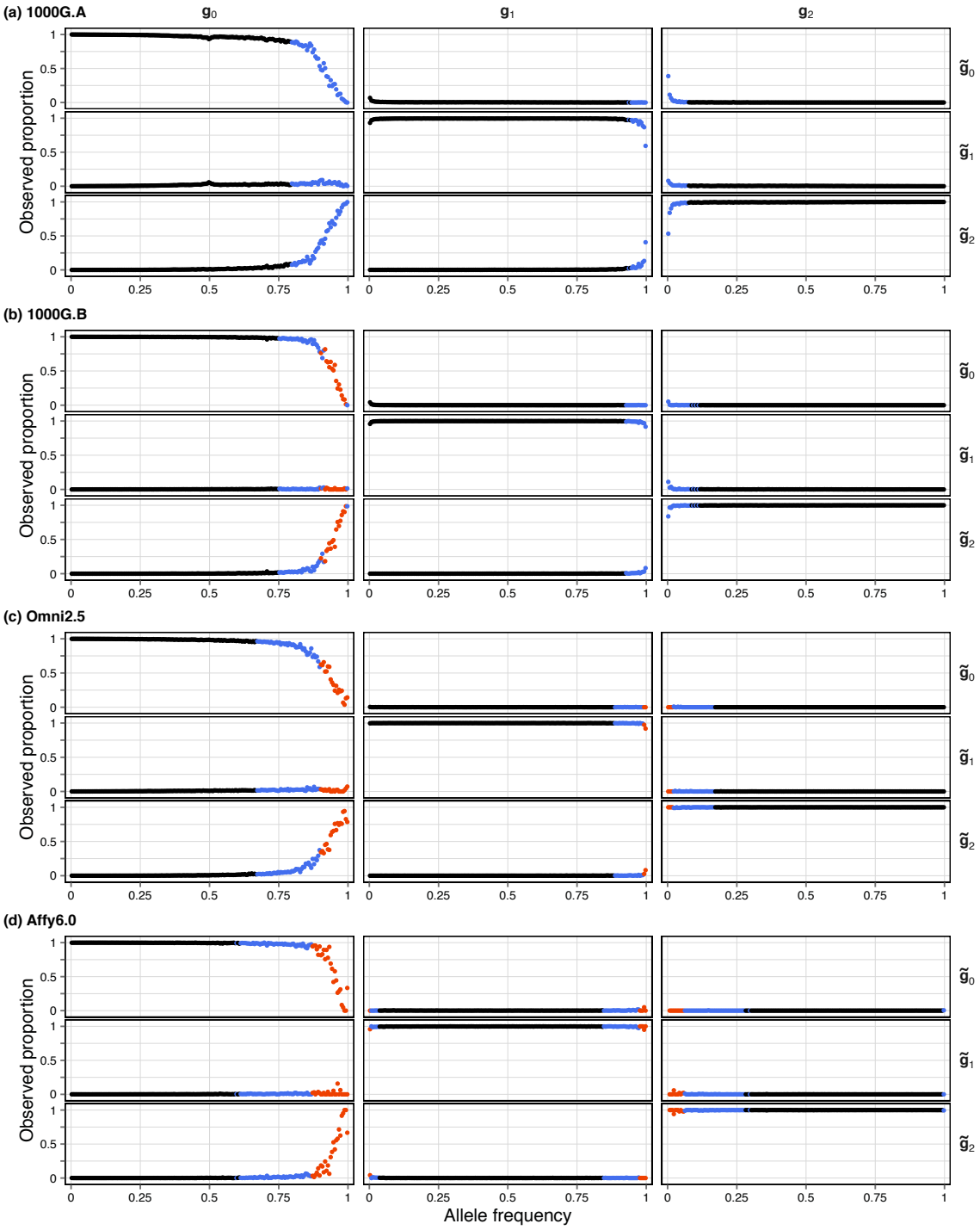


Figure 4.5: Frequency-dependent distribution of genotype penetrance in sequencing and genotyping data. For each true genotype class (columns), the fraction of g_j observed as $\tilde{g}_i$ (rows) was calculated per allele frequency bin, to estimate the frequency-dependent distribution of genotype penetrance 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.

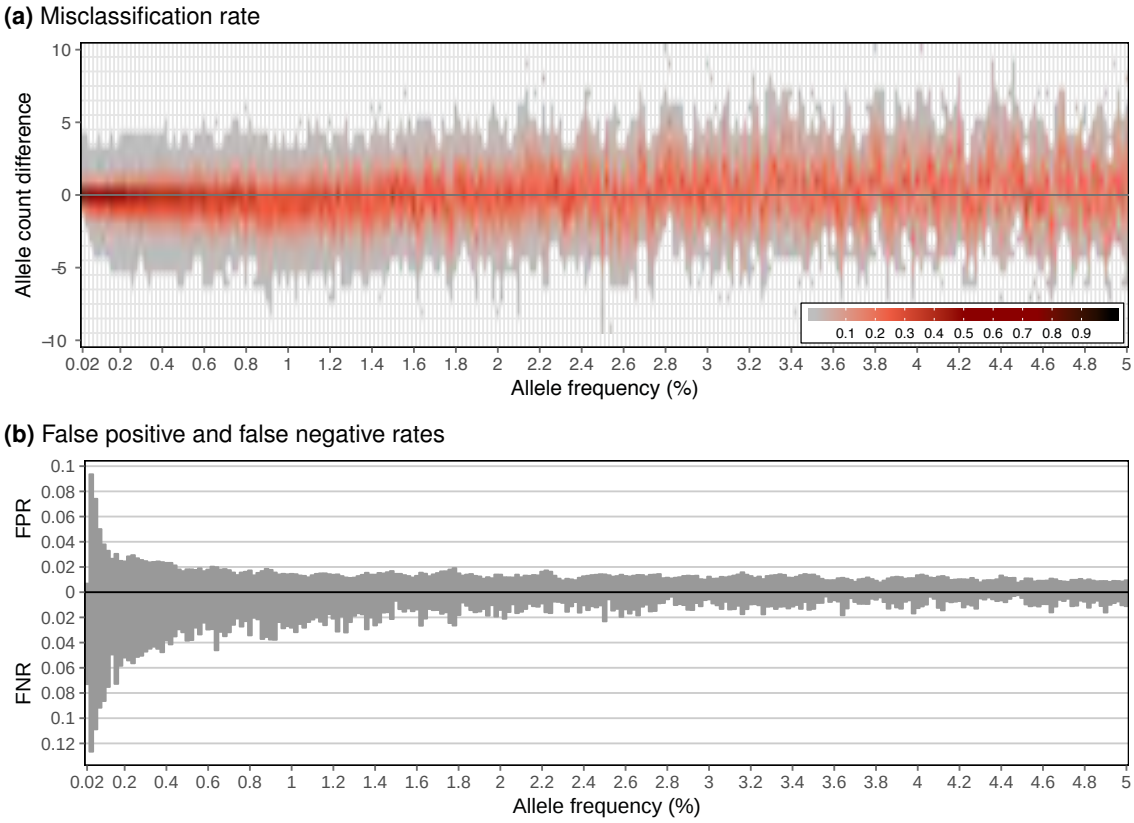


Figure 4.6: Misclassification of target sites in presence of genotype error. Simulated data were modified such that realistic distributions of genotype error were induced. Panel **(a)** indicates the rate at which alleles were observed at different frequencies after the inclusion of error. The proportion 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.

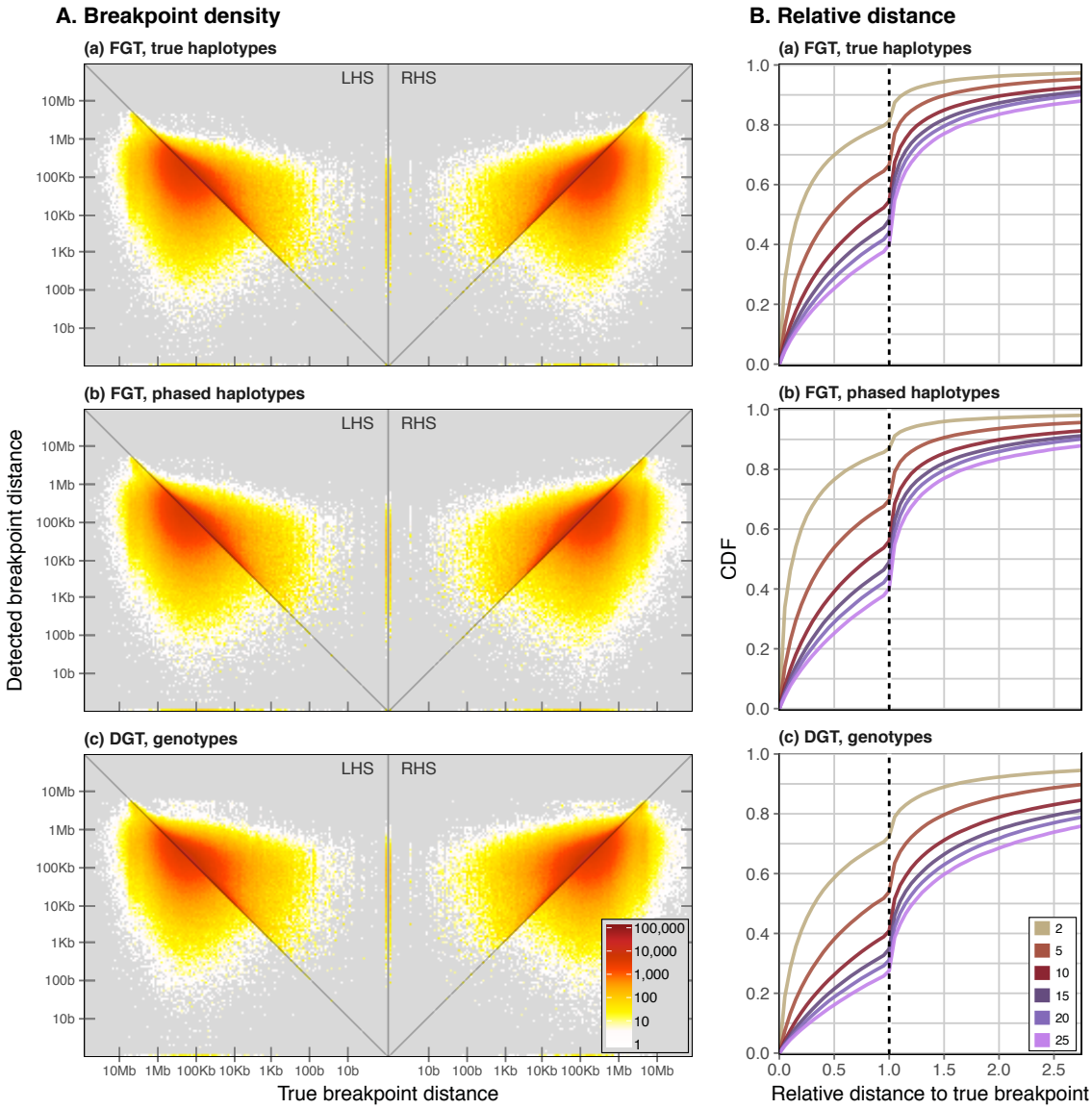


Figure 4.7: Accuracy of IBD detection using *tidy* after inclusion of genotype error. Simulated data after the inclusion of error was analysed using the FGT on true haplotypes (a), phased haplotypes (b), and the DGT on genotype data (c). Panel **(A)** shows the density of true and detected breakpoints 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 $k \in \{2, 5, 10, 15, 20, 25\}$.

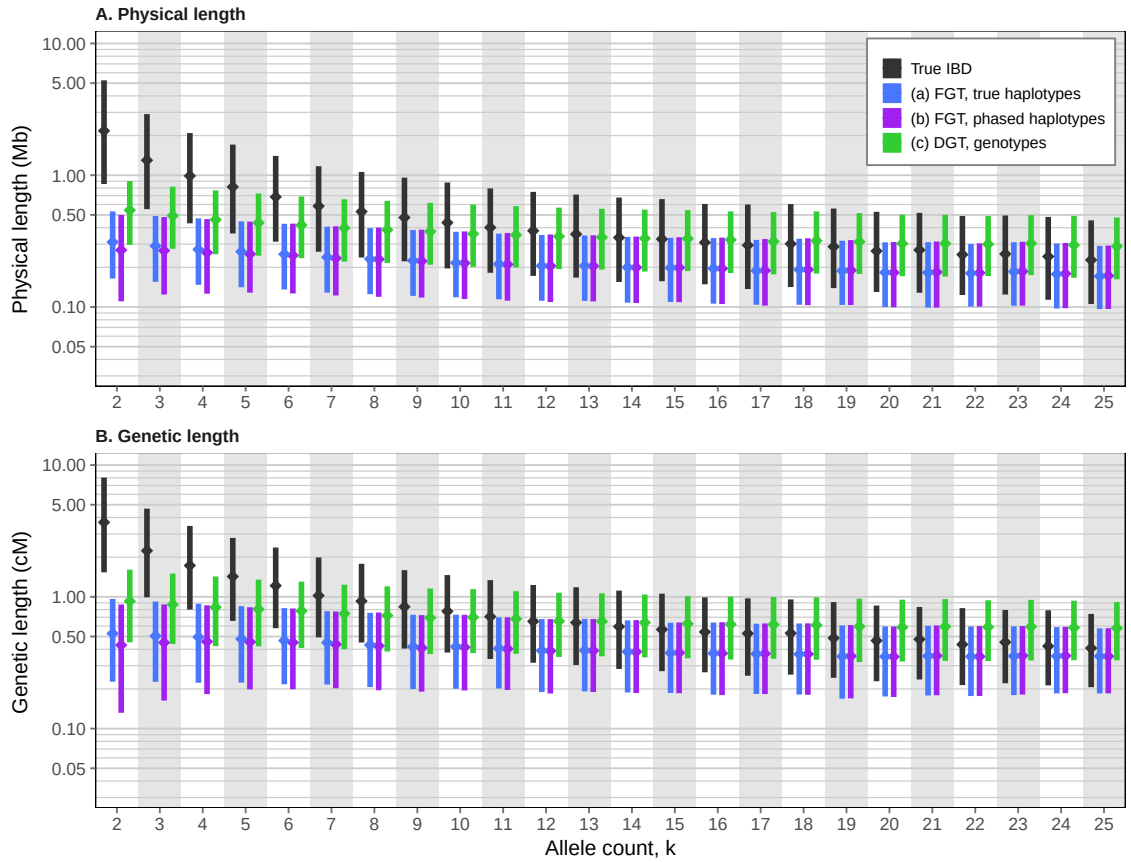


Figure 4.8: Length distribution of IBD segments using *tidy* after includion of genotype error. The distribution of physical (A) and genetic (B) segment length is shown by allele count (f_k category). Results are shown for three approaches; (a) FGT on true haplotypes, (b) FGT on phased haplotypes, and (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.

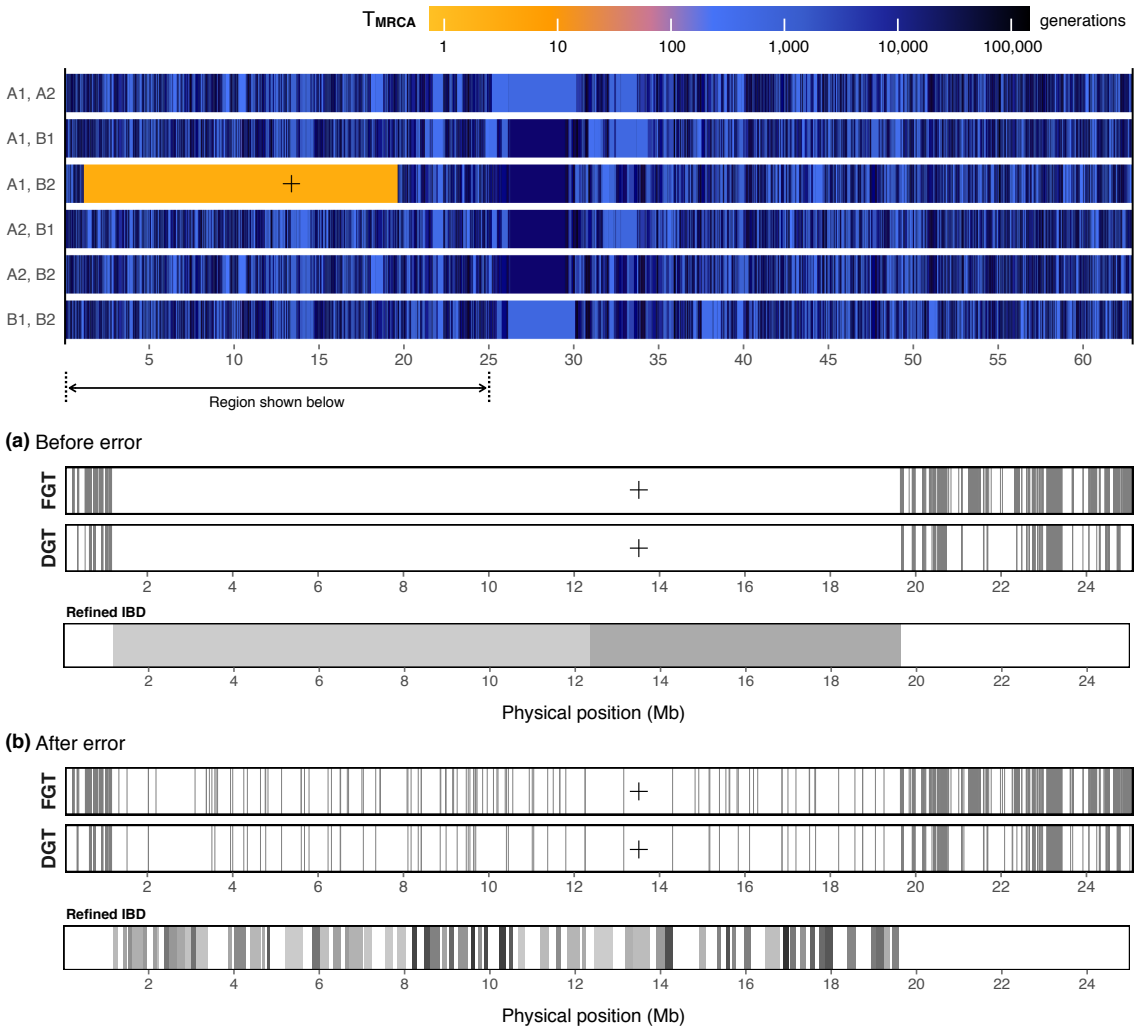


Figure 4.9: Example of the effect of genotype error on IBD detection. One allele and a pair of individuals sharing that allele were randomly selected and the underlying IBD structure for all six possible pairs of the four chromosomes was determined from simulation records (*top*). The figure shows the "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]

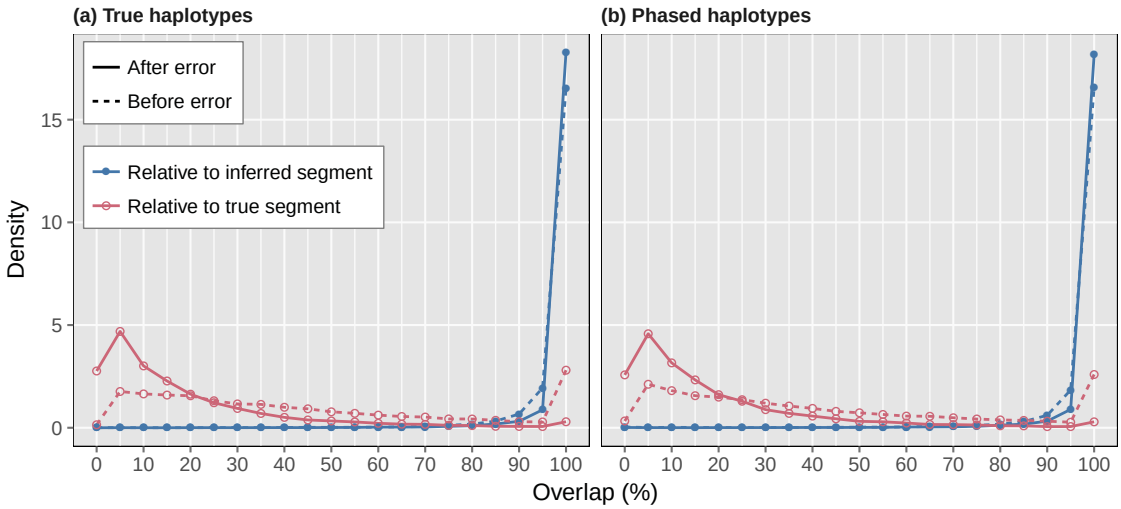


Figure 4.10: IBD segment overlap inferred using *Refined IBD* after integration of error. The proportion of overlap was measured by aligning each inferred IBD segment to the set of unique true segments determined for a given pair. Interval comparisons with zero overlap were ignored. The results shown 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.

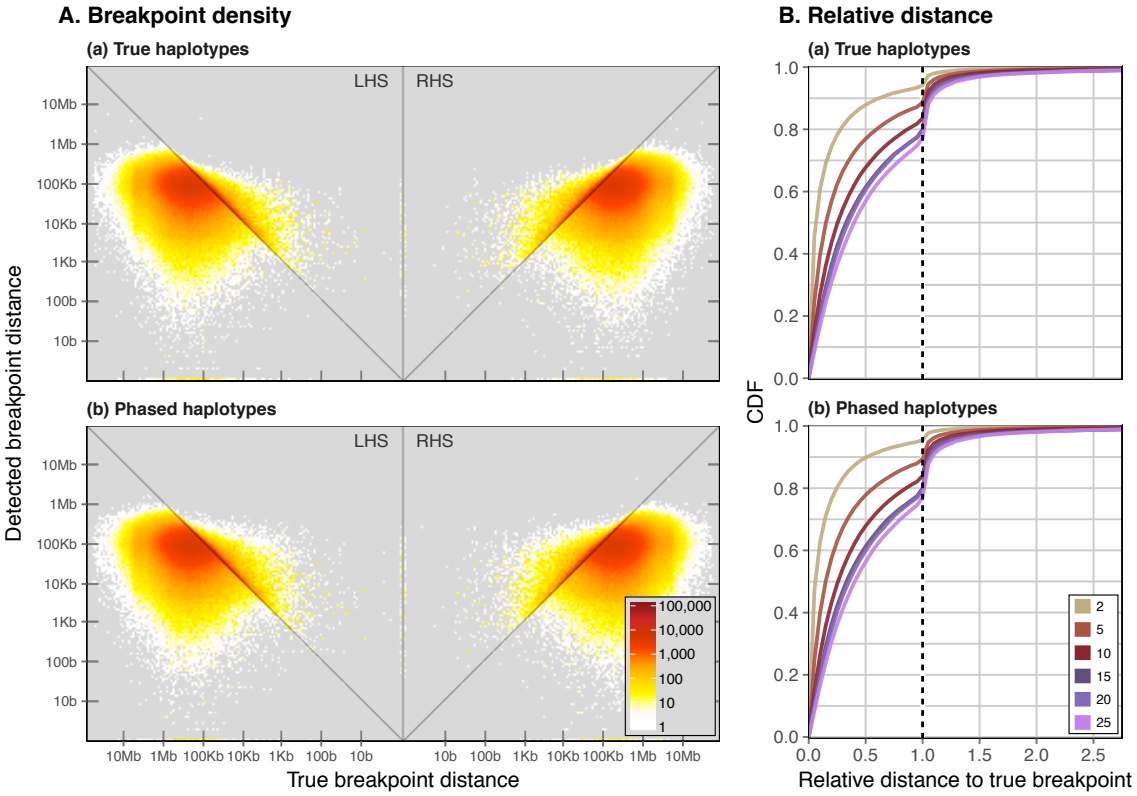


Figure 4.11: Accuracy of IBD detection using *Refined IBD* after integration of error. Panel **(A)** shows the density of true and detected breakpoint distance to the matched target position. Panel **(B)** shows the cumulative distribution function (CDF) of the relative distance by f_k category where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

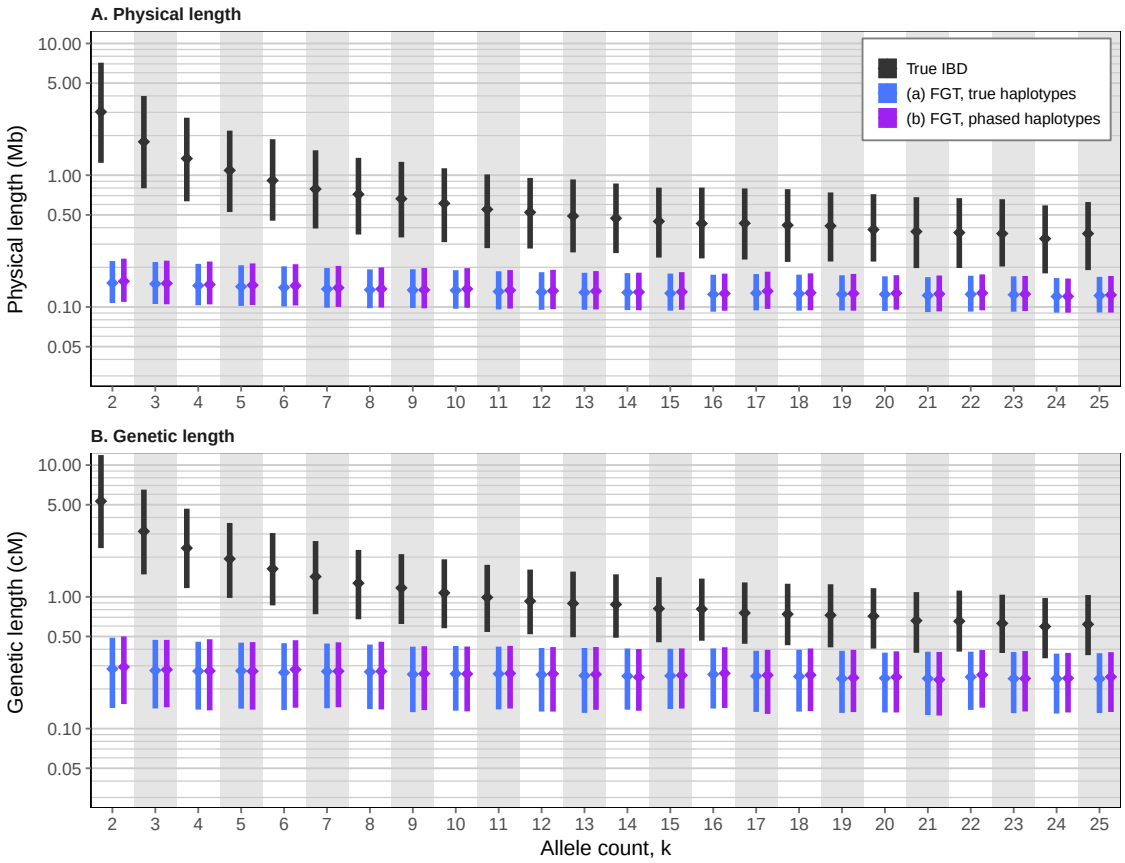


Figure 4.12: IBD length detected using *Refined IBD* after integration of error. The distribution of physical (A) and genetic (B) segment length is shown by allele count (f_k category). Results were obtained using `Refined IBD` in `Beagle 4.1`, on true and phased haplotype data; *i.e.* Approaches (a) 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 the 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.

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

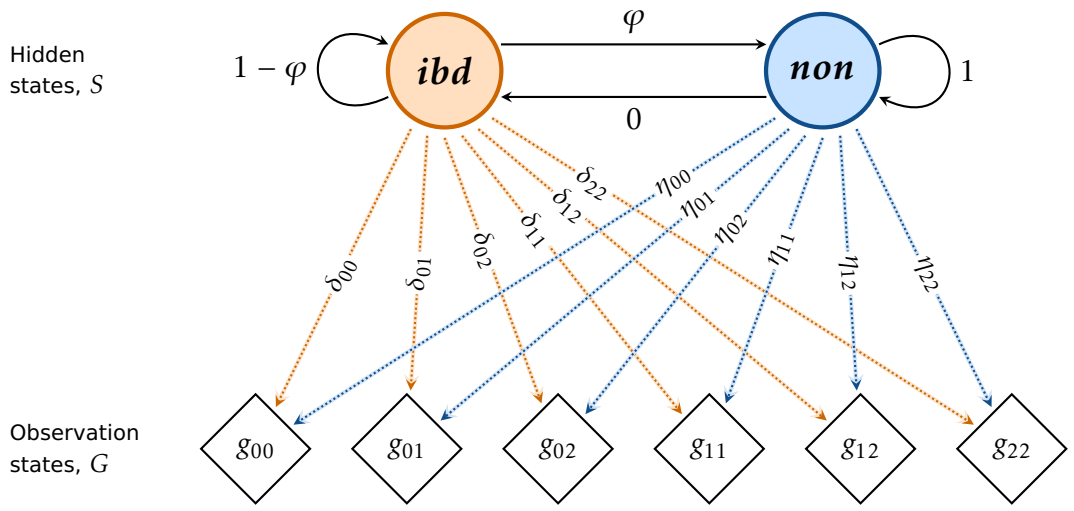


Figure 4.13: Illustration of the Hidden Markov Model for IBD inference. Two hidden states are assumed to generate the observations in a Markov process; *ibd* and *non*. Transitions from each state into any state are indicated by *solid* lines. The probability of transition from *ibd* to *non* is denoted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

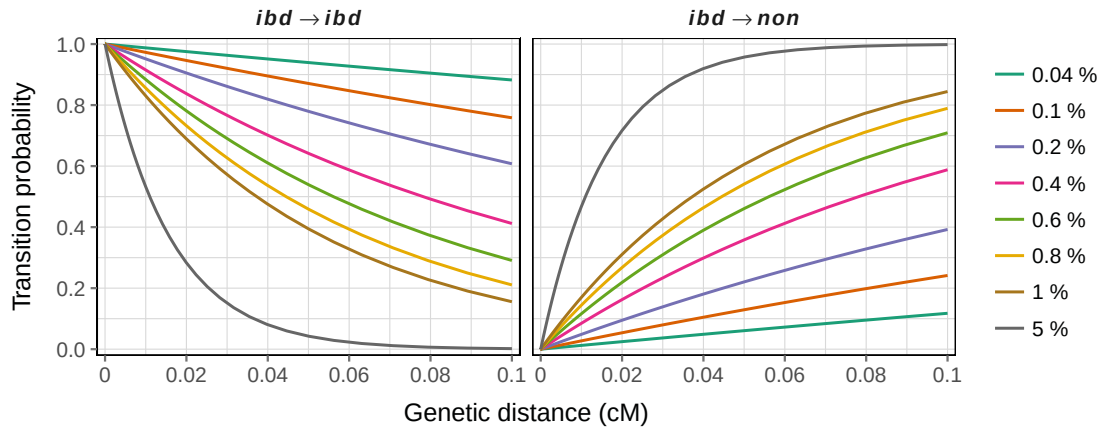


Figure 4.14: Probability distribution of transition dependent on IBD. The probability of transition was modelled dependent on the genetic distance between a particular site and the previous site and the expected age of the focal allele. The frequency of the focal allele determines its expected age, which 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 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*.

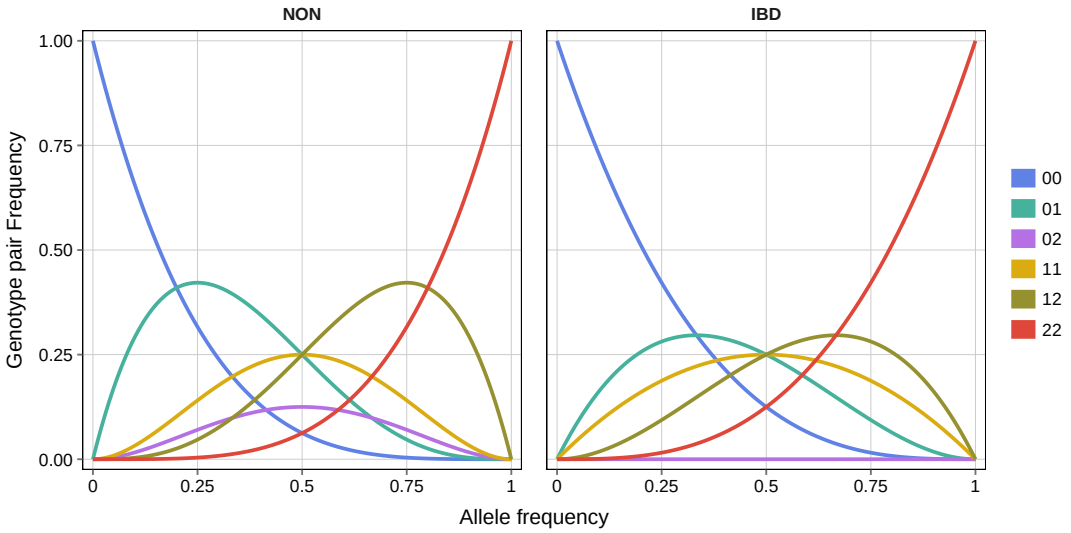


Figure 4.15: Expected frequency distribution of genotype pairs under non-IBD and IBD. Proportions were calculated using Equation (4.8) and Equation (4.9) in both hidden states, *non* and *ibd*, respectively. Colours distinguish the six possible genotype pairs, given by g_{k_1,k_2}, as indicated.

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

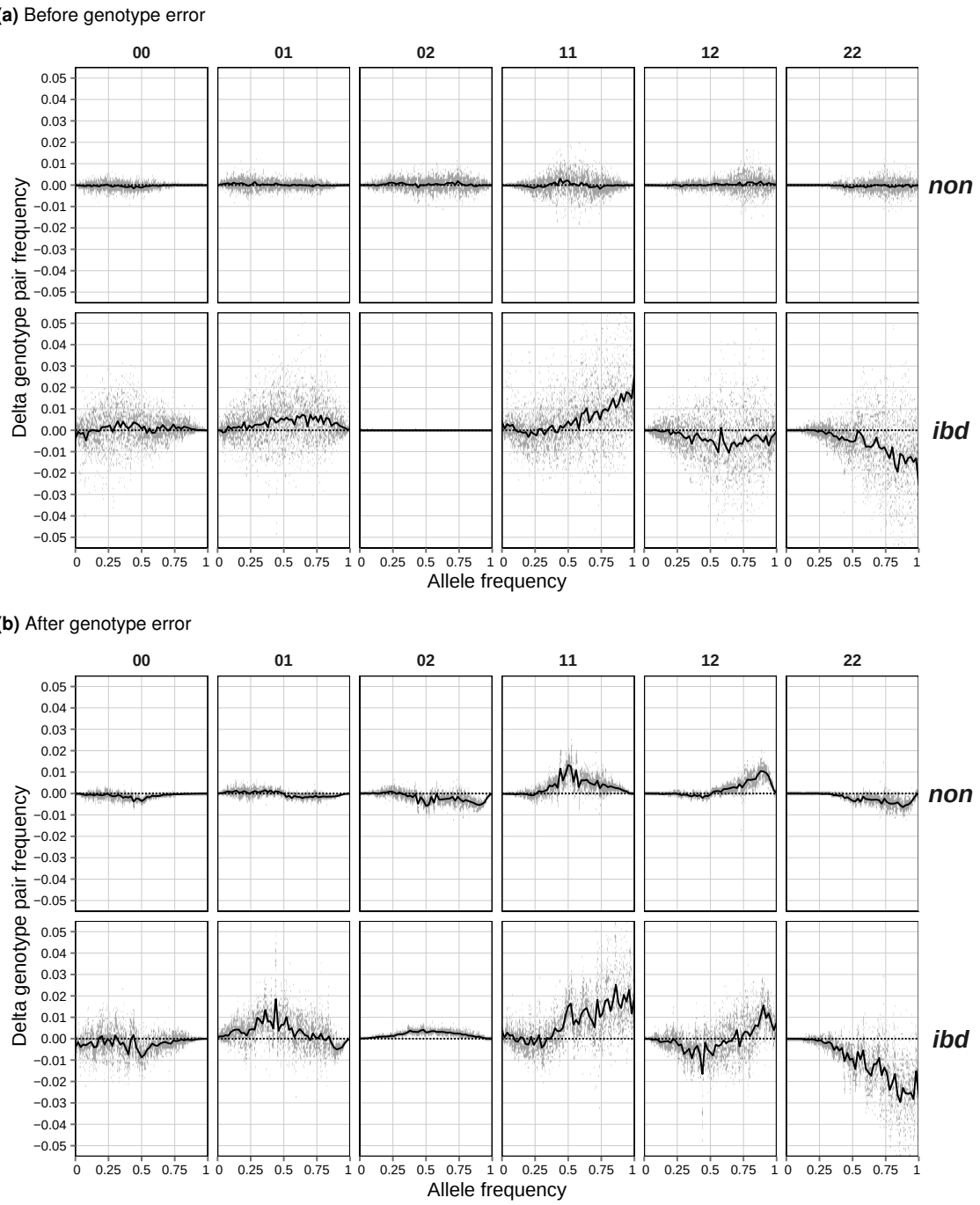


Figure 4.16: Difference between empirical and expected proportions of genotype pairs. In total, 500,000 segments were sampled in *non* and *ibd* as determined from coalescent records. Segments were aggregated by allele frequency to calculate empirical proportions for each of the six possible genotype 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.

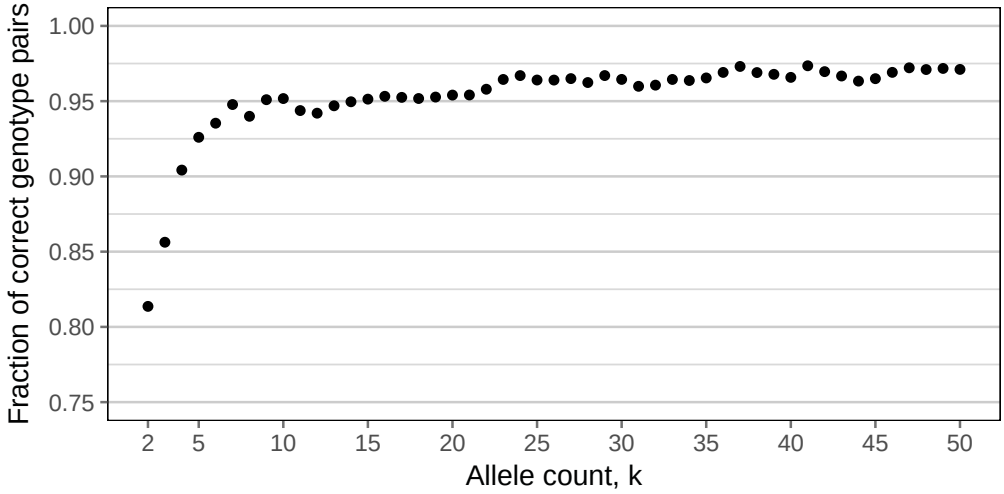


Figure 4.17: True positive rate of identified genotype pairs at focal sites. Pairwise shared genotypes at focal f_k variants with $k > 1$ were compared between datasets before and after error. The true positive rate was determined for each k. Results are shown for k in $[2, 50]$, which corresponds to 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 (5.27), page 198. 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.

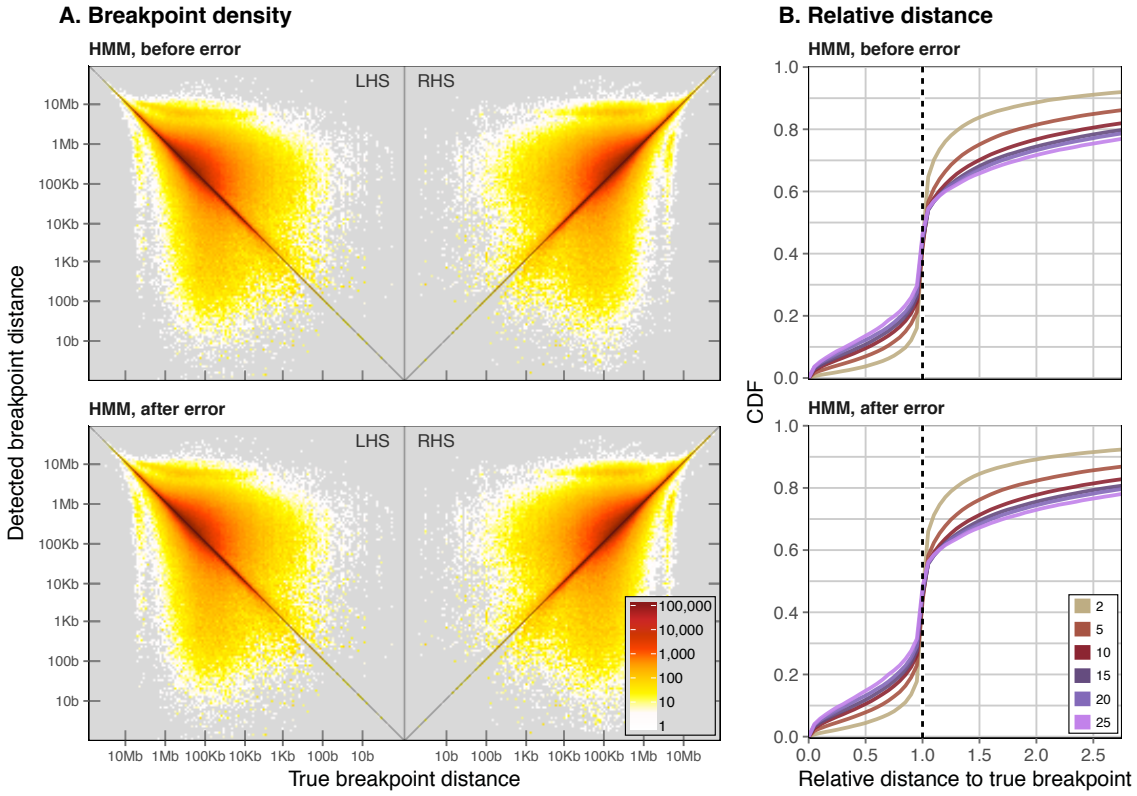


Figure 4.18: Accuracy of breakpoint detection using the HMM on simulated data before and after error. Panel **(A)** shows the density in terms of the physical distance between true and detected breakpoints and the position of a given focal allele. Panel **(B)** shows the physical length in terms of the 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).

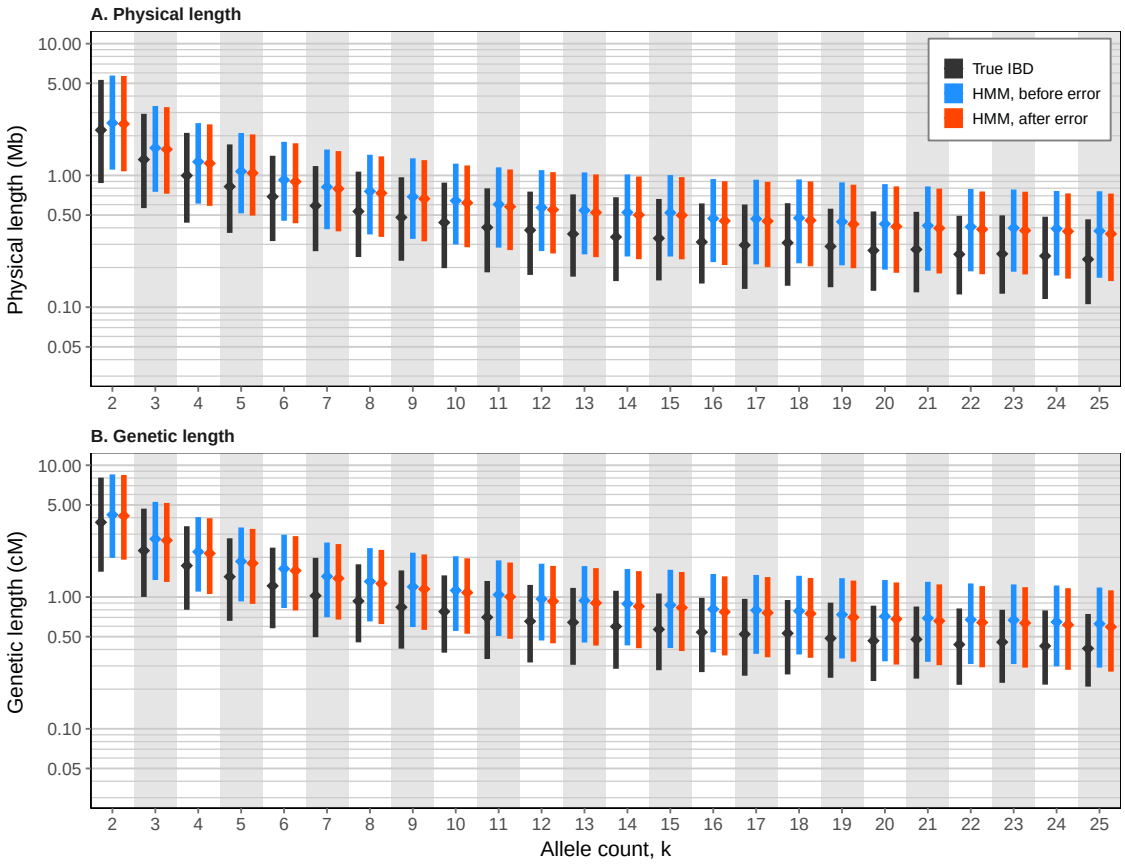


Figure 4.19: Shared haplotype lengths inferred using the HMM on simulated data before and after error. The HMM-based approach for targeted IBD detection was applied to simulated data before and after the integration of error; *i.e.* datasets D and D^*, respectively. Lengths were compared to the corresponding 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.

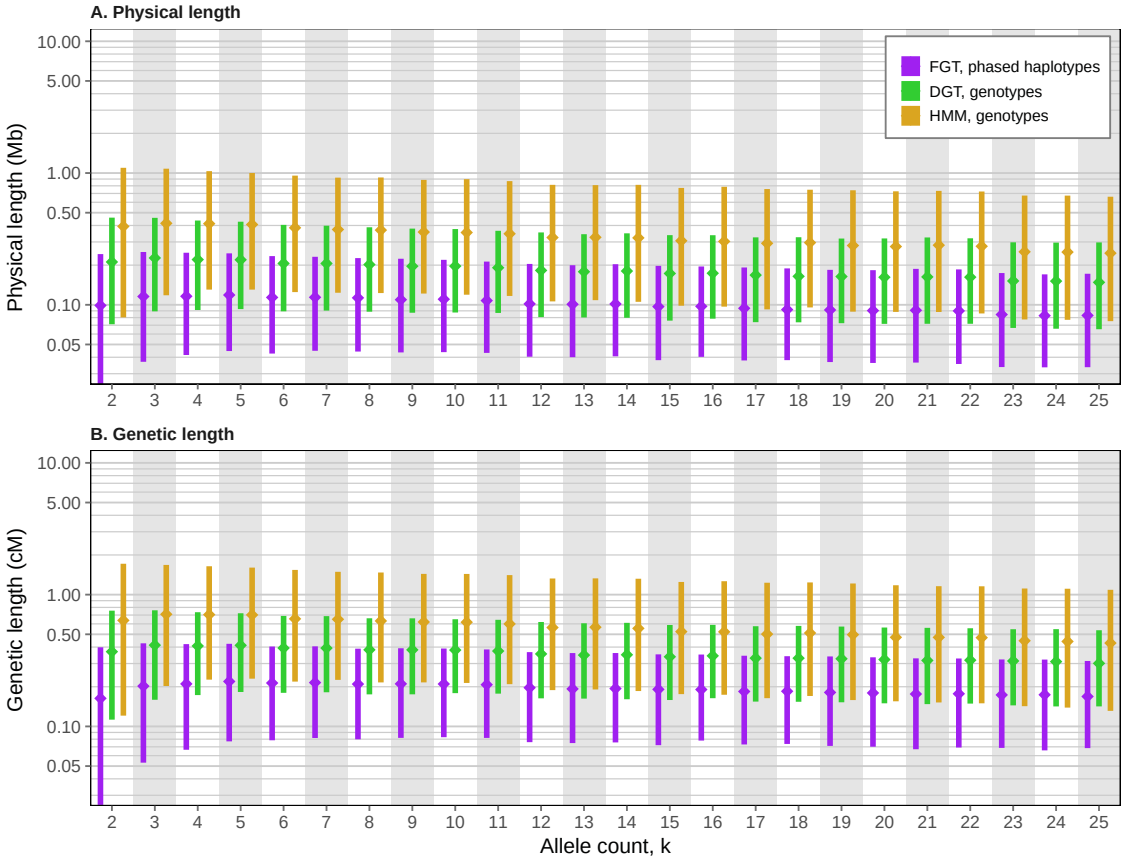


Figure 4.20: IBD inference using the Hidden Markov Model on 1000 Genomes data, chromosome 20. IBD detection using the HMM-based method was performed under the empirical error model defined on genotype data from the 1000 Genomes Project. The resulting length distribution is compared to previous results 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.

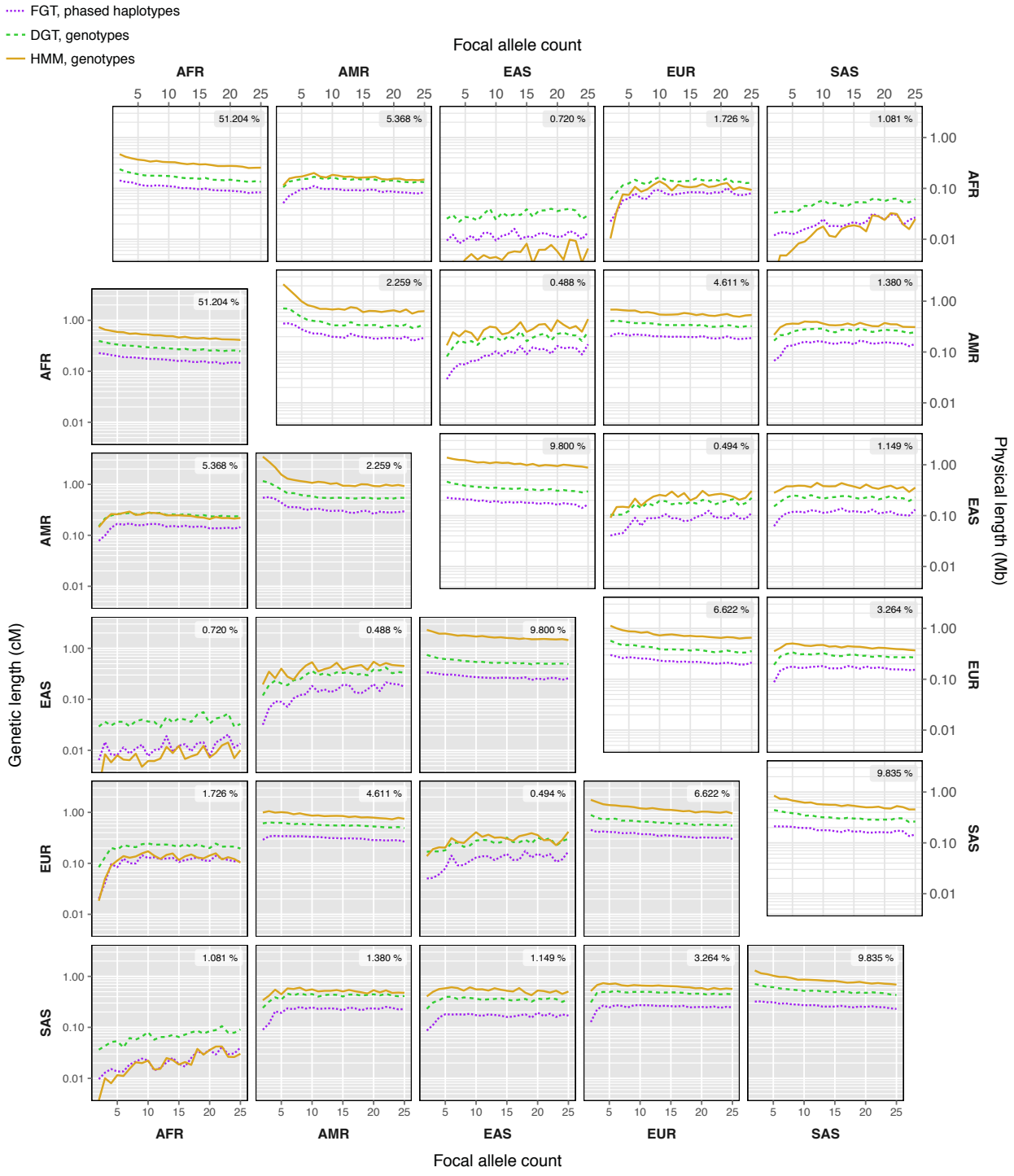


Figure 4.21: Inferred shared haplotype lengths by population in 1000 Genomes, chromosome 20. The distributions of physical (upper triangle) and genetic (lower triangle) length by frequency of the focal allele are shown for alleles shared between pairs of individuals from the same population (*diagonal* 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, chromosome 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; for example, IBD segments identified by f_2 variants were overall higher in accuracy compared to f_{15} or higher

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

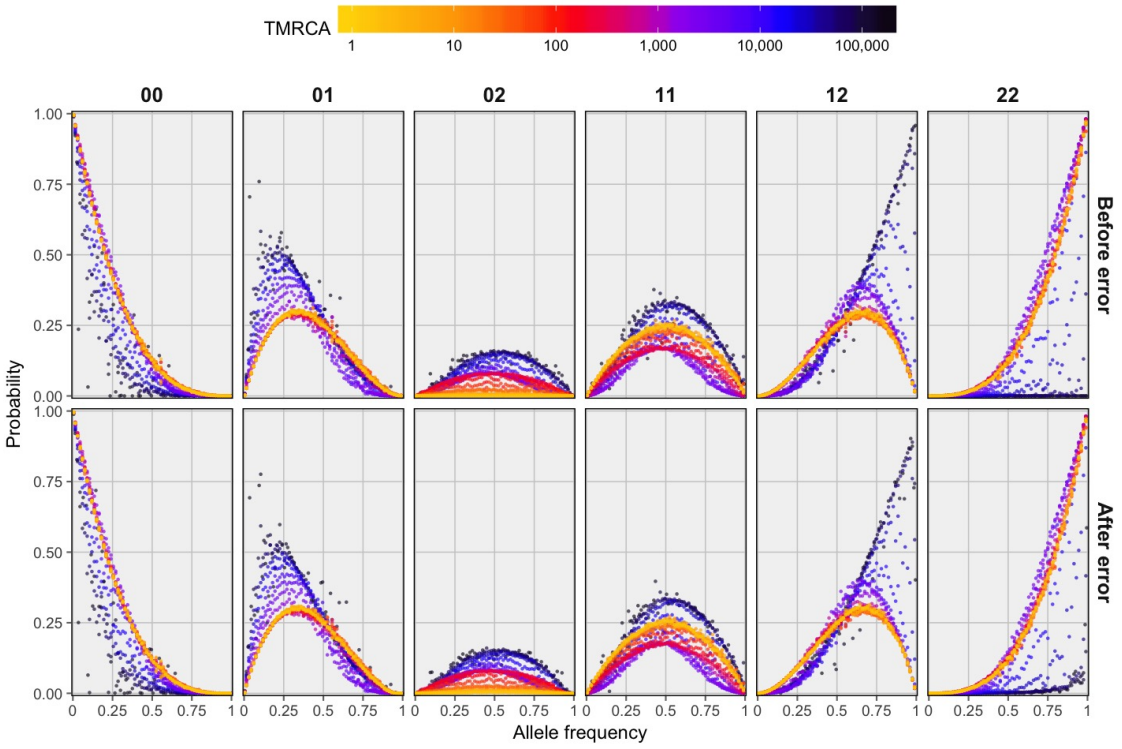


Figure 4.22: Empirical emission probabilities of genotype pairs observed at different T_{MRCA}. The relative proportions of genotype pairs observed in IBD segments are distinguished by time to the most recent common ancestor (T_{MRCA}) for a given pair of haplotypes, which was derived from simulation 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%.

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.

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.

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

— Elon Musk

Abbreviations

1000G	1000 Genomes Project
CDF	Cumulative distribution function
DGT	Discordant genotype test
FGT	Four-gamete test
FNR	False negative rate
FPR	False positive rate
GWA	Genome-wide association
HMM	Hidden Markov Model
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IPG	Illumina Platinum Genomes Project
NGS	Next-generation sequencing
PCR	Polymerase chain reaction
RMSLE	Root mean squared logarithmic error
SNP	Single-nucleotide polymorphism
$\mathbf{T_{MRCA}}$	Time to the most recent common ancestor
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., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Angela Center, Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., and Majoros... (2001). The Sequence of the Human Genome. *Science*, **291**(5507), 1304–1351.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

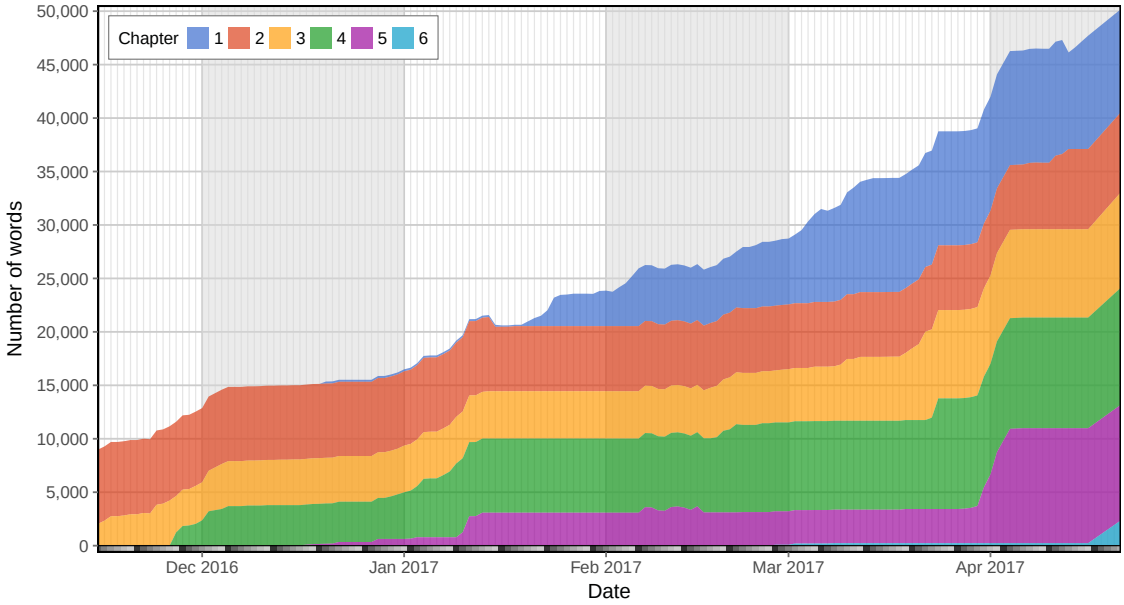
Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, **16**(2), 97–159.

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.

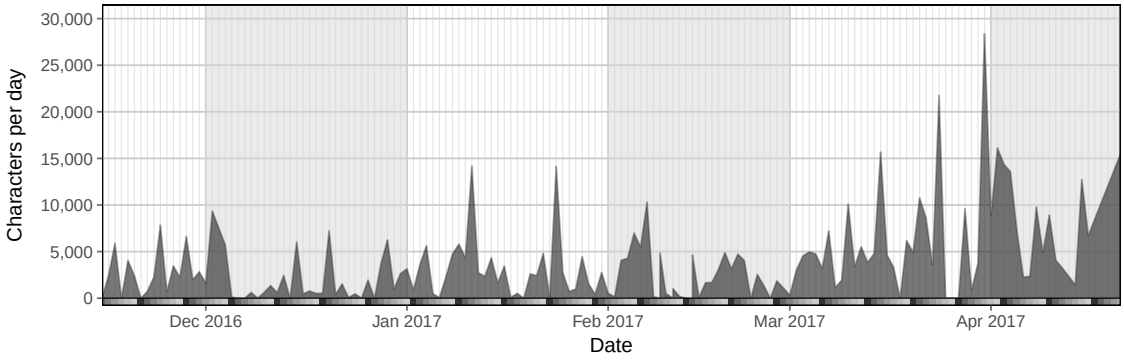
Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, **111**(4), E455–64.

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

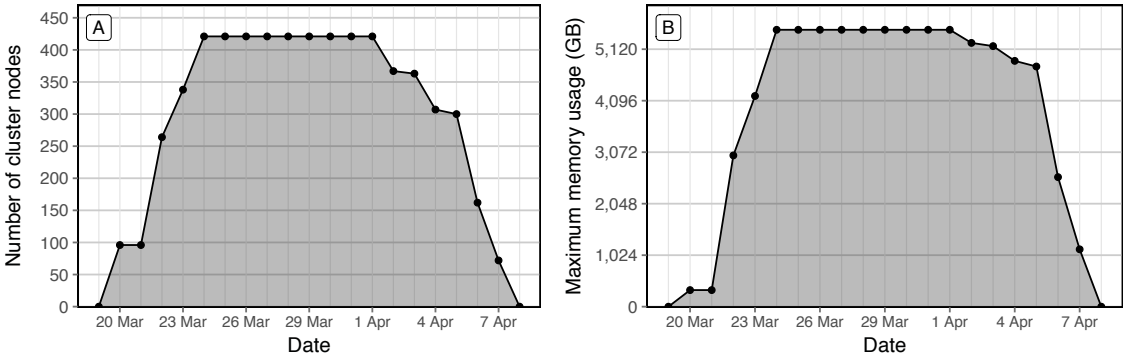
— Adam Savage



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 the **(A)** number of nodes used and **(B)** daily maximum of computer memory on the cluster of the Wellcome Trust Centre for Human Genetics.