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

— Lewis Carroll, *Alice in Wonderland*

3

Using rare variants to detect haplotype sharing and identity by descent

Contents

3.1 Introduction

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

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

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

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

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

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

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

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

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

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

3.2 Rare variants as indicators of haplotype sharing by descent

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

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

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

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

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

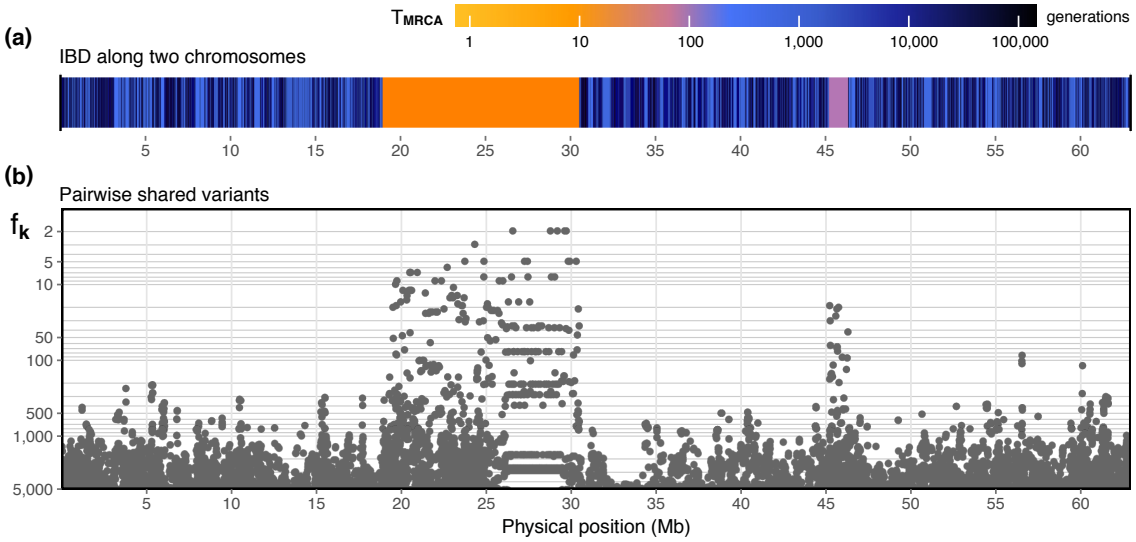
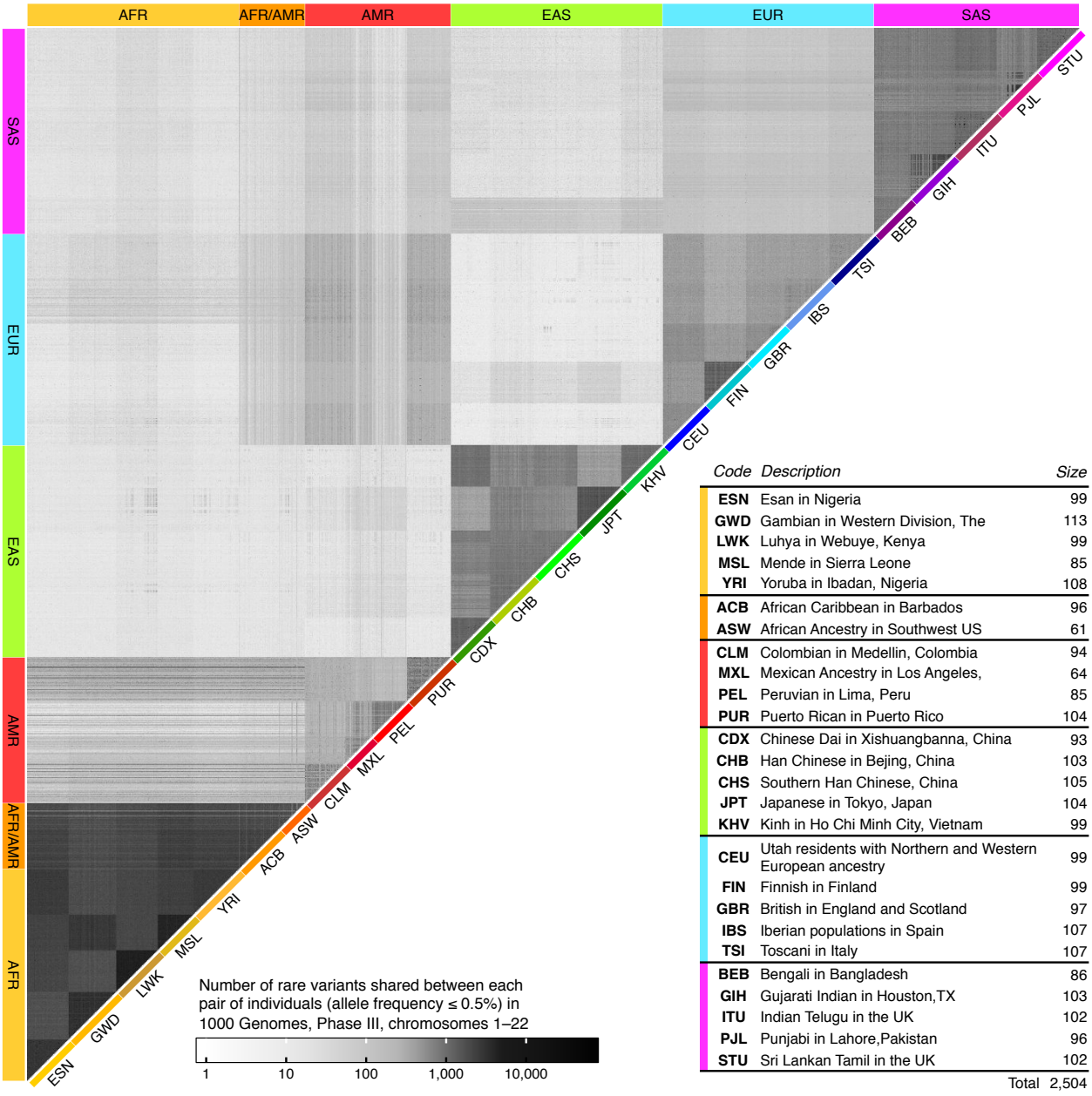


Figure 3.1: IBD structure and pairwise variant sharing. A dataset of N = 5,000 haplotypes was simulated under the coalescent using `msprime` (Kelleher *et al.*, 2016). IBD status was determined from simulated genealogies for a pair of chromosomes selected at random from the set of chromosomes that shared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

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

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



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

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

3.3 IBD detection around rare variants

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

3.3.1 Inference of historical recombination events

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

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

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

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

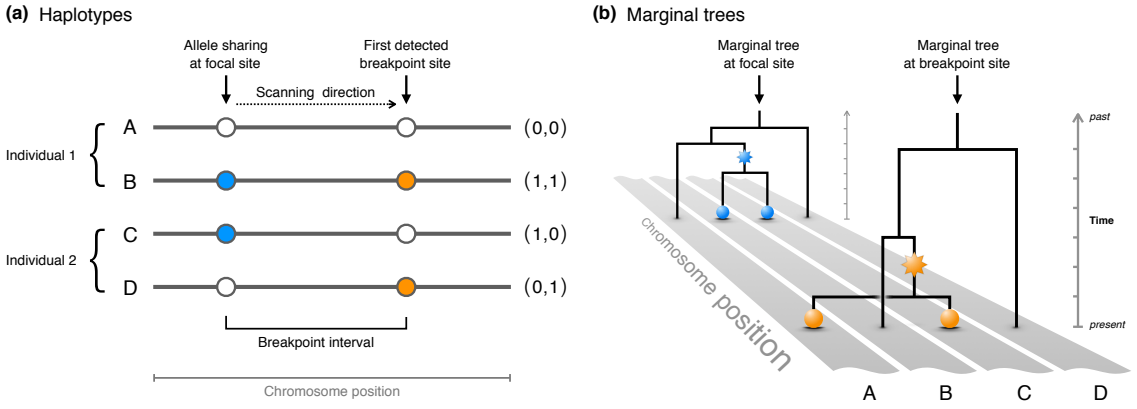


Figure 3.3: Breakpoint detection using the four-gamete test (FGT). Panel **(a)** shows the four haplotypes (gametes) in a pair of two diploid individuals (*horizontal lines*). The focal allele (*blue*) on haplotypes B and C is shared by both individuals. A breakpoint interval is detected if all four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

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

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

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

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

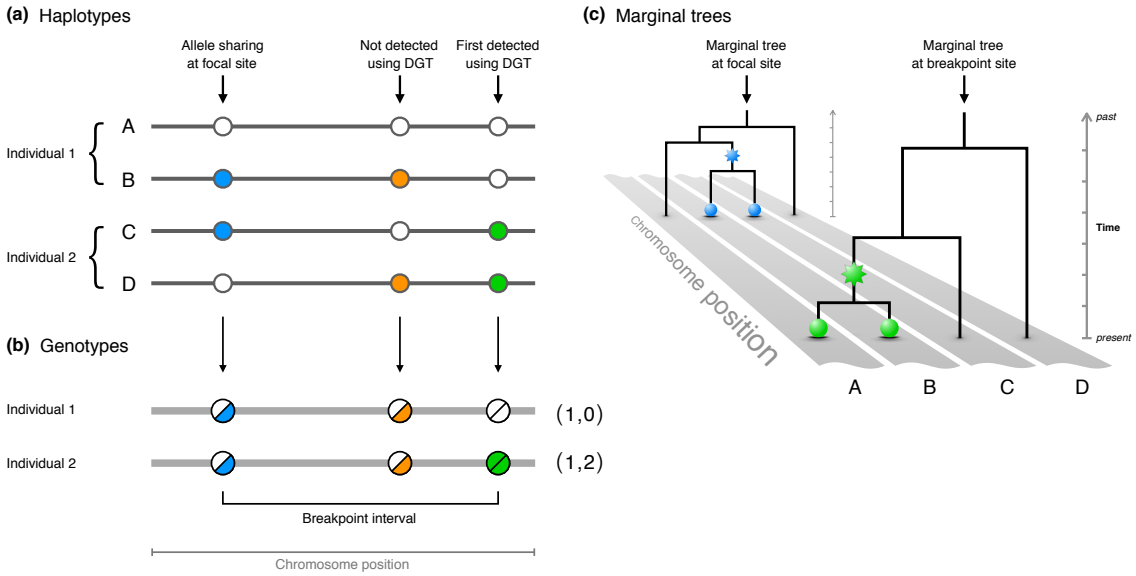


Figure 3.4: Breakpoint detection using the discordant genotype test (DGT). Unlike the FGT, which requires haplotype information, the DGT identifies a breakpoint interval using genotype data. This representation extends the example shown for the FGT in Figure 3.3 (page 82). For comparison, Panel **(a)**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

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

3.3.2 Description of the algorithm

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

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

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

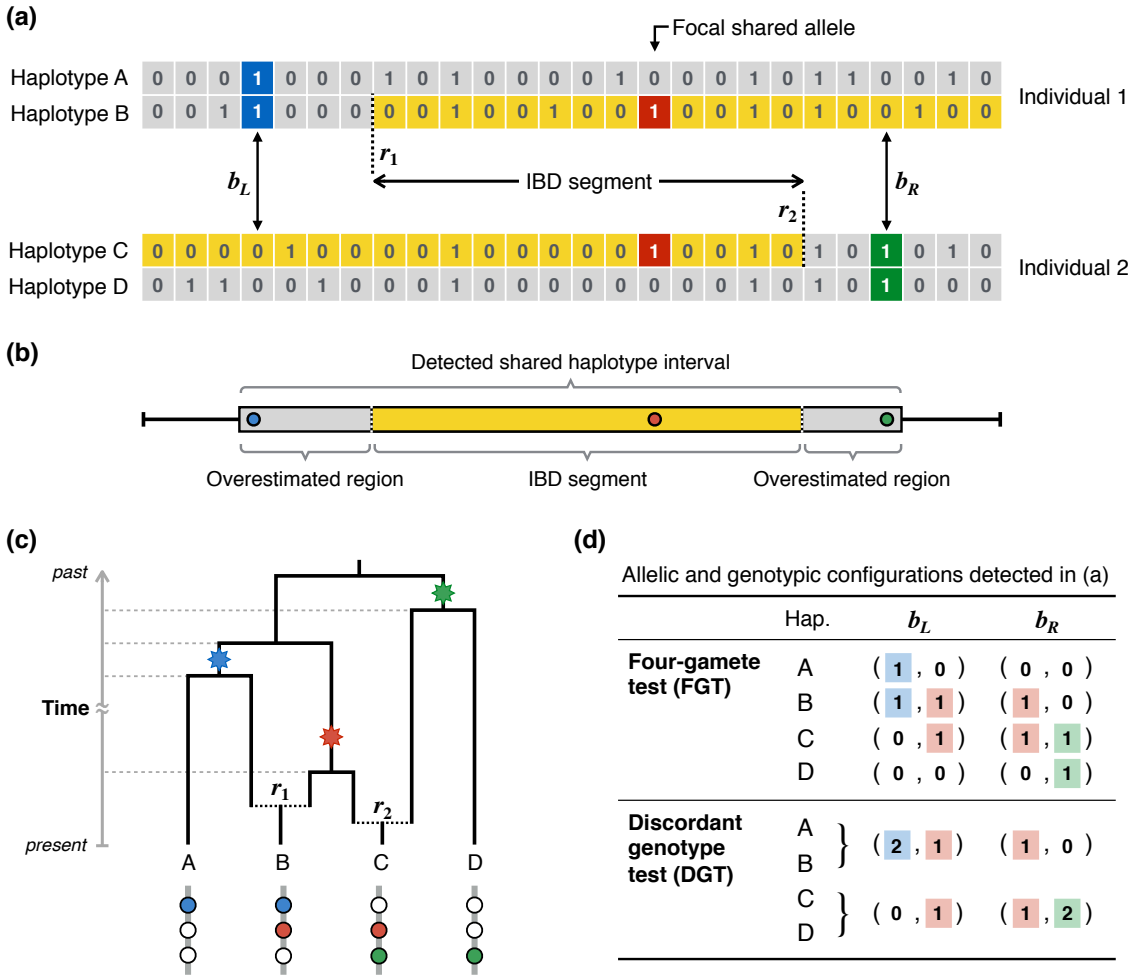


Figure 3.5: Illustration of shared haplotype detection in a pair of diploid individuals. Panel **(a)** shows two individuals composed of haplotypes A and B, and haplotypes C and D, respectively. Each haplotype is represented as a sequence of observed allelic states, where 0 and 1 denote the ancestral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

3.3.3 Anticipated limitations

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

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

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

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

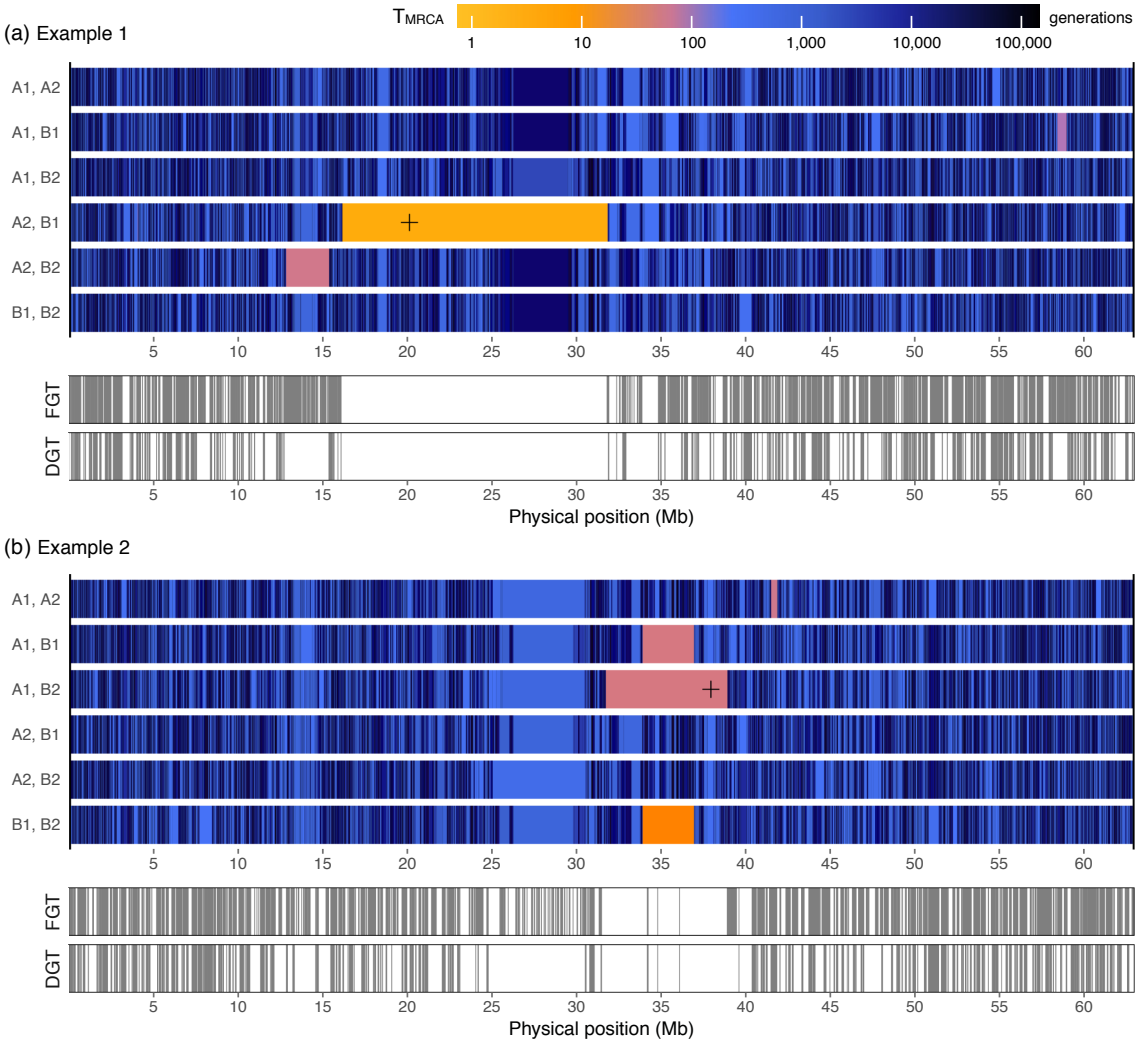


Figure 3.6: Examples of the underlying IBD structure in each pair of four chromosomes. The true, underlying IBD structure is shown for each possible pair among four chromosomes in two diploid individuals; two examples are shown. Each chromosome is labelled by its occurrence in individuals A and B, where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

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

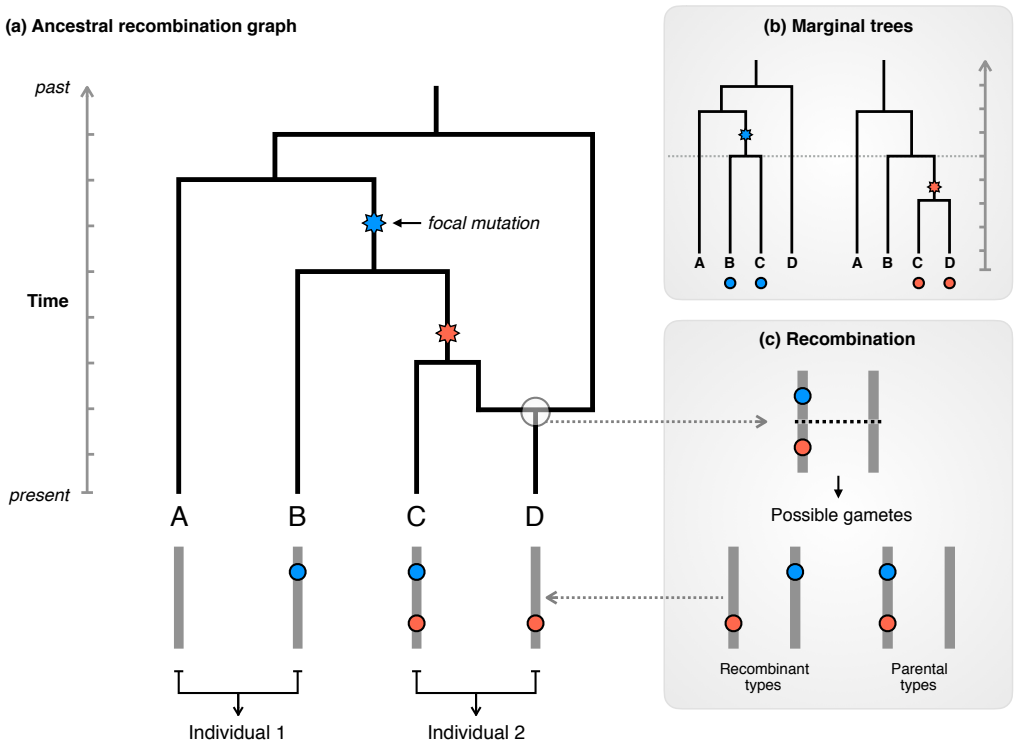


Figure 3.7: Recombination outside the focal sub-tree. The example shows two diploid individuals that share a focal allele (*blue*) on haplotypes *B* and *C*, which form a sub-tree within the larger genealogy. Panel **(a)** shows the ancestral recombination graph (ARG) of a possible recombination history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

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

3.4 Evaluation

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

3.4.1 Data generation

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

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

3.4.1.1 Demographic model

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

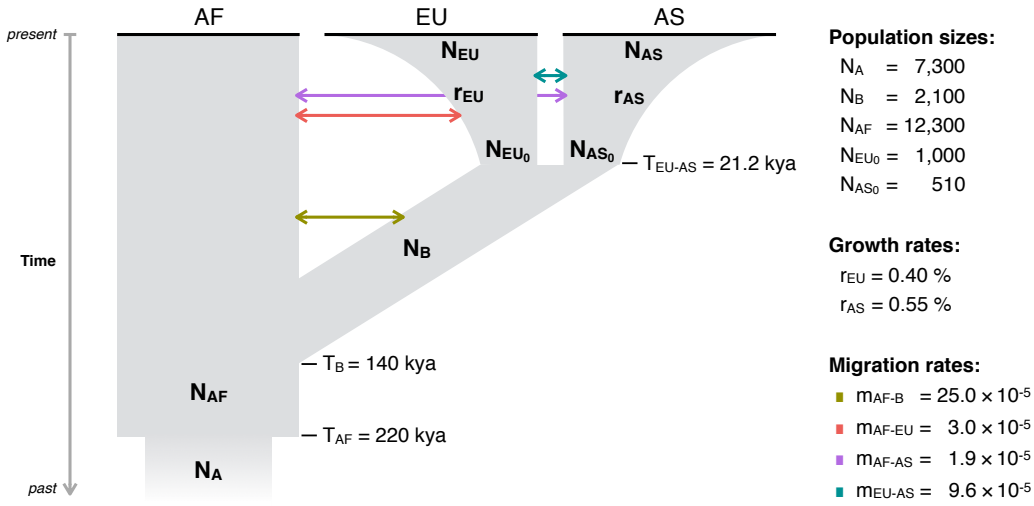


Figure 3.8: Demographic model used in simulations. Three populations were modelled, African (AF), European (EU), and Asian (AS), which derive from an ancestral population (A). Both EU and AS experienced a bottleneck with subsequent exponential growth following the out-of-Africa expansion of a founder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

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

3.4.1.2 Simulated dataset

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

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

3.4.2 Accuracy analysis

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

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

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

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

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

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

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

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

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

(b) FGT on *phased* haplotypes, and

(c) DGT on genotype data.

3.5 Results

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

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

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

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

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

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

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

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

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

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

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

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

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

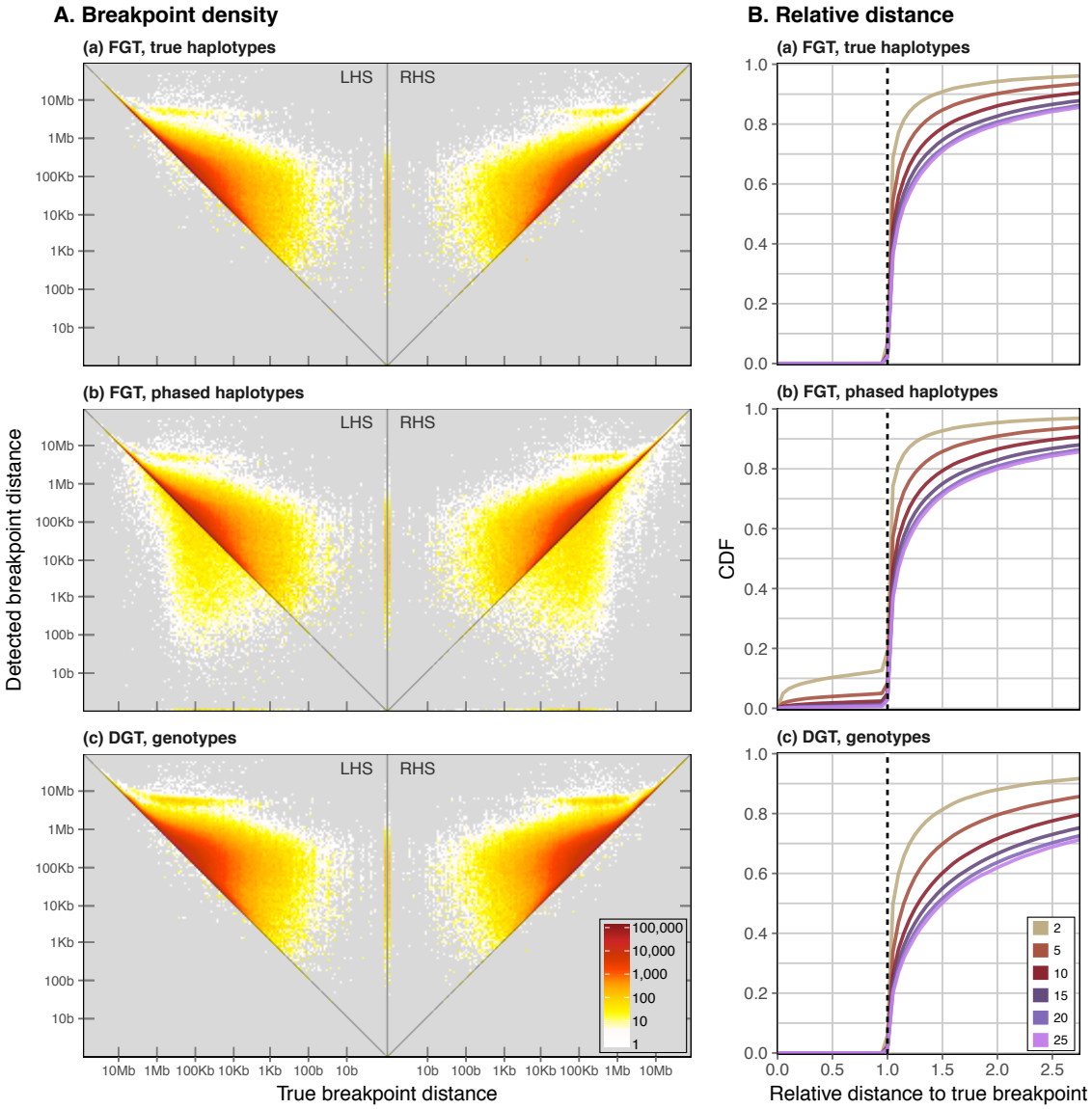


Figure 3.9: Accuracy of breakpoint detection in simulated data. Breakpoints detected in f_k pairs at $k \in \{2, \ldots, 25\}$ are compared to true IBD breakpoint sites, after removing boundary cases in either the detected or true dataset. Segments were inferred using the FGT on true haplotypes **(a)**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

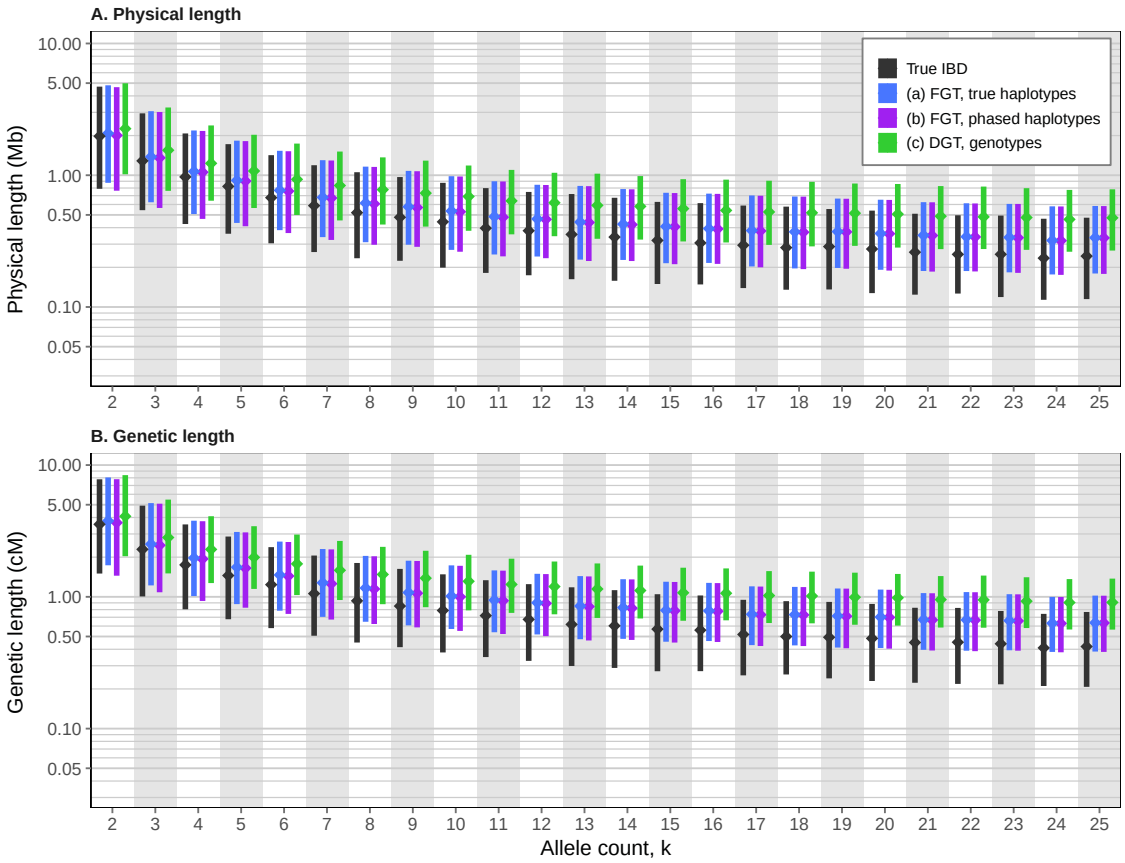


Figure 3.10: IBD segment lengths inferred in simulated data. The distribution of median physical and genetic length of detected IBD segments is shown by allele frequency of the focal variant ($f_{[2,25]}$). IBD detection was performed using the FGT on true and phased haplotypes, as well as the DGT on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

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

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

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

3.5.0.1 IBD detection using the *Refined IBD* method

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

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

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

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

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

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

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

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

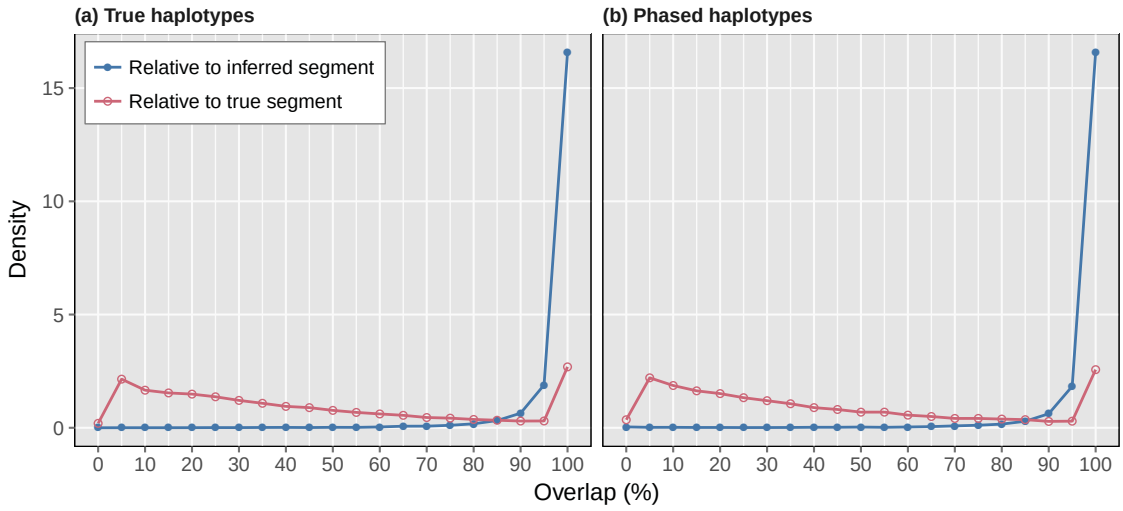


Figure 3.11: IBD segment overlap inferred using *Refined IBD* in Beagle 4.1. Each of the inferred IBD segments were aligned with each of the true segments determined for a given pair to measure the proportion of total base overlap; interval comparisons with zero overlap were ignored. The results shown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

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

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

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

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

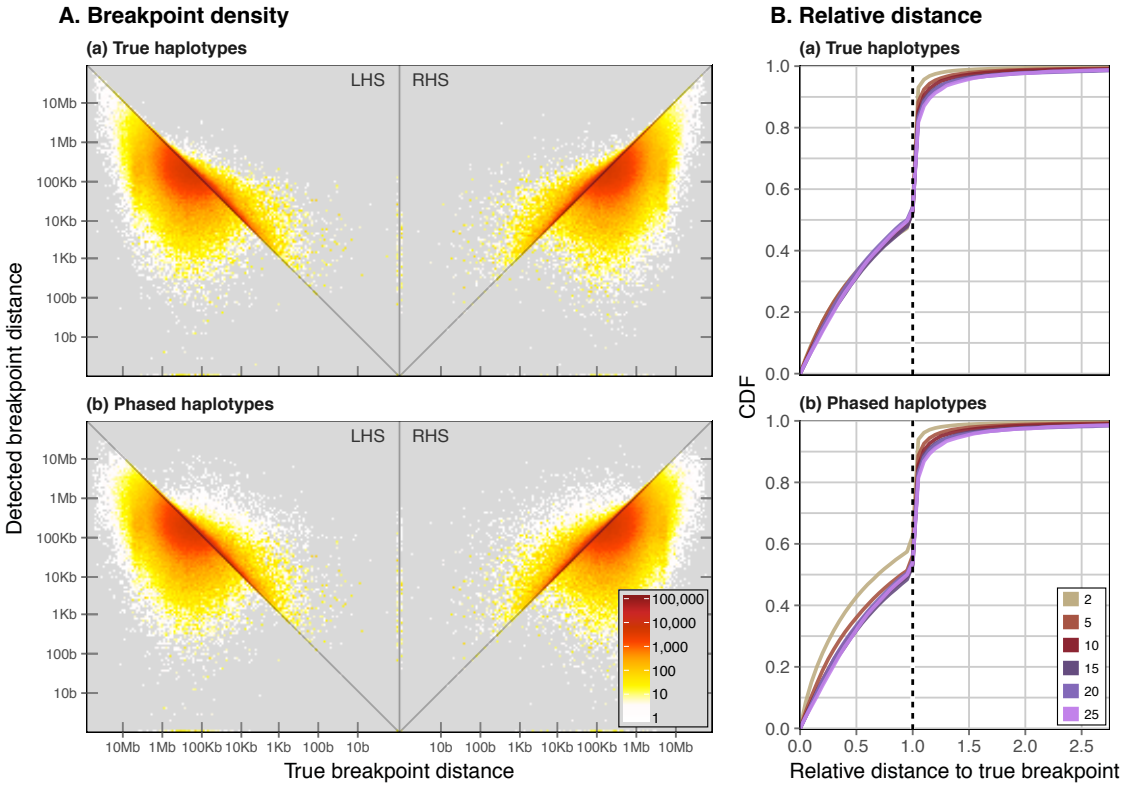


Figure 3.12: Accuracy of breakpoint detection using *Refined IBD* in Beagle 4.1. Results are shown for shared haplotype segments inferred using the `Refined IBD` method; after the detected segments were matched to the set of true segments based on a given target site being contained within the interval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

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

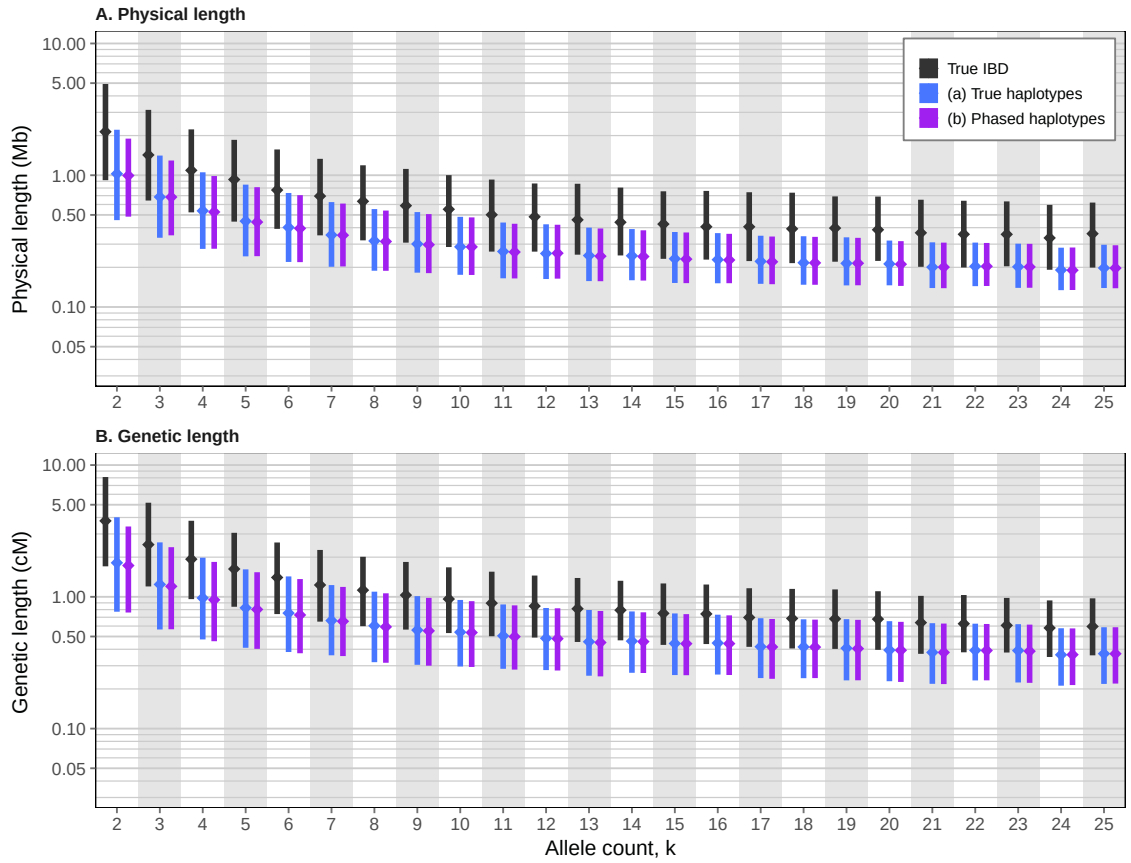


Figure 3.13: IBD segment lengths inferred using *Refined IBD* in Beagle 4.1. The distribution of median physical *(A)* and median genetic *(B)* segment length is shown by allele count (f_k category). IBD segments were inferred using the `Refined IBD` algorithm implemented in `Beagle 4.1`, using true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

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

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

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

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

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

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

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

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

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

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

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

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

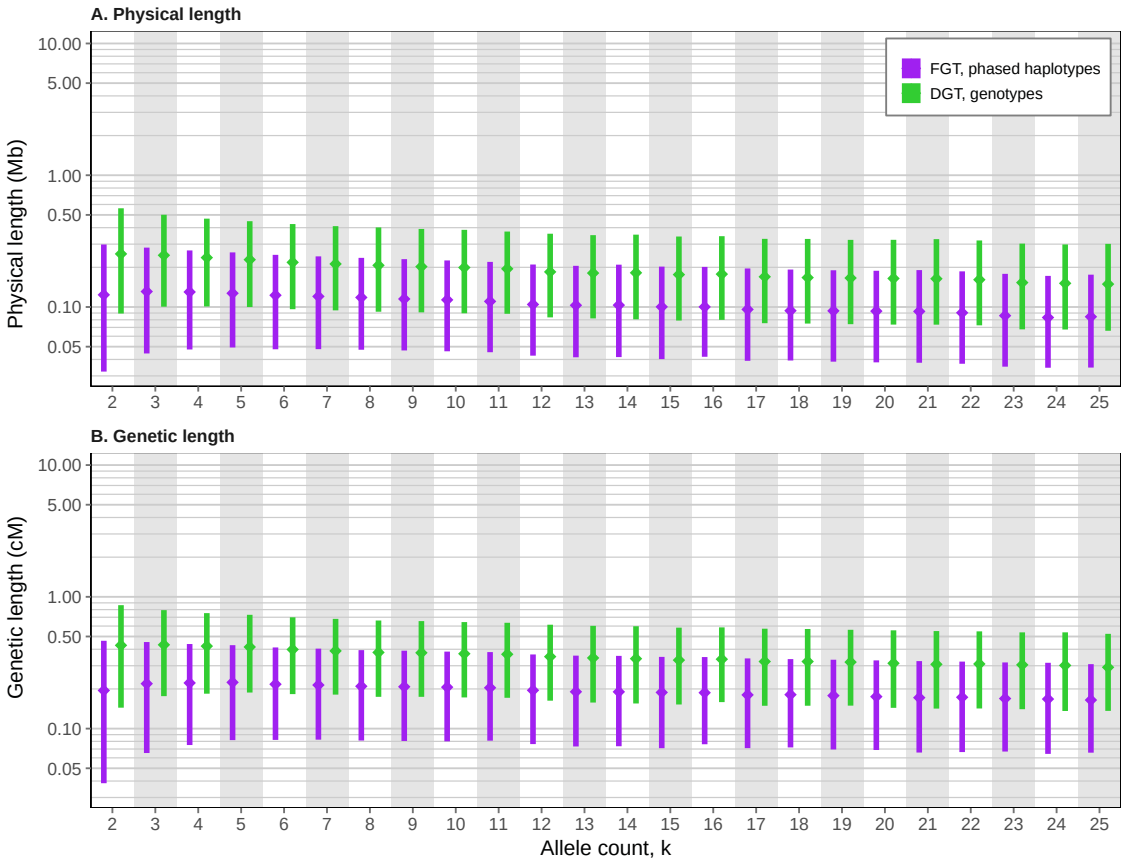


Figure 3.14: Distribution of inferred IBD lengths in 1000 Genomes data, chromosome 20. Results are shown for the detected physical and genetic lengths of shared haplotype segments by f_k, using chromosome 20 in the final release dataset of 1000 Genomes Project Phase III, including $N = 2{,}504$ individuals.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

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

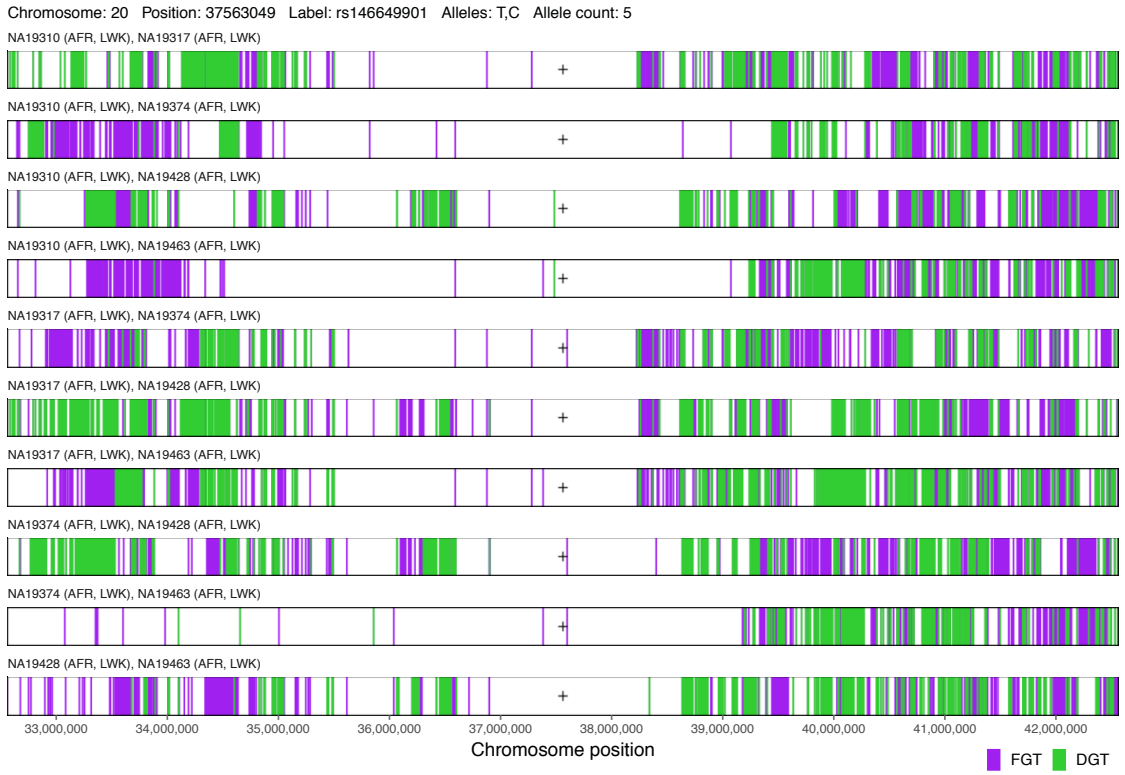


Figure 3.15: Example of breakpoints detected in 1000 Genomes, chromosome 20. One of the target sites analysed was randomly selected and breakpoint detection was performed for all pairs of individuals sharing the focal allele. The plot shows all breakpoints detected using the FGT and DGT relative to the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

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

3.6 Discussion

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

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

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

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

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

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

— Elon Musk

Abbreviations

1000G	1000 Genomes Project
ARG	Ancestral recombination graph
CDF	Cumulative distribution function
DGT	Discordant genotype test
FGT	Four-gamete test
HapMap	International HapMap Project
IBD	Identity by descent
LD	Linkage disequilibrium
LOD	Logarithm of odds
LR	Likelihood ratio
LRP	Long range phasing
Mb	Megabase
MRCA	Most recent common ancestor
RMSLE	Root mean squared logarithmic error
SNP	Single-nucleotide polymorphism
$\mathbf{T_{MRCA}}$	Time to the most recent common ancestor
VCF	Variant Call Format
WGS	Whole-genome sequencing

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

— E. O. Wilson

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Onofrio, R. C., Richter, D. J., Birren, B. W., Daly, M. J., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–861.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., 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.

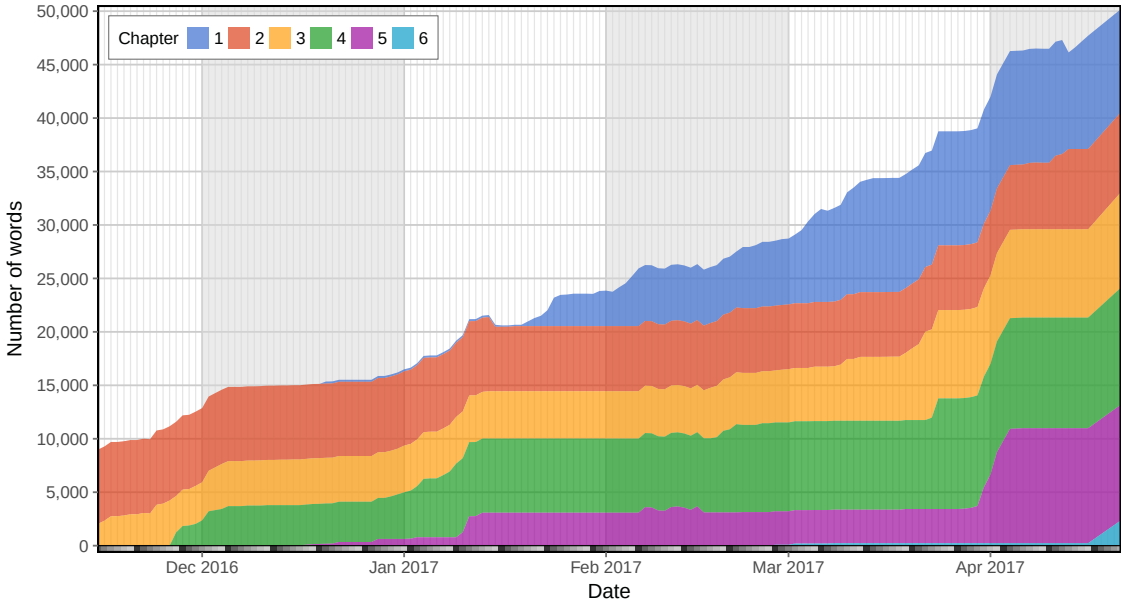
Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, **16**(2), 97–159.

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.

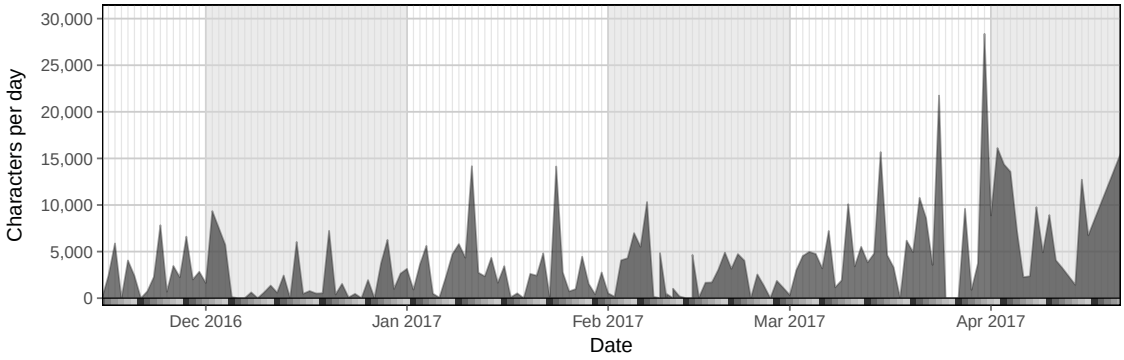
Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, **111**(4), E455–64.

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

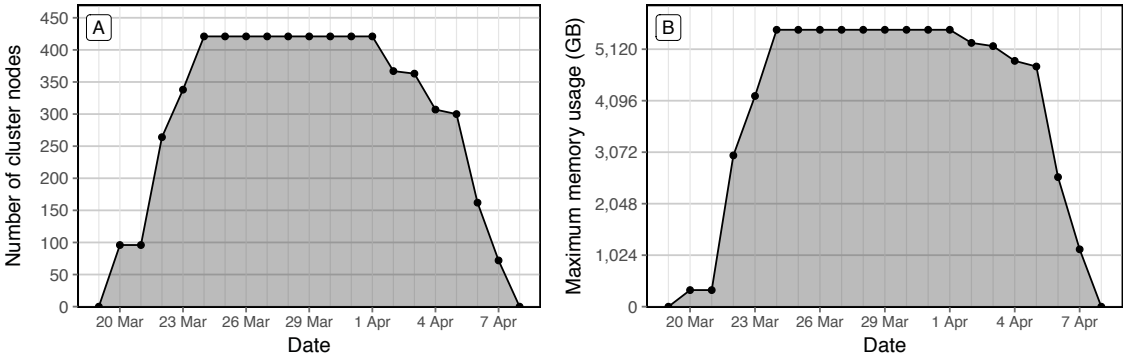
— Adam Savage



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.