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

— Charles Darwin, The Descent of Man

# 1 Introduction

**Contents**



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

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

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

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

## 1.1 Aims and structure of this thesis

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

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

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

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

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

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

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

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

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

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

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

## 1.2 Basic concepts and terminology

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

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

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

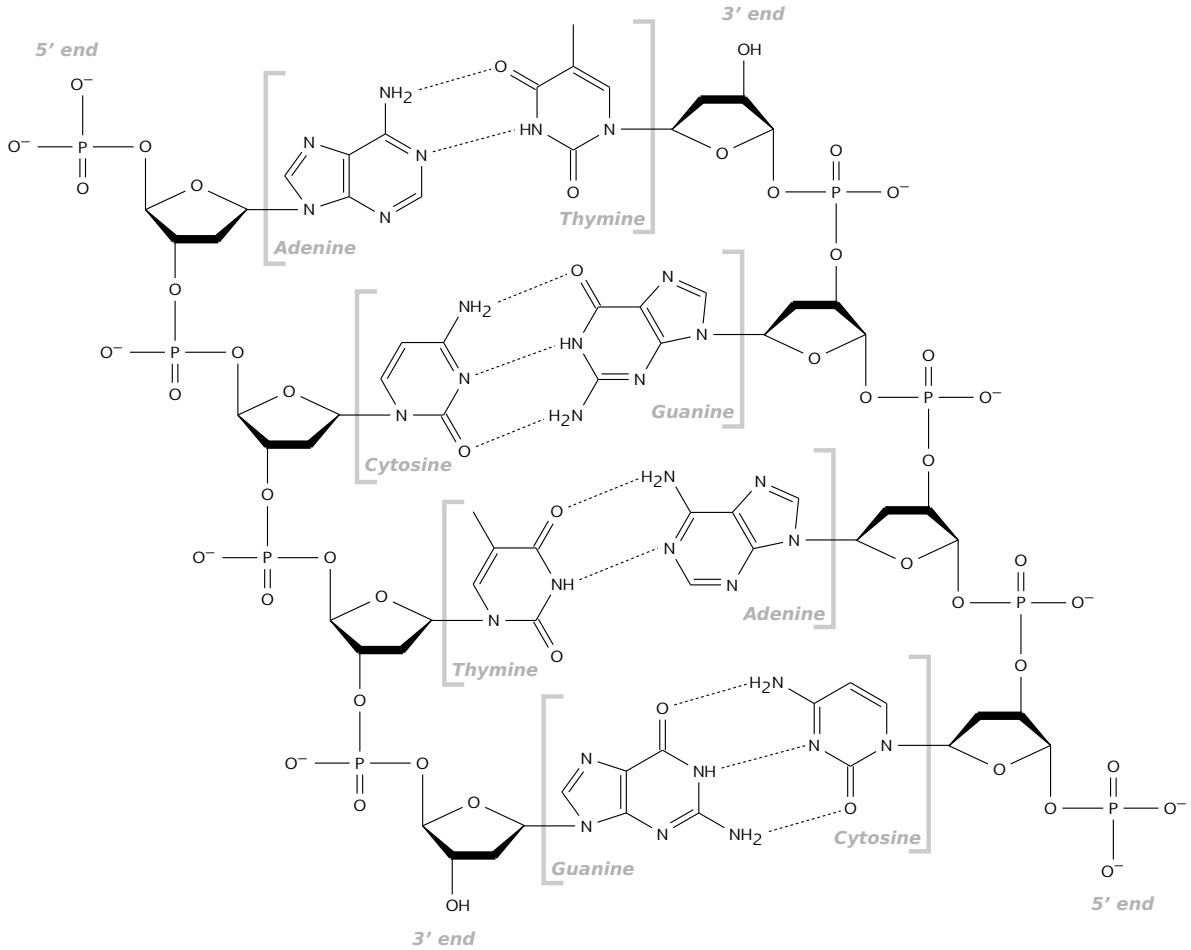



**Figure 1.1: The chemical structure of DNA.** The DNA molecule is a long chain polymer of individual nucleotide building blocks. Each nucleotide is composed of a phosphate residue, a deoxyribose sugar (pentose), and a nucleobase (adenine, guanine, cytosine, or thymine). The phospho-deoxyribose subunits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

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

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

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

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

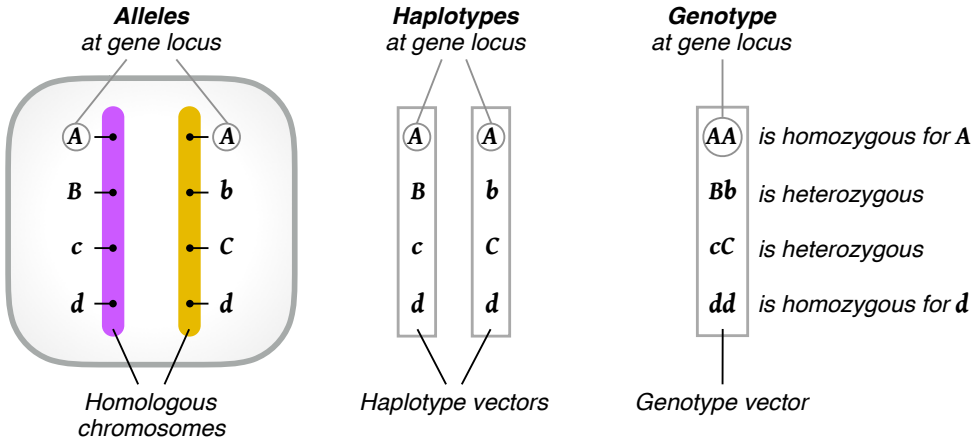



**Figure 1.2: Alleles, haplotypes, and genotypes.** A pair of homologous chromosomes is shown (*left*) on which four gene loci are highlighted; labelled as $A$, $B$, $C$, and $D$. Maternal and paternal chromosomes are shown in *purple* and *yellow* (arbitrarily coloured). Each gene may have two allelic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

### 1.2.1 Mutation

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

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

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

### 1.2.2 Recombination

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

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

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

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

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

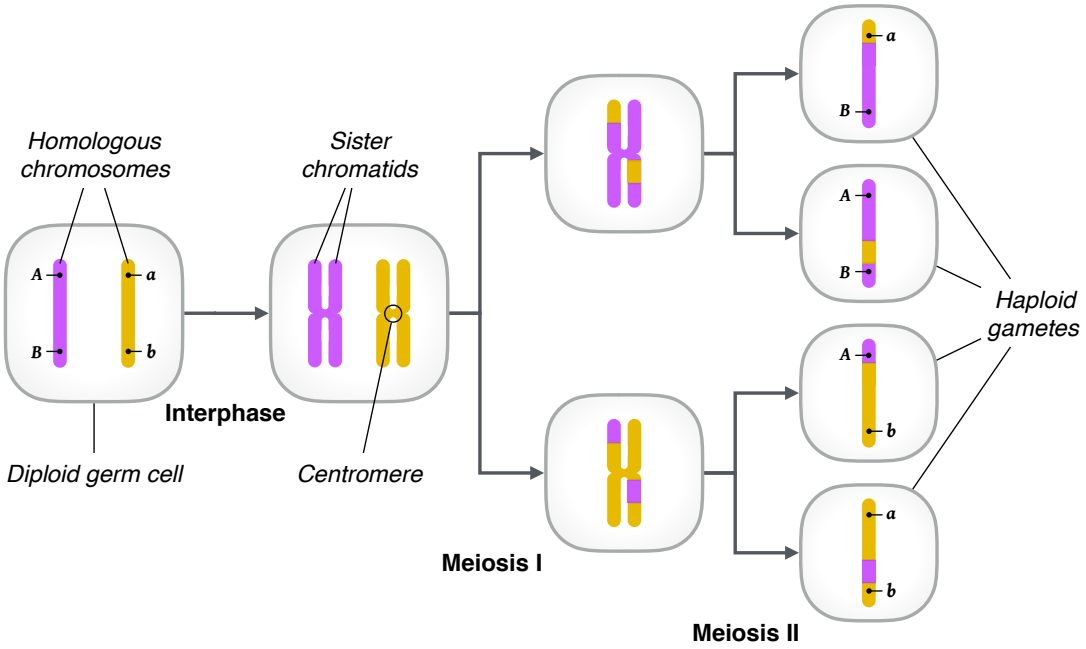



**Figure 1.3: Illustration of recombination during meiosis.** One pair of homologous chromosomes is shown at the beginning of the meiotic cell cycle (*left*). Maternal and paternal chromosomes are shown in *purple* and *yellow* (arbitrarily coloured). The allelic configuration at two sites is indicated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

#### 1.2.2.1 Genetic linkage

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

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

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

## 1.3 Models in population genetics

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

### 1.3.1 Wright-Fisher model

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

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

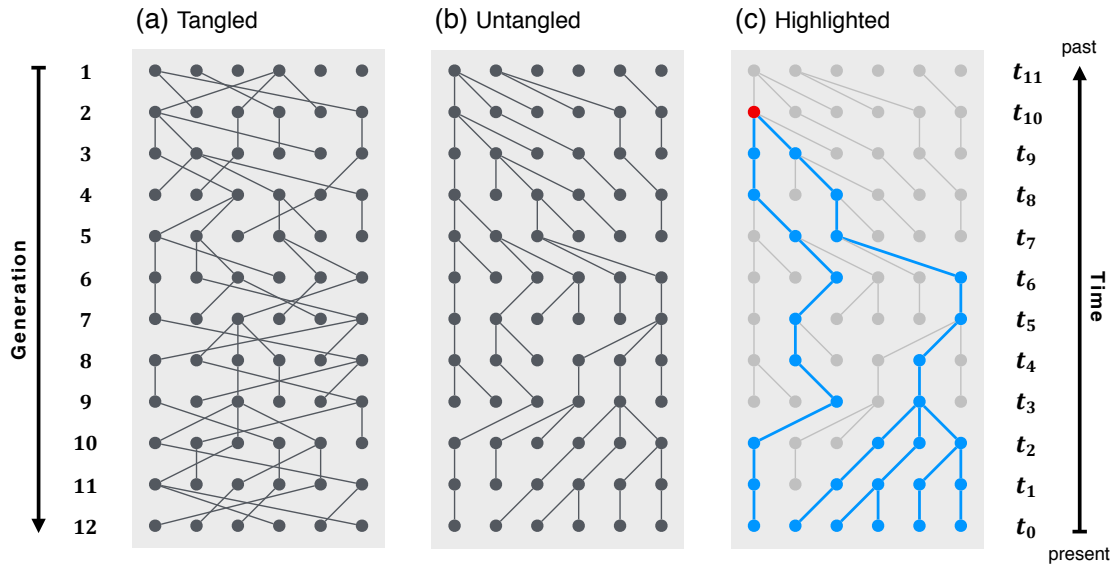



**Figure 1.4: Example genealogy in a Wright-Fisher model.** A population of size $N = 6$ is shown in Panel **(a)**, which is observed over 12 generations. In the neutral Wright–Fisher model, one individual is chosen at random (with replacement) in each generation to produce offspring for the next generation,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

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

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

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

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

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

and the variance is given by

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

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

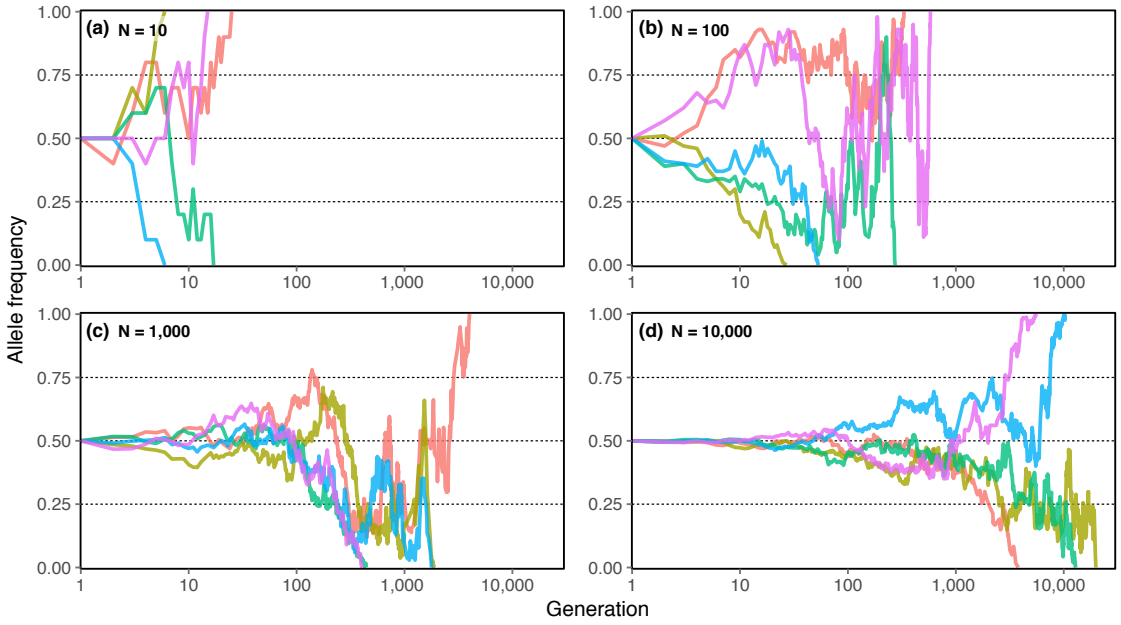



**Figure 1.5: Allele frequency changes over time simulated under the Wright-Fisher model.** A haploid population was simulated under four different constant values of population size, $N$, as indicated in each panel. The change in allele frequency is shown by generation. For each value of $N$, five replicate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

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

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

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

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

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

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

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

### 1.3.2 Coalescent theory

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

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

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

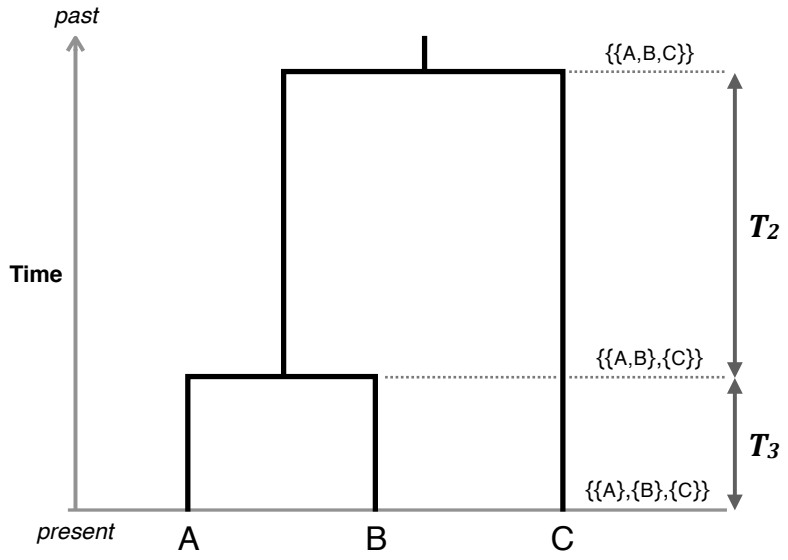



**Figure 1.6: Topology of a genealogical tree in the coalescent.** The genealogical relationship of three haploid individuals is shown, $A$, $B$, and $C$, which represent separate lineages at present, but where $A$ and $B$ are the first to coalesce (back in time). The waiting time between successive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

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

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

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

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

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

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

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

and the variance is

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

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

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

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

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

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

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

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

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

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

ancestor. It follows that

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

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

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

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

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

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

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

and the variance is

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

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

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

the expectation can be expressed as

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

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

#### 1.3.2.1 Effective population size

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

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

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

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

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

#### 1.3.2.2 The coalescent with mutation

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

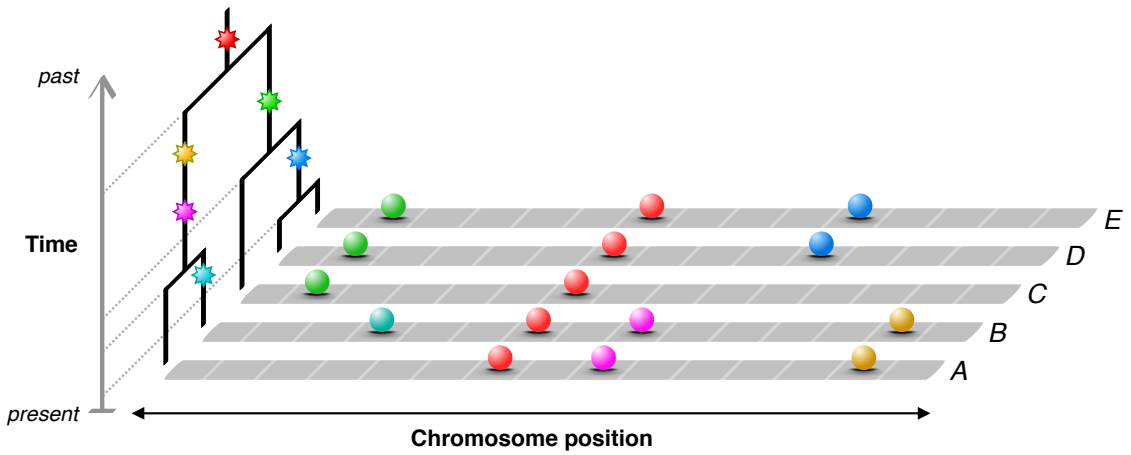



**Figure 1.7: Mutation events on a genealogical tree in the coalescent.** The genealogy of a sample of five haploid individuals ($A - E$) is shown on the left. The time of each coalescent event is indicated by a *dotted* line. Mutation events (*stars*) are placed along the branches of the tree. Each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

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

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

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

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

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

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

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

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

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

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

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

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

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

By rearrangement, the following equation can be obtained;

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

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

#### 1.3.2.3 The coalescent with recombination

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

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

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

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

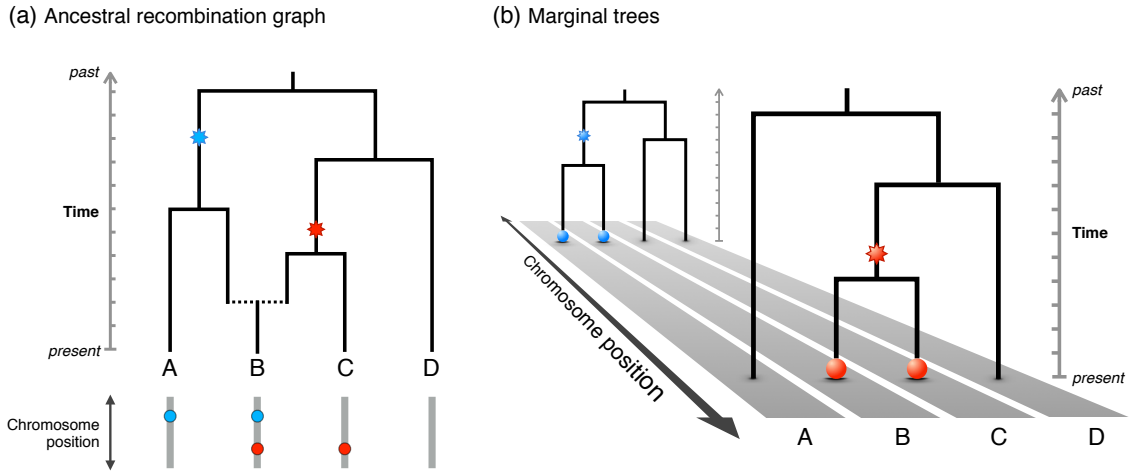



**Figure 1.8: Illustration of the ancestral recombination graph.** Panel **(a)** shows the ARG for a sample of four chromosomes, labelled by $A$, $B$, $C$, and $D$. The *dotted* horizontal line denotes the time of a recombination event between chromosomal lineages. Mutation events are shown as *stars*.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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 1.4 Advances in high-throughput genomic technologies

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

### 1.4.1 Next-generation sequencing

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

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

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

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

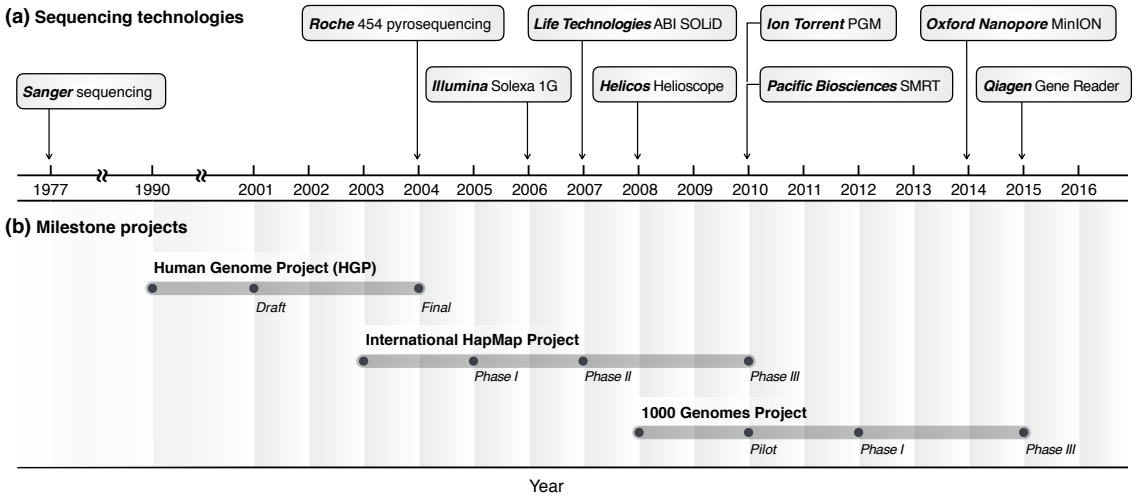



**Figure 1.9: Timeline of sequencing technologies and milestone projects.** Panel **(a)** shows the year of commercial introduction of successfully established next-generation sequencing (NGS) platforms until 2016, following the introduction of the Sanger *et al.* (1977) sequencing method. Panel **(b)**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

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

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

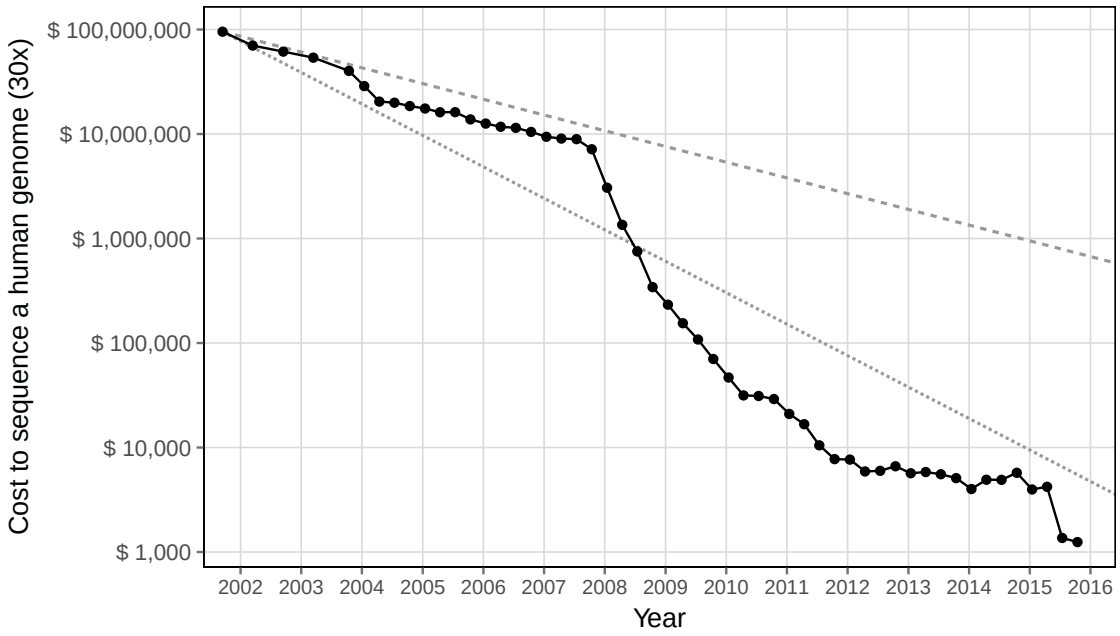



**Figure 1.10: Timeline of cost reduction in DNA sequencing.** Technological improvements in whole-genome sequencing have led to drastic reductions in cost while simultaneously improving accuracy and speed of data generation. The plot shows the development of price per human-sized genome sequenced at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

### 1.4.2 Exploration of the human genome

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

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

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

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

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

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

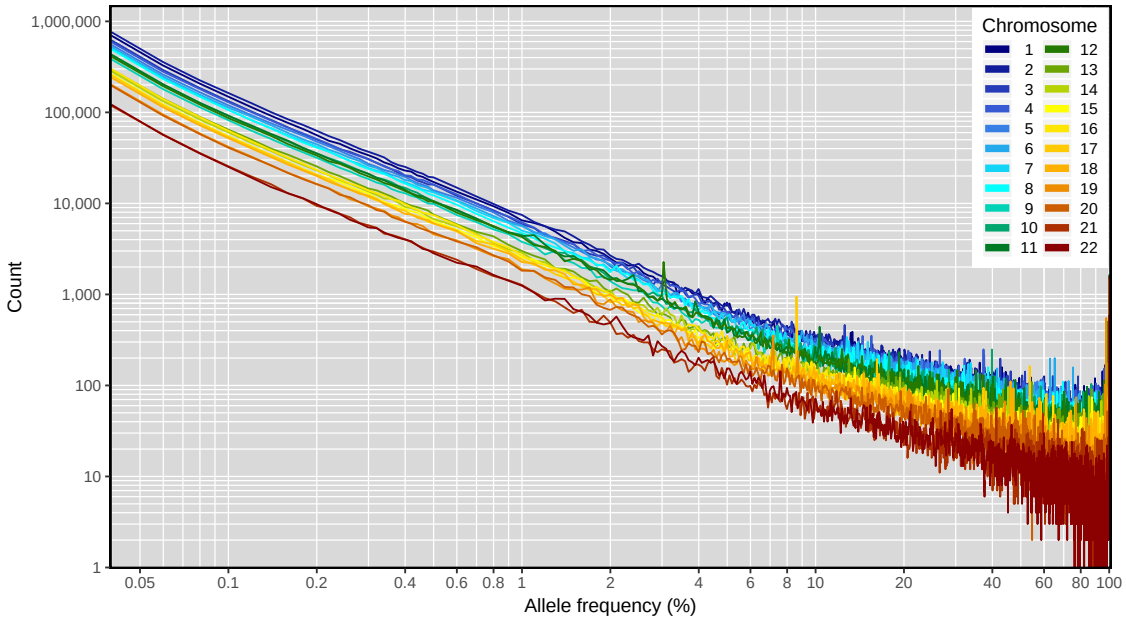



**Figure 1.11: Allele frequency spectrum in the 1000 Genomes Project.** The allele frequency distribution is shown per chromosome (1–22) for all variants contained in the final release dataset of 1000G Phase III. Singletons (private mutations observed only once in the sample) were excluded. Note that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

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

## 1.5 Genome-wide association studies

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

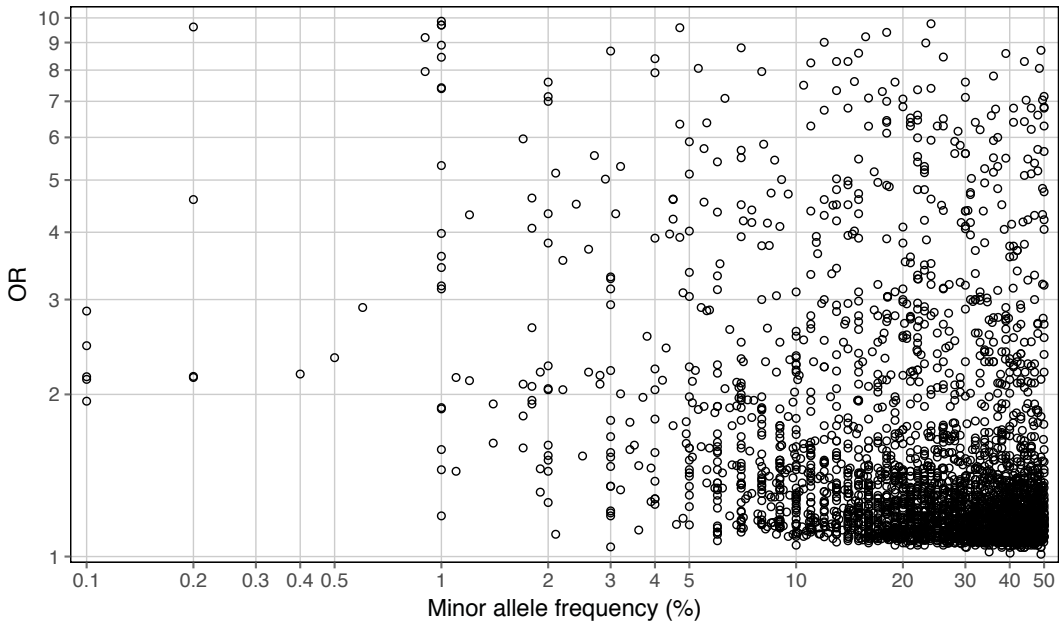



**Figure 1.12: Significant risk-associated variants listed in the NHGRI-EBI Catalogue.** Results are shown for 3,186 unique variants which were reported as being significant at $p$-value $\leq 5 \times 10^{-8}$ and for which odds ratio (OR) values were available in the database. Note that different studies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

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

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

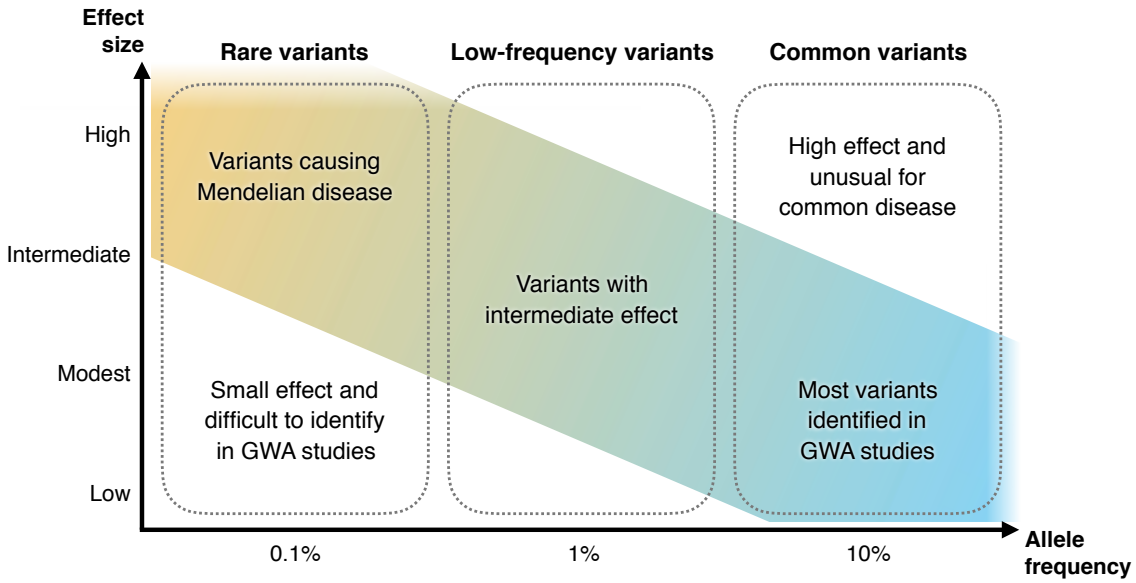



**Figure 1.13: Risk-related variants by allele frequency and effect size.** Rare, low-frequency, and common variants are distinguished by (minor) allele frequency. Note that frequency values are only indicated as approximate guides. Figure adapted from McCarthy *et al.* (2008, Box 7) and Manolio *et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

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

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

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

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

## 1.6 Identity by descent

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

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

### 1.6.1 Single-locus concept

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

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

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

### 1.6.2 Genealogical concept

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

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

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

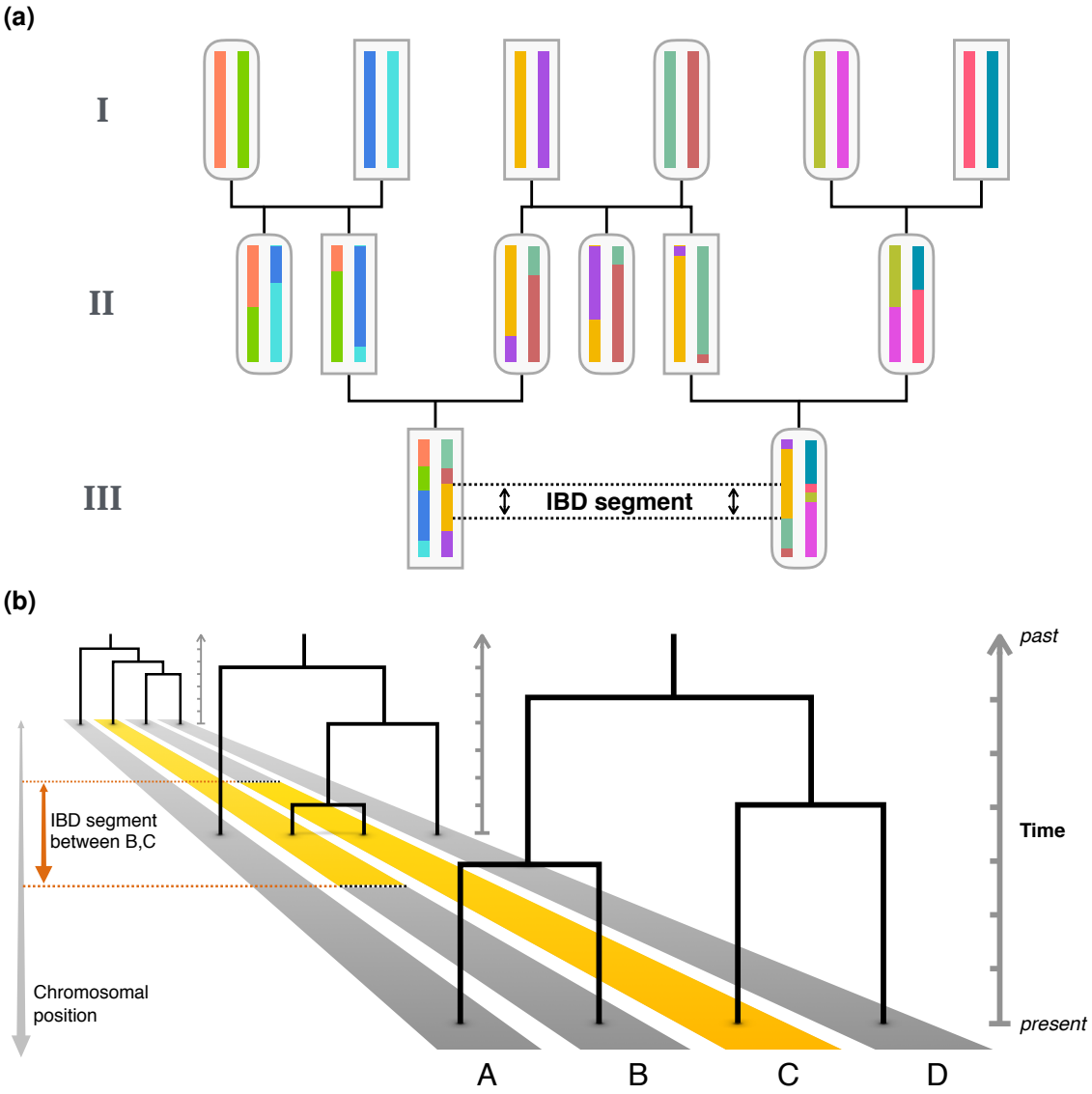



**Figure 1.14: Illustration of haplotype sharing by descent.** Panel **(a)** shows a three-generation pedigree; generation I consists of the founders of the pedigree. The two individuals shown in generation III are first-degree cousins. Male and female individuals are distinguished by square and round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

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

## 1.7 Allele age estimation

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

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

### 1.7.1 Theoretical results

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

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

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

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

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

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

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

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

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

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

### 1.7.2 Application in human disease research

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

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

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

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

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

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

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

*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 |
| **cM** | CentiMorgan |
| **DNA** | Deoxyribonucleic acid |
| **EBI** | European Bioinformatics Institute |
| **GWA** | Genome-wide association |
| **HapMap** | International HapMap Project |
| **HGP** | Human Genome Project |
| **HMM** | Hidden Markov Model |
| **HWE** | Hardy-Weinberg equilibrium |
| **IBD** | Identity by descent |
| **IBS** | Identity by state |
| **LD** | Linkage disequilibrium |
| **MAF** | Minor allele frequency |
| **Mb** | Megabase |
| **MRCA** | Most recent common ancestor |
| **NGS** | Next-generation sequencing |
| **NHGRI** | National Human Genome Research Institute |
| **OR** | Odds ratio |
| **RFLP** | Restriction fragment length polymorphism |
| **RNA** | Ribonucleic acid |
| **SNP** | Single-nucleotide polymorphism |
| $\mathbf{T_{MRCA}}$ | Time to the most recent common ancestor |
| **WGS** | Whole-genome sequencing |

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

— E. O. Wilson

# Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

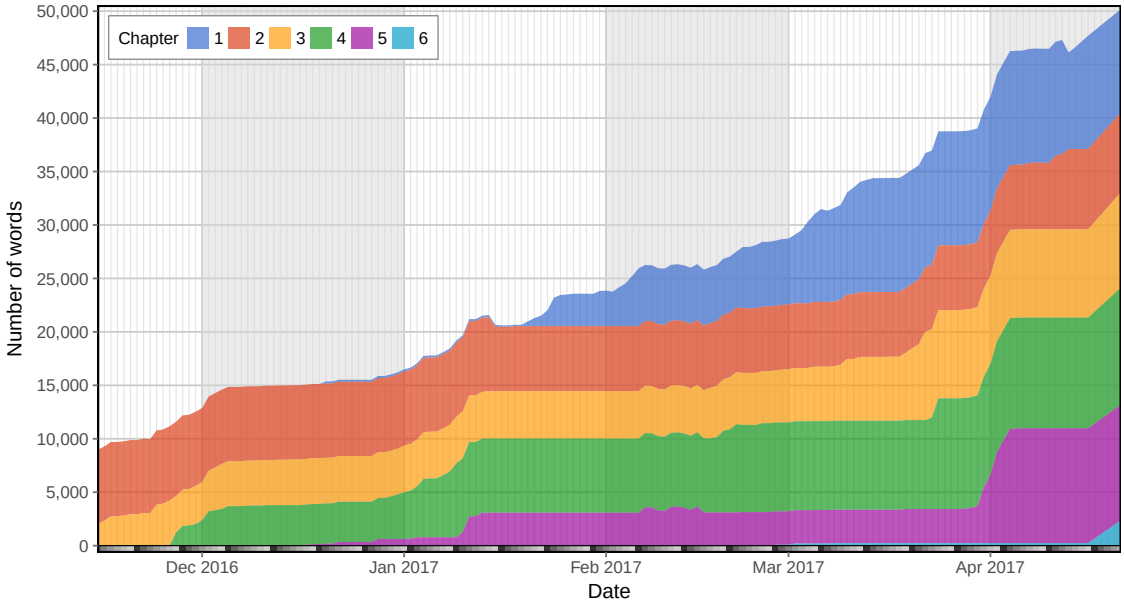
Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, **16**(2), 97–159.

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

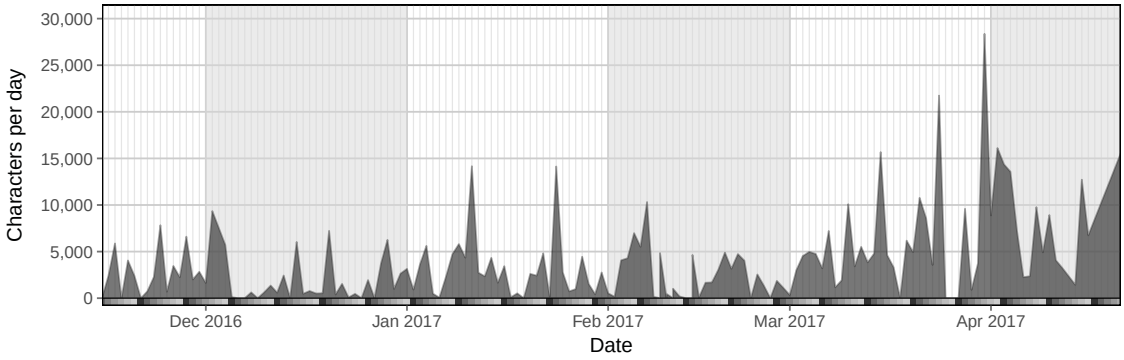
Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, **111**(4), E455–64.

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

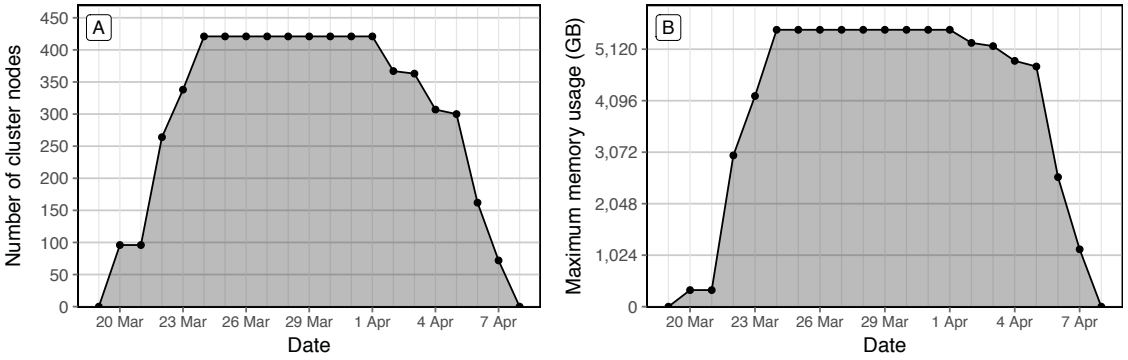
— Adam Savage




**Supplementary Figure 1:** Word count over time during thesis writing period. Shown for the time since I automatically generated daily backups and until the submission of this thesis.




**Supplementary Figure 2:** Number of characters written per day. Note that all characters in each LaTeX file were counted.




**Supplementary Figure 3:** Computer cluster usage one month before the submission date of this thesis. Indicated by the **(A)** number of nodes used and **(B)** daily maximum of computer memory on the cluster of the Wellcome Trust Centre for Human Genetics.