Population Genetics of Identity By Descent

Pier Francesco Palamara

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Population Genetics of Identity By Descent

Pier Francesco Palamara

Recent improvements in high-throughput genotyping and sequencing technologies have afforded the collection of massive, genome-wide datasets of DNA information from hundreds of thousands of individuals. These datasets, in turn, provide unprecedented opportunities to reconstruct the history of human populations and detect genotype-phenotype association. Recently developed computational methods can identify long-range chromosomal segments that are identical across samples, and have been transmitted from common ancestors that lived tens to hundreds of generations in the past. These segments reveal genealogical relationships that are typically unknown to the carrying individuals. In this work, we demonstrate that such identical-by-descent (IBD) segments are informative about a number of relevant population genetics features: they enable the inference of details about past population size fluctuations, migration events, and they carry the genomic signature of natural selection. We derive a mathematical model, based on coalescent theory, that allows for a quantitative description of IBD sharing across purportedly unrelated individuals, and develop inference procedures for the reconstruction of recent demographic events, where classical methodologies are statistically underpowered. We analyze IBD sharing in several contemporary human populations, including representative communities of the Jewish Diaspora, Kenyan Maasai samples, and individuals from several Dutch provinces, in all cases retrieving evidence of fine-scale demographic events from recent history. Finally, we expand the presented model to describe distributions for those sites in IBD shared segments that harbor mutation events, showing how these may be used for the inference of mutation rates in humans and other species.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful to my advisor, Itsik Pe'er, for his endless guidance and support, for having spent five years teaching me so much, and for having given me the opportunity to work in such a stimulating environment. His enthusiasm has been a great source of inspiration. I certainly owe much to all members of the lab, past and present, and the students I worked with. I'm particularly grateful to Sasha, Snehit, Eimear, Vlada, Anat and Yufeng, for their friendship and for countless inspiring discussions. I am very grateful to my thesis committee members for their advice and the time dedicated to my work, and to the many collaborators of these years. I thank the great group of people I met back in my RoboCup days, with whom I made the first scientific steps that influenced me so much.

To my family, Beatrice, Antonio, Gian Marco, and Susanna, for their constant support and their unlimited love.

Chapter 1

Introduction

In a famous paper published in 1965, Gordon Moore, currently co-founder and Chairman Emeritus of Intel Corporation, predicted that the number of transistors on integrated circuits would double approximately every two years, as a result of decreased production costs [Moore and others, 1965]. During the past five decades of technological development, this prediction has been closely matched by empirical data, and Moore's law, as the conjecture is often referred to, is expected to last for a few more years. After the announced completion of the human genome project, in 2001 [Lander *et al.*, 2001; Venter *et al.*, 2001], the development of DNA sequencing technologies has followed a similar trend, with the average cost for obtaining a full human genome DNA sequence dropping exponentially at a rate that closely matched Moore's law. With the transition from Sanger-based sequencing technologies to 'next-generation' sequencing, in 2008, the cost of DNA sequencing had a further, dramatic drop, outpacing Moore's law and bringing the price of a single human genome from 2001's ∼\$3 billion to a few thousand dollars in little more than a decade [NHGRI, 2013].

While the speed at which large volumes of high-resolution DNA sequences are being produced exacerbates issues related to data handling (e.g. hardware storage, processing power), the availability of several fully sequenced individuals from multiple populations worldwide, together with phenotypic information, has enabled data-driven studies of the origins and diversification of human populations, including genomic signatures of evolutionary events [Pool *et al.*, 2010], discovery of genetic markers responsible for the heritability of common traits [Hindorff *et al.*, 2009], and the development new tailor-made diagnostic and therapeutic tools based on an individual's genetic makeup [Hamburg and Collins, 2010]. Achieving these tasks by analyzing such large volumes of data involves relying on statistical and computational methods to develop new specific tools that are simultaneously efficient, making minimal use of computational resources, and effective, successfully extracting and elaborating information for the question at hand. To this extent, a widespread analysis paradigm consists in working on specific "features", or summary statistics obtained from the DNA sequences of the analyzed cohort. These are chosen to succinctly capture the most relevant aspects of the data, while allowing efficient downstream analysis. Choosing the right genomic features is extremely important, as a particular summary statistic may not carry substantial information to address specific questions, while in other cases the relevant genomic features may be hard to access, or require intractable computational efforts.

In this thesis, we focus on developing new models and methodologies for genetic analysis that are based on a specific genomic feature that was recently made available due to technological and computational advances, namely the sharing of long-range haplotypes across purportedly unrelated individuals from a study cohort. These are chromosomal segments that are transmitted from the genome of common ancestors to sets of individuals. Such common ancestors may have lived a large number of generations in the past, so that the co-inheriting individuals may not be aware of their genetic relationship, being therefore purportedly unrelated. Since these segments are copied almost identical from the transmitting common ancestors, they are generally referred to as "identical-by-descent" (IBD) segments, although a small number of mutations and other rare genomic events may occur on the segments during the transmission process. The detection of IBD segments in large datasets of purportedly unrelated individuals (henceforth simply referred to as *unrelateds*) was recently made possible due to (1) advances in the resolution and number of genomic sequences that modern technologies can produce (2) the development of computational methods that are able to phase (i.e. separate an individual's maternal and paternal copies of a diploid chromosome into two distinct sequences) and efficiently locate these IBD segments in a computationally tractable way. A more detailed introduction of the basic concepts underlying IBD segments is provided in Section 1.1.4, and recent review on the subject can be found in [Browning and Browning, 2012; Thompson, 2013].

The reminder of this chapter provides a brief overview of basic definitions and fundamental concepts of population genetics and identity-by-descent. Chapter 2 reports results of the analysis of several densely typed human datasets (HapMap 3, Jewish Hapmap), where descriptive statistics of IBD sharing across unrelateds were shown to capture relevant features of a population's recent evolutionary and demographic history. This preliminary analysis motivated investigating the formal link between IBD sharing and demographic history, which is introduced in Chapter 3, and used to infer population size fluctuations in several synthetic and real populations. In Chapter 4, the framework of Chapter 3 is extended to allow for inferring recent demographic events in demographic models that include several demes, and migration across them. This extension is used to analyze recent demographic events using sequences of 250 families from several Dutch provinces (the Genome of the Netherlands Project). In Chapter 5, the proposed model is further extended to include the occurrence of mutation events within IBD segments. These mutations are informative about the distance to transmitting common ancestors, and can be used in the study of mutation rates and several other applications. We finally provide a brief discussion of the presented work in Chapter 6.

1.1 Population genetics

Long before James Watson and Francis Crick presented the double helical structure of DNA [Watson and Crick, 1953], statisticians of the past century had laid the theoretical foundations of population genetics, which is aimed at providing mathematical support to describing the dynamics of key genetic quantities resulting from the interbreeding of organisms in a sexual population. The pioneering work of Sewall Wright, John B. S. Haldane and Ronald A. Fisher, generally considered the fathers of population genetics, has now been further developed for more than a century, and theoretical predictions of these models have recently been extensively validated by empirical evidence in thousands of genome sequences from diverse populations in different species. In this section, we provide a brief introduction of the basic concepts of population genetics that will be used in the remainder of this thesis, namely the coalescent process and identity-by-descent. Comprehensive introductions to the concepts here briefly illustrated can be found in textbooks such as [Hartl, 1988; Hartl and Clark, 1997; Hein *et al.*, 2004; Wakeley, 2009]. The presented overview is in some cases a summary of the material that can be found in these books.

1.1.1 Basic definitions

Deoxyribonucleic acid, or DNA, is hereditary material coded using an alphabet of four chemical bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Each base couples with its complement (adenine with thymine and cytosine with guanine), forming *base pairs* which are attached to a sugar and a phosphate molecules to form *nucleotides*. A sequence of nucleotides is arranged in a double helix structure which coils around proteins called *histones* to form *chromosomes*, basic physical units found in the nucleus of cells. Humans have 23 such chromosomes, of which 22 are of the same kind in males and females (*autosomes*), while one, the sex chromosome, may differ. Two copies of each chromosome are stored, one inherited from each parent, making humans a *diploid* organism (as opposed to *haploid*, where one copy of each chromosome is stored, or *polyploid*, which may have multiple copies). Diploid individuals produce *gametes* (egg and sperm cells) for sexual reproduction. These contain a single copy of each chromosome formed by mixing the two existing copies during the process of *meiosis*. For the purpose of this thesis, two main events occurring during meiotic division will be discussed: *mutation* and *recombination*.

Mutation occurs when errors are randomly made during the copying of genetic material when the haploid gametes are formed. Mutations involving the change of a single base pair are called *point mutations*. While mutations can occur at other stages of the cell life cycle, those occurred during the production of germ cells, which will be passed down to offsprings, are called *germline* mutations. As these mutations are not present in the parents, they are often referred to as *de novo* mutations. Point mutations are extremely rare, with an estimated genome wide rate of $\sim 1.1 \times 10^{-8}$ [Roach *et al.*, 2010] per nucleotide, per generation (with variation that may depend, among other things, on the father's age at conception [Kong *et al.*, 2012; Sun *et al.*, 2012]). Since a haploid copy of the genome is composed of $\sim 3,200,000,000$ bases (or 3.2 giga base pairs), however, the average diploid genome is expected to harbor around 70 de novo mutations. We note that several other types of rare alterations may occur during meiosis (e.g. insertion, deletion, inversion of genetic material), however these are not relevant for this thesis work, and will not be discussed.

Abstracting from biological mechanisms, a germ cell is created during meiosis by copying consecutive base pairs of a randomly chosen copy of each chromosome (maternal or paternal), until the chromosome end is met or a recombination event occurs. The occurrence of a recombination event between two adjacent nucleotides interrupts the copying process of the currently chosen chromosome (maternal/paternal), and starts the copying of DNA material from the other chromosome of the diploid individual (paternal/maternal) to the haploid gamete, thus potentially creating a patchwork of the original two chromosomes. In a population, recombination results in the shuffling of genetic variation which is created by mutation events. Similarly to mutations, meiotic recombination events are rare, occurring at an average rate[1] of $\sim 1.3 \times 10^{-8}$ between pairs of neighboring nucleotides. The probability of a recombination event occurring is far from uniform across the genome, as specific genomic regions may harbor increased recombination rates (hotspots), while others may have little or no recombination occurring (coldspots). The reconstruction of a mapping between physical genomic location and recombination probability (genetic map) has been extensively studied in both families and using population-level datasets of unrelated individuals [Hudson, 2001; Kong *et al.*, 2002]. The length of genetic maps is measured in Morgans (M), or centimorgans (cM). A centimorgan is defined as 1% chance of observing a recombination event during a meiosis (one generation).

In the remainder of this work, a specific genomic location may be referred to as a *site* or, equivalently, a *locus* (plural: *loci*), or a *gene* (the latter typically indicating a region whose DNA content encodes a protein). Due to the occurrence of mutations, different

[1] average rate computed from autosomal genetic map of the 1,000 genomes project available at `http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated_SHAPEIT2.html`

versions of a locus or of a gene may exist in a population. These are referred to as *alleles*. When several loci are simultaneously considered and they all belong to a single chromosome (e.g. maternal/paternal), these constitute a *haplotype*. Haplotypes need not be adjacent sites, and may consist of a sparse subset of loci from a genomic region. Datasets will be distinguished in *SNP array* data and *whole-genome sequencing* data. SNP array data results from genotyping technologies that do not read the entire genome of the analyzed individuals, but rather only focus on subsets of genomic sites that are known to harbor single-point mutations that reached high frequency in certain human populations, and are informative for medical genetics purposes or to discriminate genomic variation across individuals. These mutations are called *SNPs*, short for *single nucleotide polymorphisms*. In this work, we will ignore those rare polymorphisms for which more than two alleles are present in the population. We will only deal with polymorphisms where two alleles are present: the *wild type*, or the *reference allele*, and the mutated allele. Whole-genome sequencing data results from the more recent high-throughput sequencing technologies, and typically results in the complete reading of a human genome. It is to be noted that both genotyping and sequencing technologies typically do not provide information on the maternal/paternal haplotypes of the analyzed individuals. Rather, for a biallelic locus they provide a *genotype*, i.e. the count of nucleotide copies that differ from the human genome reference sequence at a specific location. For a diploid individual, these counts take values $0, 1$ or 2. The process of reconstructing haplotypes from genotype information is called *phasing*, or *haplotyping*. While phasing approaches are not directly discussed in this work, the ability to correctly phase genotypes into haplotypes is fairly important for the material presented in this thesis, and a review of methods for computational phasing can be found in [Browning and Browning, 2011b].

1.1.2 Population models

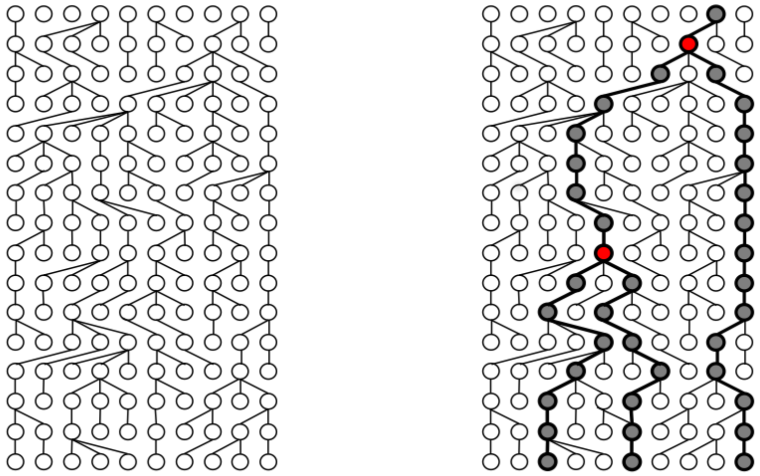
The distribution of genetic variability found in modern day populations is strongly influenced by demographic history. Events such as migrations and population size fluctuations determine the rate at which new mutations spread, and the frequencies of these mutations may differ substantially across different cohorts. Several idealized populations models have been developed in order to study quantities such as the frequency and distribution of genetic variation. In all cases, the goal is to simplify the relevant biological processes to achieve mathematical tractability while maintaining the highest level of realism.

The Wright-Fisher model [Fisher, 1930; Wright, 1931] is arguably the most important and widely used population model. A number of assumptions are made in a Wright-Fisher population:

1. Generations do not overlap. All individuals in the population die at the same time, and a new generation is created.
2. The population size remains constant in time. At each generation the number of individuals is the same as in the previous generation.
3. All individuals in a population have a single chromosome, and do not need another individual to reproduce to the next generation (asexual, haploid).
4. There is no recombination (or only one site is considered). When reproduction occurs, the entire genetic material is copied to an individual of the next generation.
5. Equality of fitness and lack of population structure. At each generation an individual may reproduce to the next generation with the same probability of all other individuals.

To create a new generation for a population of size N, an individual is sampled from the previous generation, with replacement. The sampling is repeated N times, until the new generation is fully defined, and the previous generation dies. An example of this process is shown in Figure 1.1. This model allows calculating several quantities of interest. First, it is possible to compute a distribution for the number of offspring that an individual has in the following generation. Since all individuals have the same chance $1/N$ of being chosen at each draw of a new individual for the next generation, the distribution for the number n of offspring at the next generation will be binomial, with mean $1/N$:

$$P(n = k|N) = \binom{N}{k} \left(\frac{1}{N}\right)^{k} \left(1 - \frac{1}{N}\right)^{N-k} \tag{1.1}$$



(a) A genealogy in the Wright-Fisher model.

(b) Highlighting the ancestry of three samples.

Figure 1.1: Details from a sample genealogy in a Wright-Fisher model (figures adapted from [Hein *et al.*, 2004]).

It follows that the expectation and variance for the number of offspring of an individual are

$$\mathrm{E}[n|N] = N\left(\frac{1}{N}\right) = 1 \tag{1.2}$$

$$\mathrm{Var}[n|N] = N\left(\frac{1}{N}\right)\left(1 - \frac{1}{N}\right) = 1 - \frac{1}{N} \tag{1.3}$$

A multinomial distribution can be used to model the joint distribution for the number of children of two individuals from a population, and, using standard properties of multinomials, we can obtain their covariance as

$$\mathrm{Cov}[n_1 n_2|N] = -N\left(\frac{1}{N}\right)^2 = -\frac{1}{N} \tag{1.4}$$

As expected the covariance decreases as N is increased, since an individual that has a large number of children does not strongly affect the number of children another individual may have if the population is not constrained to be small. If an allele is carried by i individuals in the population, the chance of finding k copies at the next generation can be computed using analogous reasoning, but the probability of a single draw in the binomial distribution will now be $p = \frac{i}{N}$. The expectation and variance will therefore be

$$\mathrm{E}[n|N,i] = N\left(\frac{i}{N}\right) = i \tag{1.5}$$

$$\mathrm{Var}[n|N,i] = N\left(\frac{i}{N}\right)\left(1-\frac{i}{N}\right) = i\left(1-\frac{i}{N}\right) \tag{1.6}$$

One important quantity that may be calculated in this model is the probability that out of two sampled individuals one carries an allele and the other does not, given that the population is of size N and that the allele has frequency i. Under the assumption that the two chromosome copies of a diploid individual are randomly sampled from a population of haploid individuals, this is the probability of finding a heterozygous site along the genome of an individual (*heterozygosity*). Assuming an allele can be of the kinds A or a, and that there are i copies of the allele A at generation 0, the initial frequency of A is $p_0 = i/N$, and the chance of sampling (with replacement) two different copies out of N individuals is

$$H_0 = P(A,a|N,i) + P(a,A|N,i) = 2p_0(1-p_0) \tag{1.7}$$

Using the random variable P_1 to represent the frequency of the allele at generation 1, the expected heterozygosity at the next generation can be computed using equations 1.5 and 1.6

$$\begin{aligned}
\mathrm{E}[H_1|N,i] &= \mathrm{E}[2P_1(1-P_1)] \\
&= 2(\mathrm{E}[P_1] - \mathrm{E}[P_1^2]) \\
&= 2(\mathrm{E}[P_1] - \mathrm{E}[P_1]^2 - \mathrm{Var}[P_1]) \\
&= 2p_0(1-p_0)\left(1-\frac{1}{N}\right) \\
&= H_0\left(1-\frac{1}{N}\right)
\end{aligned} \tag{1.8}$$

Indicating that heterozygosity is expected to decrease, and it is expected to do so faster in small populations. Note that, applying the result of equation 1.8 recursively for g generations, heterozygosity is expected to have an exponential decay

$$\mathrm{E}[H_g|N, i] = H_0 \left(1 - \frac{1}{N}\right)^g \approx H_0 \ e^{-g/N} \tag{1.9}$$

We note that while the Wright-Fisher model is the most widely adopted, other models have been proposed and are in some cases more convenient in terms of realism or mathematical tractability. One notable example is the Moran model [Moran, 1958; Moran, 1962], which will be however omitted as not relevant for this work.

1.1.3 The coalescent

In a series of papers published in 1982, Kingman has shown that a stochastic process named *the coalescent* is able to describe the genealogical dynamics emerging from several idealized population models, including the Wright-Fisher model [Kingman, 1982b; Kingman, 1982c; Kingman, 1982a]. In the coalescent, the ancestral lineages of a set of considered individuals from a population are traced backwards in time, allowing for a quantitative description of key genealogical events that only requires keeping track of such subset of lineages.

1.1.3.1 The basic coalescent

If we trace the ancestral lineages of two individuals from a Wright-Fisher population back in time, repeatedly sampling a random ancestor from the previous generation, a common ancestor will be found when both individuals happen to sample the same parent (i.e. these lineages *coalesce*, as in the example of Figure 1.1b). The chance a parent is chosen by one of the individuals is N^{-1}, and since both individuals choose independently, the chance both individuals choose the same parent is N^{-2}. Since there are N parents to choose from, the chance a common ancestor will be found at a given generation is $N \times N^{-2} = N^{-1}$. The waiting time (in generations) to the most recent common ancestor ($TMRCA$) can therefore be expressed using a geometric distribution with parameter N^{-1}

$$P(g = k|N) = \left(1 - \frac{1}{N}\right)^{k-1} \frac{1}{N} \tag{1.10}$$

If we are tracing n individuals from the current generation, a total of $\binom{n}{2}$ pairs of ancestral lineages are followed, and we are interested in the time to the first coalescence of such lineages. The chance that no coalescence occurs during one generation is now

$$\frac{N-1}{N}\frac{N-2}{N}\ldots\frac{N-n+1}{N} = \prod_{i=1}^{n-1}\left(1 - \frac{i}{N}\right) = 1 - \sum_{j=1}^{n-1}\frac{j}{N} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{1.11}$$

If we ignore the term in $\mathcal{O}\left(\frac{1}{N^2}\right)$, which is negligible for large population sizes, the probability of a coalescent event in the previous generation is then $\binom{n}{2}\frac{1}{N}$ and again, using a geometric distribution

$$P(g = k|N) \approx \left[1 - \binom{n}{2}\frac{1}{N}\right]^{k-1}\binom{n}{2}\frac{1}{N} \tag{1.12}$$

Note that we can switch to a continuous time approximation, using the exponential distribution in lieu of the geometric distribution.

$$P(T = t|N) \approx \binom{n}{2}\frac{1}{N}e^{-\binom{n}{2}\frac{1}{N}} \tag{1.13}$$

Simulating the genealogy for a sample of n individuals in a population of size N is easy using this formulation, and it involves repeatedly sampling coalescent times from exponential distributions with parameters $\binom{n_t}{2}\frac{1}{N}$, for $n_t = n, n-1, \ldots, 2$, reflecting the decreasing number of ancestral lineages as pairs of individuals find common ancestors.

Again, a number of relevant genealogical quantities can be expressed in this model. Since we are using an exponential distribution, the expected time to the first coalescence event for these samples is $\mathrm{E}[T_n] = \left[\binom{n}{2}\frac{1}{N}\right]^{-1} = \frac{2N}{n(n-1)}$, and the variance is $\mathrm{Var}[T_n] = \left[\binom{n}{2}\frac{1}{N}\right]^{-2} = \frac{4N^2}{n^2(n-1)^2}$. We can now compute the expected TMRCA for all these samples by summing the expected times for the occurrence of $n-1$ coalescence events, which occur with linearly decreasing rate as pairs of lineages coalesce

$$\mathrm{E}[T] = \sum_{i=2}^{n} \mathrm{E}[T_i] = 2N \sum_{i=2}^{n} \frac{1}{i(i-1)} = 2N\left(1 - \frac{1}{n}\right) \tag{1.14}$$

And the variance can be similarly computed by summing the independent variances of each coalescence event. Using similar principles, we may also compute the expected total branch length for a tree representing the genealogy of these samples as

$$\mathrm{E}[L] = \sum_{i=2}^{n} i\,\mathrm{E}[T_i] = 2N \sum_{i=1}^{n-1} \frac{1}{i} \approx 2N \log n \tag{1.15}$$

1.1.3.2 The coalescent with mutation

The coalescent process is suitable to include mutation events, and therefore study the distribution of genetic variation in idealized populations. The *infinite sites assumption*, due to Motoo Kimura [Kimura, 1969], allows to simplify calculations in this context. Under the infinite sites assumption, whenever a mutation occurs, it always results in a new mutated site (i.e. it is impossible that a site that is already mutated in the population mutates again). Since the chance of two mutations affecting the same site is inversely proportional to the number of available sites, assuming an extremely large genome results in an infinitesimal probability for this event. This assumption is justified by the observation that the number of mutated sites in human populations is relatively small compared to the number of available sites in the genome (i.e. human DNA sequences are largely identical).

Consider n individuals who have a genome of s sites, and a genealogical tree of total length L representing the coalescent history of these individuals along their entire sequence (as later described, the occurrence of recombination events may result in different trees for different genomic regions, but no recombination is assumed for now). A mutation may occur independently, with a small probability μ at any meiotic copy of each nucleotide. In this scenario, the total number of mutation events can be modeled as a Poisson distributed random variable, with mean $\mu s L$. Furthermore, due to the infinite sites assumption, each mutation event occurring along the genealogy is harbored by a distinct site. The number of total mutation events will therefore be equivalent to the number of mutated sites. Using Equation 1.15 to express the expected volume of the genealogical tree and defining $\theta = 2N\mu$, the following estimator can be obtained:

$$\hat{\theta}_s = \frac{m}{\sum_{i=1}^{n-1} \frac{1}{i}} \tag{1.16}$$

Where m is the observed number of mutated sites in the analyzed sample. The parameter θ is referred to as the scaled mutation rate, as it includes the value of the population size N. Such an estimator, often referred to as Watterson's estimator, allows inferring the size of the population based on the observed number of mutated sites in a group of sequences, assuming a Wright-Fisher population model, and for a given value of μ. Because a Wright-Fisher model is only approximating the real genealogical process that results in the observed distribution of mutation events, the recovered population size has to be viewed as a projection of the true genealogical process onto the idealized Wright-Fisher population. A population size inferred using similar estimators is generally referred to as *effective population size* ([Wright, 1931]). Using classical estimators such as Watterson's, the effective population size of all humans has been inferred to be $N_e \approx 20,000$ haploid individuals [Takahata, 1993]. However note that several possible definitions of effective population size exist [Ewens, 2004], depending, among other things, on which summary statistics are used to match real and idealized populations (e.g. the number of segregating sites in the case of Watterson's estimator). Inferring the effective population size will be a central task in the remainder of this work, and new estimators of N_e will be derived in Chapter 3. The assumption of constant population size used in the Wright-Fisher population model will often be relaxed (thereby describing effective population sizes as a function of the considered genealogical time), and new summary statistics obtained from the genetic data will be employed to achieve higher resolution into the recent past of a studied cohort.

1.1.3.3 The coalescent with mutation and recombination

To conclude this overview of the coalescent process, we include the modeling of recombination events along the sequences during the genealogical process, introduced in [Hudson, 1983]. As in the case of mutations, recombination events may occur between any pair of sites at any transmission of the genetic material (provided the recombination rate between these sites is positive). While mutation does not affect the tree structure of the genealogy, however, recombination does.

Again, consider n sequences of length s sites. Assume recombination occurs at the same rate for all pairs of sites, and that a sequence has a total chance of recombining of ρ per generation. Under these conditions, if a recombination event occurs, the exact location can be randomly sampled along the sequence. It is possible that, for long chromosomal regions that have high recombination rates, more than one recombination occurs during one generation. Again, however, we measure time in the continuous space, so that only one recombination event is allowed to occur at a time, but the number of recombination events occurring during a unit interval of time may be greater than one. The effect of a recombination event occurring between the sites s_i and s_{i+1} is to break one of the ancestral lineages that we are tracing backwards in time. This creates two lineages, one harboring the ancestral material in the range $[1, s_i]$, and the other carrying the ancestral material in $[s_{i+1}, s]$. After a recombination event occurs, the number of ancestral lineages being traced increases by one. This turns the genealogical structure representing the cohort's genetic history from a tree into a graph, as shown in Figure 1.2. This graph structure, which may assume very complex forms for large sample sizes and long genomic regions, is called the *ancestral recombination graph* (ARG), introduced in [Griffiths and Marjoram, 1997].

While some quantities may still be derived analytically, the ARG is a fairly complex mathematical object, and it often requires the use of numerical sampling for its use in quantitative analyses. A possible sampling algorithm for the ancestral recombination graph operates as follows:

1. Initialize the number of ancestral sequences to $k = n$, the samples from the current

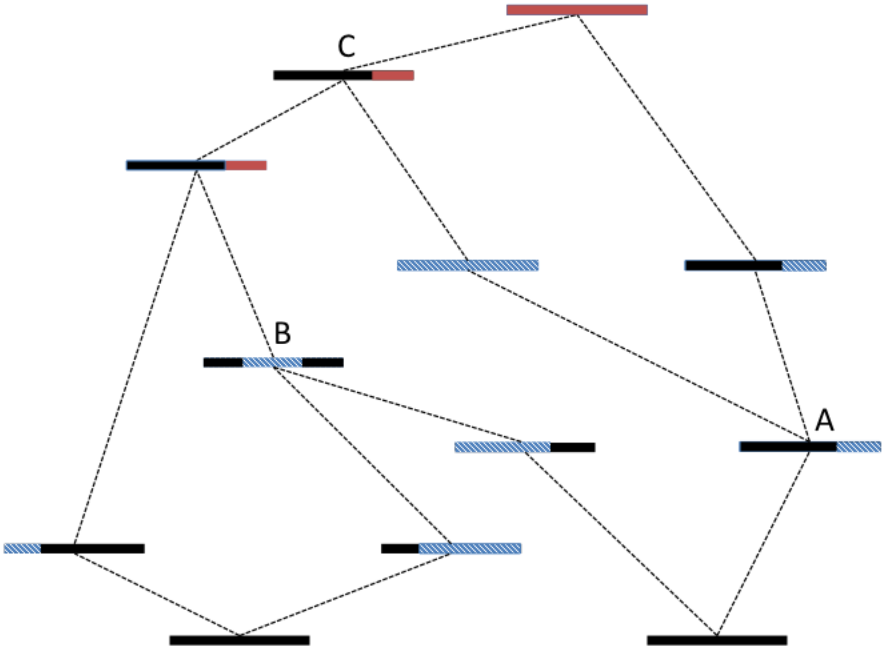


Figure 1.2: A sample genealogical history including recombination events (figure adapted from [Hein *et al.*, 2004]).

generation.

2. Recombination occurs with rate $k\rho$, while coalescence occurs with rate $\frac{\binom{k}{2}}{N}$, and the time distribution to the first event is exponential in both cases. To sample the time to the first occurrence of an event (either recombination or coalescence), sample from an exponential distribution with rate $k\rho + \frac{\binom{k}{2}}{N}$.

3. The event is a recombination with probability $Nk\rho/\binom{k}{2}$, a coalescent otherwise. Draw a uniform value between 0 and 1 to select the type of event.

4. Handle the sampled event: if it is a coalescent, randomly choose two lineages and merge them; update $k = k - 1$. If it is a recombination, sample a random lineage and break it at a uniformly chosen point along the genome; update $k = k + 1$.

5. If $k > 1$, go to step 2.

Figure 1.2 shows an example of running such an algorithm. Note that although the number of traced lineages may grow through recombination events, the algorithm is expected to converge, since individuals are eliminated through coalescent events at a rate that is quadratic in k, and created through recombination at a rate that is linear in k. As in the case of no recombination, sequences can be generated after having sampled an ancestral recombination graph, by introducing mutations over the graph edges using the same procedure that was discussed in Section 1.1.3.2.

1.1.3.4 Approximations of the coalescent

The algorithm shown in the previous section for sampling from the coalescent with recombination process may be improved in several ways. A first possible improvement follows from the observation that some lineages that are created and traced during the sampling process are not affecting the final sequences. Consider for example the ARG of Figure 1.2. The event marked with the letter “A” is a recombination that creates a lineage that does not contain any genetic material inherited by present-day individuals. This lineage will increase the coalescent rate, and will eventually be absorbed during the coalescent event marked with the letter “C”. The creation and the absorption of this lineage have no effect on the genealogy, and may be omitted. Since the exponential distributions that are used to model the timing of these events are memoryless, it turns out that omitting these events does not affect the distribution of the sampled ARG structures. To avoid tracing these lineages, therefore, it is sufficient to modify the algorithm so that if a recombination would produce a lineage that carries no ancestral material, no action is taken.

A number of additional improvements can be developed for this basic algorithm, an extensive discussion is beyond the scope of this work. It is however worth mentioning an approximation of the ARG generation algorithm that resulted in substantial further development. The algorithm described in the previous section operates backwards in time (“vertical algorithm”), starting from a set of individuals in the present generation and sampling ancestors or splitting recombinant lineages until a single common ancestor is found. Alternatively, it is possible to sample from the same space of ancestral recombination graphs by moving along the chromosome (“horizontal algorithm”), rather than backwards in time. Such horizontal algorithm, which was developed in [Wiuf and Hein, 1999] and is here omitted for brevity, has a computational complexity that is comparable to that of the horizontal version (depending on which improvements to the basic version are considered). It is however appealing because several methods in computational genetics analyze DNA sequences moving from left to right (or right to left), assuming an underlying Markovian process and relying on computational machinery such as Hidden Markov Models to perform inference of relevant features. The version introduced in [Wiuf and Hein, 1999], however, violates Markovian properties, as ARGs are intrinsically not Markovian when analyzed horizontally. This is due to the presence of nodes such as the one marked with letter “B” in the example of Figure 1.2, where a lineage with a “gap” is created from the coalescence of two lineages whose ancestral material does not overlap. The existence of this kind of coalescent events requires keeping track of the entire history of genealogical events in an algorithm that moves horizontally across the genome, therefore violating a key Markovian property that requires the distribution of future states to be only dependent on recent states. In a seminal paper by Gil McVean [McVean and Cardin, 2005], it was noted that the effects caused on commonly used summary statistics by the coalescence of lineages that with "gaps" in their ancestral material are negligible. The sequentially Markovian coalescent (SMC), introduced in [McVean and Cardin, 2005], provides an approximation of Wiuf and Hein's horizontal algorithm that substantially simplifies the computation of ARGs. This approach has been recently used in a variety of genomic applications, some of which found application in the reconstruction of demographic events, and will be briefly discussed in Chapter 6. Many of the methods described in this thesis are related to the SMC model, depending on the definition of IBD (see Section 1.1.4.1).

We conclude by noting that approximations of the vertical algorithm have also been developed. In [Parida *et al.*, 2011], for instance, a similar approximation is made to limit coalescent events to those lineages that have an overlapping region of ancestral material, preventing the formation of gaps as the one seen in the example of Figure 1.2.

1.1.4 Identity by descent

In this section we will introduce the basic concepts related to the co-inheritance of identical-by-descent (IBD) haplotypes that are relevant to the development of this work.

Consider the structure represented in Figure 1.3. In this sample pedigree a pair of fourth degree cousins share two common ancestors that lived five generations in the past. These diploid ancestors each have two copies of their autosomal chromosomes, represented using colored bars. At each generation, the offspring inherit a chromosome copy from each of their two parents. Such inherited copies result from the meiotic events that generate germ cells, during which recombination may break down and mix the original chromosome copies present in the diploid parents. In the depicted pedigree, individuals from the population mate with individuals that are direct descendants of the pair of common ancestors living five generations in the past. It is assumed that the genetic material of these external individuals (founders) is unrelated to that of the pair of ancestors. After five generations, the pair of extant fourth degree cousins happen to both inherit stretches of the colored chromosomes from their common ancestors. The blue stretch of chromosome, in particular, overlaps in a region, which constitutes an identical-by-descent segment, or haplotype.

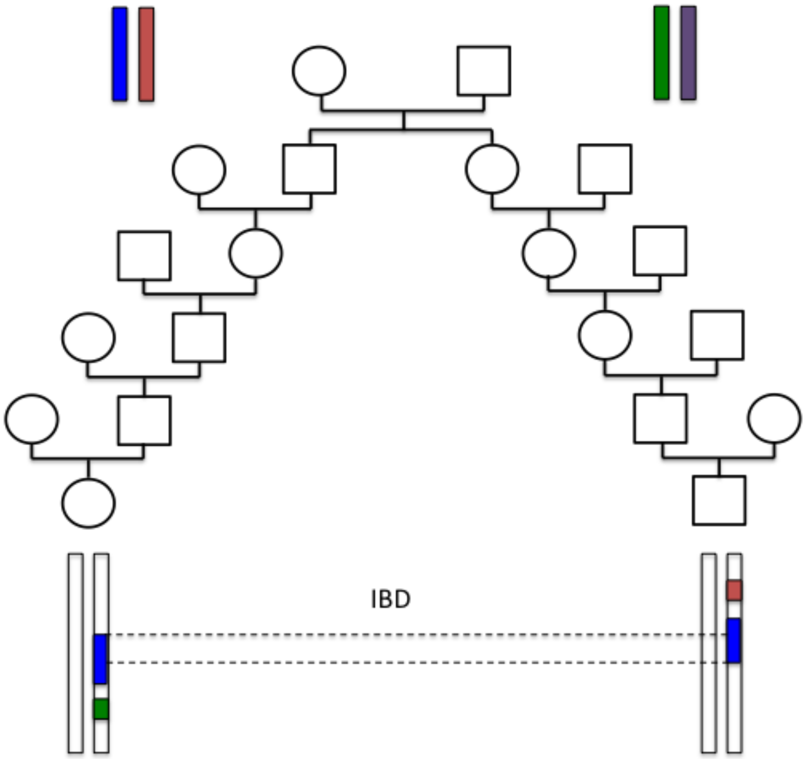


Figure 1.3: A pedigree structure where two fourth degree cousins co-inherit an IBD segment from ancestors that lived five generations in the past (figure adapted from [Browning and Browning, 2012]).

Identical-by-descent haplotypes have been extensively studied in the context of pedigree structures, particularly in early genotype-phenotype association studies, which generally involved information about the family structure of the analyzed samples ([Spielman *et al.*, 1993]), therefore several quantities regarding IBD haplotypes can be derived from pedigrees. Some basic quantities can be easily derived as follows. Consider a pair of siblings sharing two common ancestors (their parents) one generation in the past, and a single nucleotide on a haplotype along their genome. Such nucleotide may have been co-inherited by both individuals from the same copy of their parental genome, with probability $1/2$ (if the copies of the father are, for instance, A and a, the two offspring will co-inherit the same copy if both choose A, or both choose a, and the same reasoning holds for the copy they inherit from the maternal side). Now consider a pair of first degree cousins descending from these siblings. One chromosomal copy for these first degree cousins will be inherited from a parent chosen from the general population. As previously assumed, these are completely unrelated individuals, and such chromosome will not harbor an IBD locus. Focusing on the chromosome that is inherited through the lineage leading the their shared common ancestors, the probability of being IBD is $1/4$. This is due to the fact that each cousin will inherit one of the four possible copies present in their grand parents, and will choose the same with probability $1/4^2 \times 4 = 1/4$. Recursively computing this probability for the following generations, we obtain that the chance that two $(k-1)$-th degree cousins that share two diploid common ancestors k generations in the past are IBD at a chosen genomic location is $(1/4)^{k-1}$. Due to the linearity of the expectation operator, this also corresponds to the expected fraction that a pair of $(k-1)$-th degree cousins will share IBD. Note that this quantity decreases exponentially in k, and indeed after a relatively small number of generations it is very common that no IBD sharing exists at all. If IBD sharing exists, however, this typically occurs through the sharing of relatively long IBD haplotypes. If a genomic locus is shared IBD by a pair of individuals, the flanking positions along the genome are in fact typically also shared IBD, because the haplotypes that are transmitted from common ancestors are delimited by recombination events. As shown in the previous section, a recombination event may occur during meiosis between any two consecutive nucleotides. These recombination events are rare, and we can modeled their occurrence using a Poisson process with exponentially distributed waiting times between arrivals. The length of an IBD haplotype that has been transmitted from common ancestors that lived k generations in the past is therefore exponentially distributed, averaging $100/(2k)$ centimorgans. The number of IBD segments that are expected to be found for $(k-1)$-th degree cousins can be similarly computed. After $2k$ generations that separate the two cousins in the pedigree, a chromosome of genetic length l Morgans is expected to be broken into $2lk$ distinct haplotypes, each representing a potential IBD segment. The probability that one such segment is co-inherited is $(1/4)^{k-1}$, resulting in an average of $2lk(1/4)^{k-1}$ IBD segments. Again, their number can be modeled as a Poisson distributed random variable.

1.1.4.1 Definition of IBD

Despite the name, IBD segments need not be identical. Mutations in IBD segments may in fact arise during transmission from a common ancestor to her descendants, as detailed in Chapter 5. Because the number of mutations per base pair is proportional to the distance, in generations, to the common ancestor, the genomic segments transmitted to a set of individuals from very recent common ancestors will be almost identical, while regions that are co-inherited from very remote ancestors will tend to have a larger number of differences per base pair. Analyses of IBD sharing in pedigrees are usually concerned with the transmission of long IBD segments through common ancestors that span a small number of generations. These segments are therefore typically long and almost identical, and short IBD segments transmitted from very remote ancestors from the general population, which are not reported in the pedigree and are not considered members of the family, are neglected. However, when IBD sharing is detected in unrelated individuals from a population, as we do in this work, haplotypes may be co-inherited from common ancestors that lived several generations in the past, and harbor a relatively higher number of mutations. Based on these considerations, we may consider several definitions of an IBD segment:

(a) A chromosomal region transmitted from a common ancestor that lived at most t_0 generations in the past (e.g. see [Chapman and Thompson, 2003]).

(b) A chromosomal region of length at least u cM that is transmitted from a common ancestor that lived at any time in the past.

(c) A chromosomal region of length at least u cM that is transmitted *without recombination* from a common ancestor that lived at any time in the past.

While (a) is suitable in cases where t_0 is known (e.g. pedigrees) or where the focus is on modeling the descent of a known set of individuals founding a population t_0 generations in the past, this definition becomes impractical in the general case of IBD segments detected in a set of unrelateds. IBD detection in unrelated individuals usually results in a list of segments that have been discovered with a relatively high level of confidence. Often times these segments will be detected on the basis of being more similar (e.g. identical by state, IBS) compared to surrounding genomic regions. These segments will typically be transmitted from one common ancestor, generally delimited by recombination events, but their length alone is insufficient to determine the age of these segments, which has large variance for all but the very long shared haplotypes. Definitions (b) and (c) are therefore more suitable for the analysis of these segments, as no value of t_0 is assumed. In practice, current IBD detection algorithms are typically only able to reliably detect segments that are longer than a certain centimorgan length threshold, which can be accommodated in definitions (b) and (c).

In the remainder of this thesis, we use definition (c), i.e. we require that an IBD segment is transmitted from a common ancestor and is delimited by any recombination events along the lineages connecting modern day individuals to the common ancestor. Note, however, that several neighboring chromosomal regions may be merged together while still being transmitted from the same common ancestor, in which case definition (b) and (c) may not entirely overlap, depending on several factors such as population size and distance to the shared ancestor. When computing distributions of IBD sharing in chapters 3, 4 and 5, we will rely on definition (c) to derive analytical results. When using coalescent simulations to create synthetic datasets used to compare predicted and observed IBD values, however, we will compute IBD segments using definition (b), i.e. we will only require that a chromosomal region is co-inherited from the same common ancestor, without restrictions on the occurrence of recombination along these lineages, unless otherwise specified. It is evident that when very short IBD segments are considered as defined in (b) or (c), these may have a fairly large number of differences due to mutations arising along the lineages leading to extant individuals. We will still refer to these segments as IBD, although the "I" of identical may be inappropriate in this case. As we consider shared segments that are transmitted from ancestors that lived a large number of generations ago, it may be more appropriate to refer to these regions as *non-recombinant*, when definition (c) is adopted.

Chapter 2

IBD sharing in contemporary human populations

As introduced in the previous chapter, the co-inheritance of long IBD haplotypes is usually a sign of recent genetic relatedness across individuals. If the most recent common ancestor of a pair of individuals is relatively remote, the chance of finding IBD segments is very small. A pair of seventh degree cousins, for instance, will typically share no IBD segments at all. If such sharing occurs, however, the IBD haplotypes tend to be relatively long (for seventh degree cousins, for instance, IBD segments are expected to be 6.25cM long, or $\sim 4.8 \times 10^6$ base pairs, assuming a recombination rate of $\sim$1.3cM/Mb). Furthermore, if a large number of individuals is analyzed, the chance of finding IBD segments may become significant. When n individuals are analyzed, there are in fact $\binom{n}{2}$ possible pairs of IBD sharing individuals. This motivated the development of several algorithms that allow detecting IBD haplotypes in large cohorts of unrelated individuals [Purcell *et al.*, 2007; Gusev *et al.*, 2009; Browning and Browning, 2010; Browning and Browning, 2011a; Browning and Browning, 2013]. At the time the work presented in this chapter was developed, a number of large SNP array datasets comprising individuals from several human populations became available. The goal this work was to mine the presence of IBD segments in such cohorts, aiming to answer questions such as

- Are IBD haplotypes commonly found in purportedly unrelated individuals?
- Does IBD sharing reflect modern day geographic origins, and does it provide more information than other available summary statistics of genetic similarity?
- Can haplotype sharing be used to investigate a population's demographic history?
- Is the signature of natural selection visible in the distribution of IBD segments?

As discussed in the remainder of this chapter, IBD sharing was found to be pervasive in large cohorts of unrelated individuals, and was shown to be informative about both demographic and evolutionary events in human populations.

2.1 World-wide sharing of IBD segments

This section reports the results of IBD analysis performed on several large SNP array datasets, namely the HapMap 3 dataset [Frazer *et al.*, 2007], the Hebrew University Genetic Resource [HUGR, 2013], and the InTraGen Population Genetics Database (Idb, [Mitchell *et al.*, 2004; Duerr *et al.*, 2006]). Abbreviations for the distinct populations contained in these datasets can be found in Table 2.1. The results reported in this chapter, together with additional details on other analyses and the utilized datasets can be found in [Gusev *et al.*, 2012]. The work reported in this section was performed in close collaboration with Alexander Gusev.

2.1.1 IBD detection

IBD sharing was detected in the analyzed datasets using the GERMLINE software package [Gusev *et al.*, 2012]. Before analyzing the available real datasets, we assessed the accuracy of GERMLINE's IBD detection using synthetic datasets obtained using the GENOME rapid coalescent-based whole-genome simulator [Liang *et al.*, 2007]. We measured the accuracy of GERMLINE's IBD discovery using standard measures of precision (fraction of discovered segments that correspond to real IBD segments) and recall (fraction of real IBD segments retrieved). A ground-truth set for IBD segments is obtained considering all identical segments in the set of simulated haplotypes. Haplotypes were merged to form synthetic genotypes, discarding phase information. GERMLINE's *haplotype* and *genotype* extension modes were tested on both perfectly phased and computationally phased data. Discovered segments of 3 cM or longer were reported. To compute recall, GERMLINE's, IBD discovery was compared with true segments longer than 3 cM. A measure of false-positive segments was computed comparing the obtained IBD matches with segments ≥ 1 cM long in the ground-truth set.

Comparing the accuracy of both *haplotype* and *genotype* extensions on simulated data, the haplotype extension mode was found to have extremely good performance on perfectly phased data, while its recall deteriorated when computational phasing was used, as a result of unreliably reconstructed haplotypes. The genotype extension mode, on the other hand, showed a high rate of false positive IBD segment ($\sim$30% of the total) and an almost perfect recall rate. The genotype extension mode was also found to be robust to variation in the simulated demographic parameters, which, as further analyzed in Chapter 3, have an impact on phasing accuracy and therefore on the performance of the haplotype extension mode. Based on these results, and because the datasets analyzed in this work included individuals from heterogeneous populations, often with small sample sizes resulting in phasing uncertainty, GERMLINE's genotype extension mode was used for IBD detection in all reported results.

2.1.2 IBD-based graph clustering recapitulates populations structure

Although the analyzed datasets were composed entirely of purportedly unrelated individuals, IBD segments were found to be ubiquitous between and across populations, as shown in Table 2.1. To allow for population-wide analysis of IBD sharing, we built a graph model where each individual is represented as a vertex, and the amount of IBD sharing between two individuals corresponds to a single weighted edge. Building such graph for the Idb dataset results in the formation of a large connected component of individuals. The occurrence of such large connected component is extremely unlikely to occur by chance, and it indicates the presence of underlying structure in the graph (p value $< 10^{-100}$ under a hypergeometric distribution). The cohort is indeed structured, and the node membership in the connected component is highly correlated with self identification as Ashkenazi Jews (99.7% of Ashkenazi individuals

Population	Samples	Average shared genome (%)	Average segment length (cM)	% of pairs sharing IBD	Cryptic relatives
Ashkenazi Jews (AJ)	397	1.73	5.51	96.9	3

(a) Samples in the HUGR dataset.

Population	Samples	Average shared genome (%)	Average segment length (cM)	% of pairs sharing IBD	Cryptic relatives
Ashkenazi Jews (AJ)	389	1.43	5.52	99.3	2
Europeans (EU)	514	0.05	4.11	36.6	3

(b) Samples in the Idb dataset.

Population	Samples	Average shared genome (%)	Average segment length (cM)	% of pairs sharing IBD	Cryptic relatives
African Americans (ASW)	42	0.14	7.08	0.3078	4
Europeans (CEU)	109	0.48	3.77	0.9886	1
Han Chinese (CHB)	82	0.46	3.66	0.9913	0
Metropolitan Chinese (CHD)	70	0.46	3.66	0.9896	2
Gujarati Indians (GIH)	83	0.78	4.26	0.9245	5
Japanese (JPT)	82	0.77	3.71	0.9997	0
Luhya in Kenya (LWK)	83	0.80	4.98	0.9924	11
Mexicans (MEX)	45	0.96	3.87	0.9939	4
Maasai in Kenya (MKK)	143	1.06	8.58	0.9379	94
Tuscans in Italy (TSI)	77	0.4	4.23	0.9679	0
Yoruba in Ibadan (YRI)	108	0.11	4.19	0.6333	2

(c) Samples in the HMP3 dataset.

Table 2.1: Description of the samples contained in the analyzed datasets and summary of IBD sharing.

are spanned by the connected component, constituting 91.5% of the component's nodes). Overall, the total genome-wide sharing for an average pair of AJ samples (54.25 cM) is considerably higher than that of EU samples (1.81 cM).

We set out to verify the presence of similar structure in IBD sharing graphs for the HMP3 dataset. The network of shared segments in HM3 (Figure 2.1) is dense within populations and geographic regions and sparse between them. We can immediately observe an abundance of recent sharing within the cohorts, particularly in the MKK and LWK Africans; the GIH Indians. Moreover, this high level of sharing is homogeneous across most of the population and not suggestive of individual cryptic relatives. Several pairs of close relatives (defined as pairs of individuals sharing at least $1,700$cM of their genome) are found within the Maasai sample. This unexpected finding will be further discussed in Chapter 3. Looking across populations, only the JPT, CHD, and CHB East Asian groups exhibit a large number of shared segments, particularly between the two Chinese populations. The few remaining segments are also overwhelmingly within continental groups, particularly between CEU and TSI.

To investigate the ability to recapitulate population structure using the observed IBD sharing, we refined the construction of the IBD graph to allow downstream clustering analysis. In the constructed IBD graph, the weight of an edge between a pair of individuals is proportional to the sum of the length (in centiMorgans) of the IBD segments shared between the individuals. To account for the higher informativeness of rarely shared regions, the sum is normalized by the region-specific frequency of sharing in the entire population. More formally, given a set of n ordered SNPs $s \in \{1 \ldots n\}$, we define a function to represent the normalized length of an interval between two SNPs as follows:

$$F(s) = \begin{cases} \frac{l(s,s+1)}{\pi(s,s+1)} & \text{if } \pi(s,s+1) \neq 0, \\ 0 & otherwise. \end{cases} \tag{2.1}$$

where $l(s, s+1)$ is the length of the segment $[s, s+1]$, and $\pi(s, s+1)$ is the number of individuals sharing the segment $[s, s+1]$. The maximum normalized length (all SNPs being shared by a pair of individuals) is then:

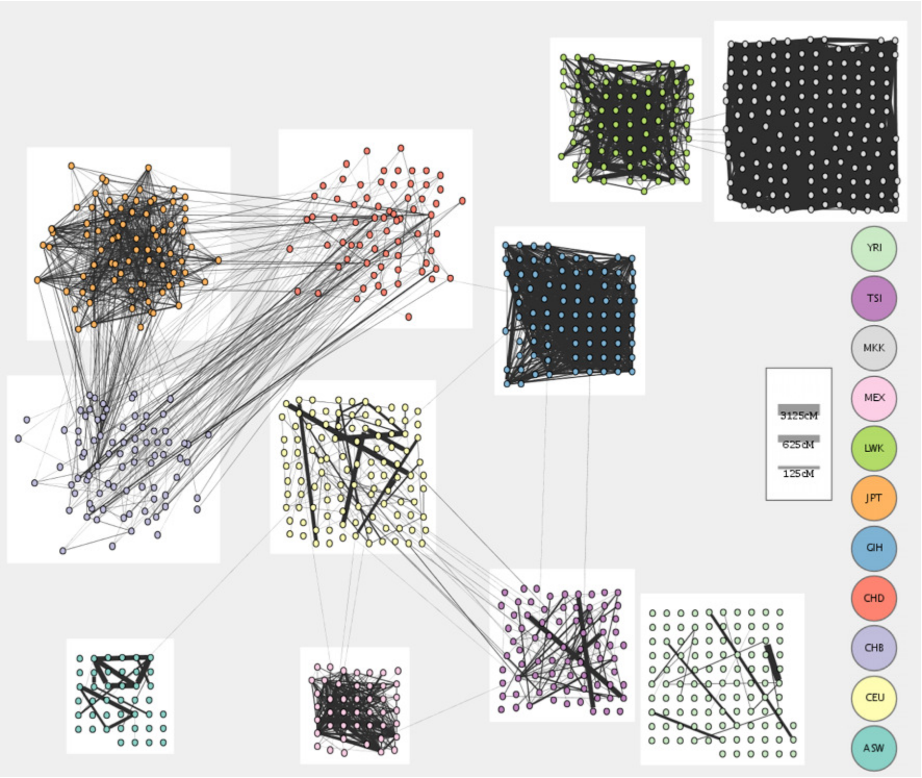


Figure 2.1: The IBD sharing graph for HMP3 samples.

$$W_{tot} = \sum_{s=1}^{n} F(s) \tag{2.2}$$

For each pair of individuals i and j sharing a set of segments K, we compute a raw edge weight normalizing the total shared length by the maximum normalized segmental length:

$$W_{ij} = \frac{1}{W_{tot}} \sum_{r \in K} \sum_{t=k_{i,r}}^{k_{e,r}} F(t) \tag{2.3}$$

Where $k_{i,r}$ and $k_{e,r}$ are the first and the last SNPs in the segment r.

The obtained value is representative of the total sharing between the two individuals and ranges between 0 (i.e., no sharing) and 1 (i.e., sharing of the whole genome). To account for the exponential decrease in the segmental length that occurs with the number of meioses, we use the weight $w_{ij} = \log(W_{ij})$ on the edges in our clustering calculations.

After constructing such graph, we performed graph clustering using the Markov Cluster Algorithm (MCL), detailed in [van Dongen, 2000]. MCL detects clusters based on the recurrence of a random walk across a weighted graph. We run MCL with default parameters as well as the *force-connected* flag which adjusts the output clusters to ensure that they are connected components. We performed the clustering in an iterative procedure that seeks to find the underlying population structure as well as identify genetic regions that are shared between clusters. The procedure starts considering all shared segments longer than 3 cM and performs the following analysis in each iteration:

1. Compute the sharing graph from the current set of shared segments. This weighted graph is then provided as input for MCL, which identifies clusters of increased relatedness.

2. Calculate the probability that a genomic locus is shared across the identified clusters, and identify any region enriched for cross-cluster sharing (1 standard deviation above the genome-wide mean).

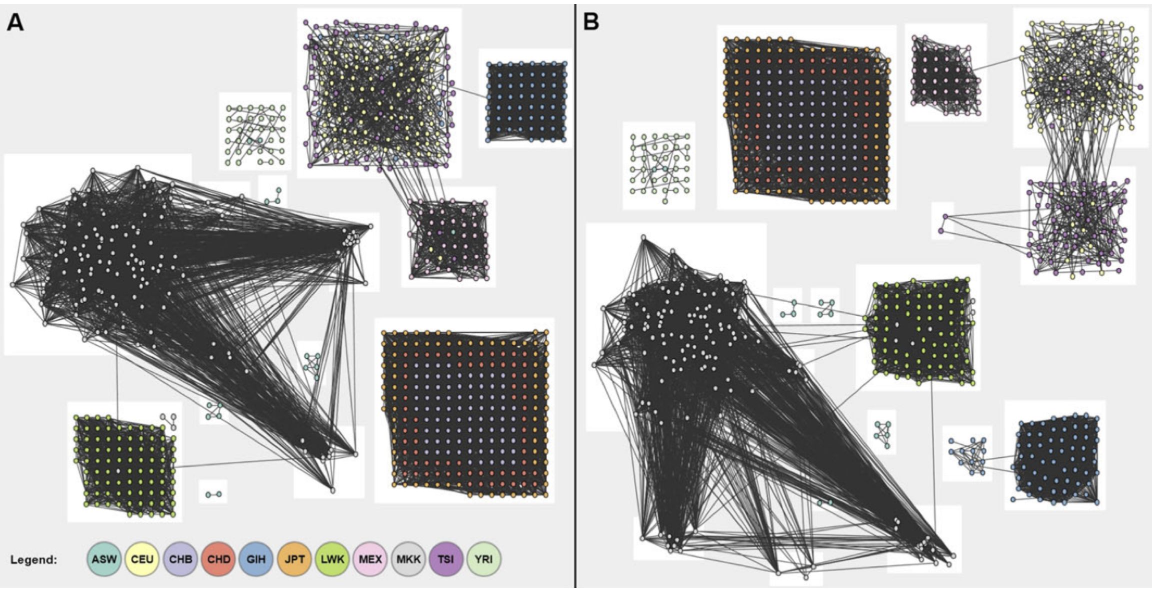


Figure 2.2: Clusters emerging from the IBD sharing graph in the HMP3 dataset reflect population structure. (A) Initial clusters from unfiltered sharing, where {GIH},{LWK},{JPT,CHD,CHB},{CEU,TSI} segregate. (B) Final clusters after cross-cluster edges have been iteratively removed, where {TSI},{CEU} newly segregated.

3. Excise all enriched cross-cluster regions as well as any affected matches that overlapped these regions and were shortened below 3 cM. The un-excised data are used as input for the next iteration.

This iterative process eventually converges when no further excision is made. Applying this procedure to the IBD sharing graph of the HMP3 dataset, we indeed recover underlying population structure. The final clusters demonstrate improved resolution between populations, with six cross-cluster regions remaining, as shown in Figure 2.2.

2.1.3 IBD sharing provides insight into recent demographic history

Further investigating the substantial IBD sharing in the Ashkenazi Jewish cohort, we examined the frequency distribution of shared IBD segments as a function of their genetic length (Figure 2.3). Based on simulations, we noticed that the slope of such distribution is not

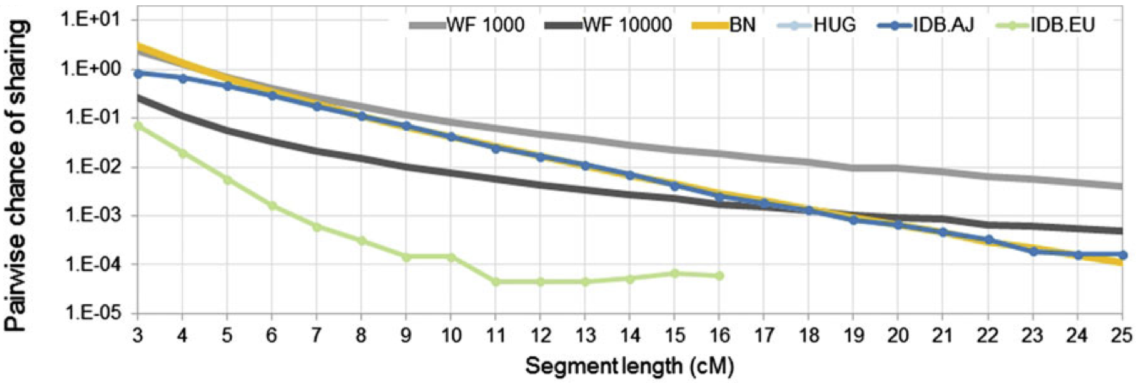


Figure 2.3: Frequency of IBD sharing as a function of genetic length in the AJ and CEU cohorts, and comparison to synthetic datasets.

compatible with the slope obtained in populations of constant size (Wright-Fisher populations). A population expansion, however, results in a steep exponential decrease compatible with what is observed in the AJ cohorts.

To obtain an initial rough estimate of an expansion rate that is compatible with the one observed in the AJ data, we considered an idealized extreme bottleneck-expansion scenario where a population is formed by one individual G generations before present, and infinite individuals from generation G to present. In such a scenario, all coalescent events happen at generation G. For a population that underwent an extreme bottleneck-expansion at generation G, two contemporary individuals are expected to share a number of segments of length l proportional to $p(1-p)^{2Gl}$, where the length is expressed in centiMorgans, and $p = 0.01$ represents the chance of a recombination event along one unit of length for a shared segment at each generation. G can be computed from N_l and N_{l+1} as:

$$\frac{N_{l+1}}{N_l} = 0.99^{2G} \tag{2.4}$$

therefore

$$G = \frac{\log(\frac{N_{l+1}}{N_l})}{2\log(0.99)} \tag{2.5}$$

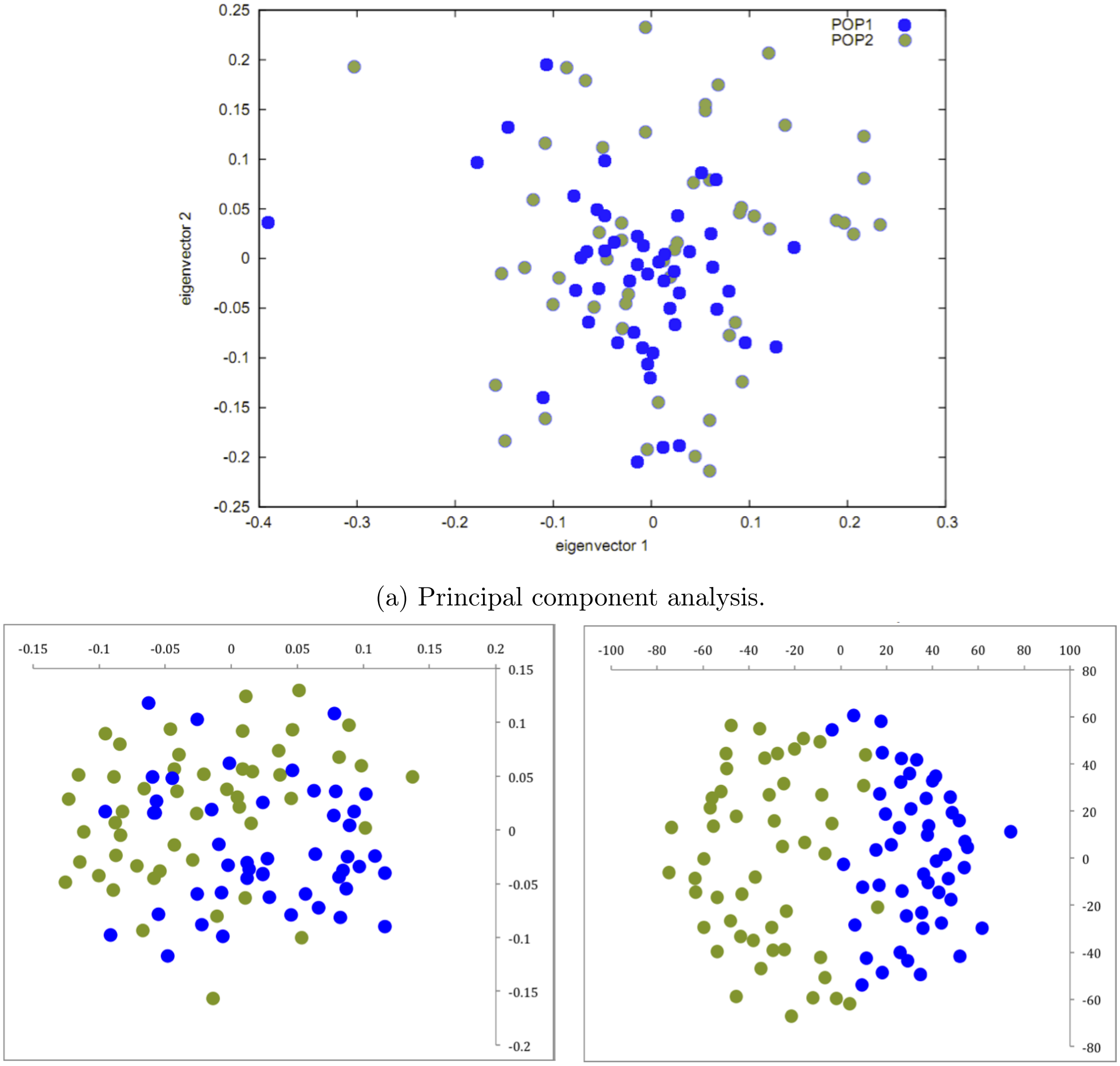
The observed exponential decay of 0.671 per cM (std 0.055) is consistent in this model with a bottleneck-expansion event occurred around 20 generations before present. We refined this estimate using extensive simulations, performing grid search in a richer parameter space (timing of the bottleneck, ancestral population size, and current population size) using a demographic model of exponential expansion (for details on these simulations, see [Gusev *et al.*, 2012]). We observe the effect of the ancestral population size are mostly noticeable on the frequency of short IBD segments, whereas the current population size mainly affects the longer segments. The timing of the bottleneck affects the entire distribution, with stronger effects on midrange segments. Our grid search suggests a rapid expansion of about 950 diploid individuals 23 generations before present to current hundreds of thousands. More complex models than those tested in this analysis may be required to explain the deviation observed for segments shorter than 5 cM (see Chapter 3). The estimated timing is compatible with a model of AJ population structure inferred from historical data in [Slatkin, 2004] and can be reconciled with previous analysis of rare mutations [Risch *et al.*, 2003] and mithocondrial data [Behar *et al.*, 2006]. Although significant admixture can be shown to influence the sharing distributions, our use of a single-population model seems reasonable due to the limited amount of recent sharing observed between European and Ashkenazi samples and by the strong similarity of the length distributions for AJ individuals sampled in Israel and USA (Idb.AJ and HUGR, see Materials and Methods). In other populations, the number of shared-segment pairs is smaller (Table 2.1) and does not yet allow for robust inference of demography.

The analysis of demographic events that occurred in the very recent history of the AJ population suggested that summary statistics of IBD sharing are informative about extremely recent demographic events. To test whether these insights may also be obtained using other methods available at the time this study was performed, we simulated a population split occurring 50 generations before present. A population of $50,000$ individuals splits into two groups of $49,000$ and $1,000$ individuals. The smaller group then exponentially expands to reach size $5,000$ individuals. We sampled 50 diploid individuals from each of these two modern groups, and analyzed realistic genotype data using several methods to investigate population structure (Figure 2.4). When principal component analysis was used to obtain a lower dimensionality projection of the data ([Price *et al.*, 2006]), little or no population structure became evident. We subsequently built a matrix representing the relatedness of individuals based on their identity-by-state (IBS), and performed multidimensional scaling using such matrix. While the subdivision of the two groups starts being visible in this case, a clear distinction is only obtained when the similarity matrix is built using IBD sharing, indicating that methods relying on summary statistics of haplotype sharing may in some cases outperform methods based on other classical genomic features.

2.1.4 Regions of increased IBD sharing are enriched for structural variation and loci implicated in natural selection

In order to examine locus-specific phenomena, we focus our analysis on local segment sharing due to intermediate and remote relatedness rather than genome-wide sharing between close relatives. IBD sharing is detected everywhere along the genome, averaging population-specific background levels (Figure 2.5). We analyzed the physical distribution of IBD sharing within and across populations, observing regions with a much higher amount of sharing than expected. Analyzing AJ samples, the most prominent such region is the human leukocyte antigen (HLA) locus. The entire segment of chromosome 6, between 25 and 35 Mb, is shared among individuals unrecombined at least 4-fold more than any other region in the genome (4.2-fold in Idb, 5.1-fold in HUGR). This is in accordance with previous observations of complex haplotype structure along the HLA locus [de Bakker *et al.*, 2006].

Examining the regions of intense sharing within HM3 populations, HLA still exhibits a very high sharing density for some of the populations: Western Europeans (CEU), Gujarati Indians (GIH), Luhya Kenyans (LWK), and Yoruba Nigerians (YRI). Additional regions along the genome exhibit notably high sharing densities within populations. Interestingly, many of these tend to also recur across unrelated individuals of different geographical origin. Segments at the recurrently shared regions in chromosomes 2, 4, and 8 are shared even across different continents of origin. Of particular interest may be the most commonly shared region, on chromosome 8*p*23.1, overlapping 5 Mb of a common inversion polymorphism, the



(a) Principal component analysis.

(b) Multidimensional scaling using IBS kernel.

(c) Multidimensional scaling using IBD kernel.

Figure 2.4: Comparison of principal component analysis and multidimensional scaling using IBS and IBD kernels for a recent split of two populations (represented by blue and green colors).

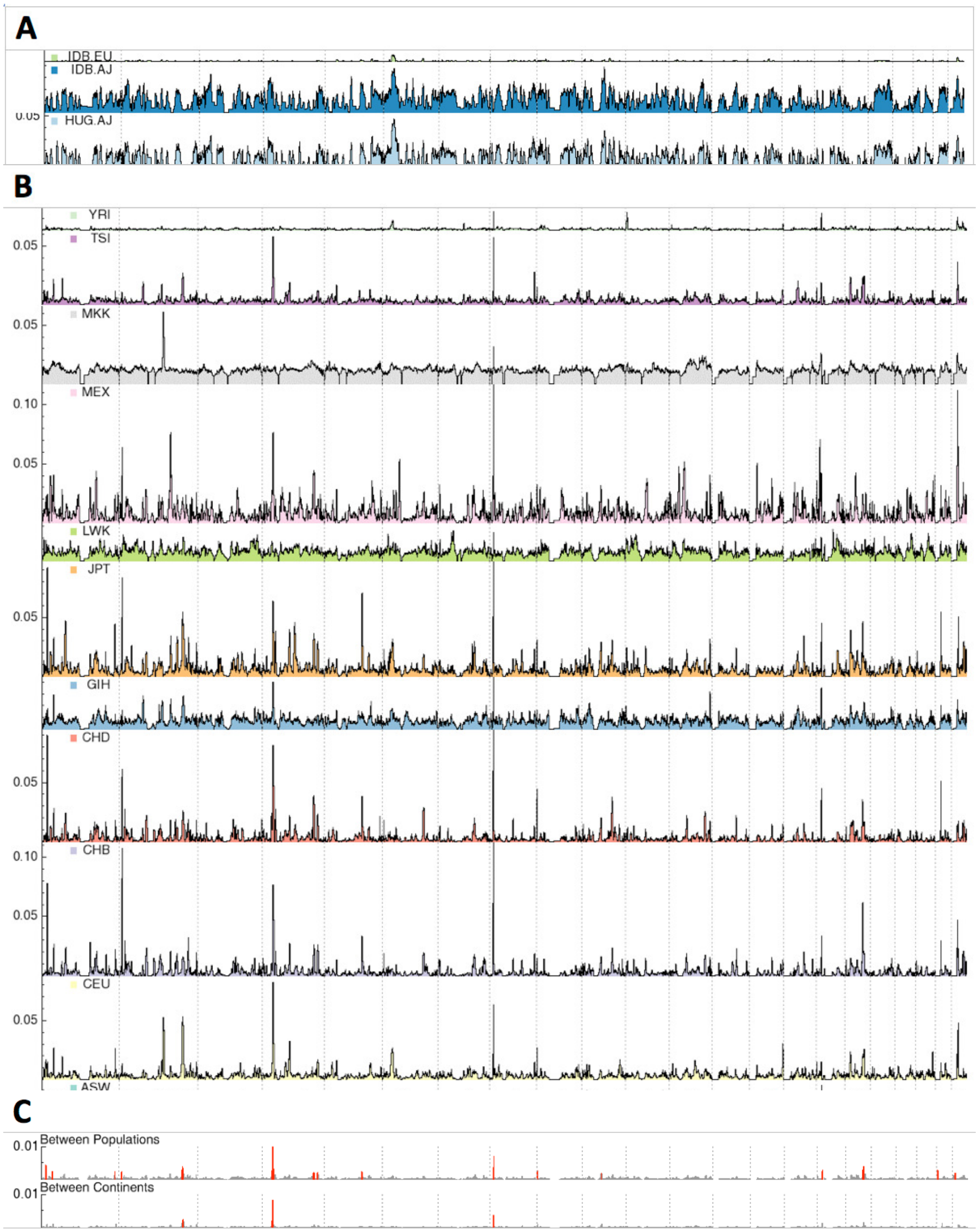


Figure 2.5: The physical distribution of IBD sharing along the genome within populations of the Idb dataset (A), the HMP3 dataset (B) and across continents/populations of the HMP3 dataset (using a different scale).

third longest reported structural variant in the entire genome [Iafrate *et al.*, 2004].

In total, the 16 cross-population commonly shared regions span only < 35 Mb ($< 0.92\%$) of the genome but account for 9.6%, 16.1%, and 18.1% of sharing within populations, between populations, and between continents, respectively. We note that these regions are not correlated to SNP density and would be unaffected by slight changes in the information content filtering. Although sharing of a region may indicate recent common ancestry, the agglomeration of shared segments at 16 loci is highly nonrandom. Biological factors or recent positive selection are possible causes of the observed reduction in haplotype diversity. Some of the identified loci correspond to previously reported regions of recent positive selection. In particular, 8 of the 16 regions were reported: $1p34.3$, $2q32.3$ [Voight *et al.*, 2006]; $4p15$ [Voight *et al.*, 2006; Sabeti *et al.*, 2002; Pickrell *et al.*, 2009]; $4q32.1$, $17q22$ [Sabeti *et al.*, 2002; Pickrell *et al.*, 2009]; $10q21.1$, $21q21.1$, $22q11.22$ [Pickrell *et al.*, 2009]; an overlap not expected by chance ($p < 0.0017$ based on permutations). Further evidence for biological retention of unrecombined ancient haplotypes, rather than random retention of new ones, comes from examining annotation for these 16 commonly shared segments. Seeking commonalities, we observe 12 of these segments to overlap structural variants that are common and long enough to have been detected in the HapMap by CGH ([Iafrate *et al.*, 2004; Perry *et al.*, 2008]). Such overlap is not expected by chance ($p < 0.00052$ in 100 longest based on permutations).

2.2 Reconstructing demographic events of the Jewish diasporas

The descriptive statistic of IBD sharing and the methods to analyze them that were developed in the previous section outline the potential of relying on shared haplotypes to gain insight into recent demographic events. In a series of three papers [Atzmon *et al.*, 2010; Campbell *et al.*, 2012; Velez *et al.*, 2012], we used these and other methods to study the signature of recent demographic variation in SNP array datasets comprising individuals from the Jewish Diaspora. The demographic events that shaped relatedness in these groups are expected to have occurred during recent millennia, and individuals from Jewish cohorts are expected to share increased IBD sharing as a result of cultural isolation following the diaspora events, motivating this analysis. In this section, we report main results and methodological development of these works, limiting the discussion to analyses of IBD sharing in these datasets. Additional analyses may be found in [Atzmon *et al.*, 2010; Campbell *et al.*, 2012; Velez *et al.*, 2012].

2.2.1 Jewish communities of the Mediterranean

Participants for this study were recruited from the Iranian (IRN, 28 samples), Iraqi (IRQ, 37 samples), Syrian (SYR, 25 samples), Ashkenazi (ASH, 34 samples), Greek Sephardic (GRK, 42 samples), Turkish Sephardic (TUR, 34 samples) and Italian (ITJ, 37 samples) Jewish communities, and included only if all four grandparents came from the same Jewish community. Subjects were excluded if they were known first- or second-degree relatives of other participants or were found to have $\hat{\pi} \geq .30$ by analysis of microarray data using the PLINK software [Purcell *et al.*, 2007]. Genotyping was performed with the Affymetrix Genome-Wide Human SNP Array 6.0 (Affy v 6). In addition to these groups, we sometimes included in the analysis a subset of populations extracted from the Human Genome Diversity Panel (HGDP), and the PopRes datasets.

IBD segments were detected with the GERMLINE algorithm in Genotype Extension [Gusev *et al.*, 2012]. The output of GERMLINE was used to detect unreported close relatives, who were omitted from the analysis. Two individuals were considered cryptic relatives if their total sharing was observed larger than $1,500$ cM and if the average segment length was more than 25 cM, suggesting an avuncular or closer relationship. The output was also used to produce sharing densities, sharing graphs, and sharing statistics.

GERMLINE output was filtered to ensure consistency across genotyping platforms and to remove noise by filtering out regions of low information content. SNP density in sliding, non-overlapping blocks across the genome was used to filter shared segments that spanned SNP-sparse regions, particularly the edges of the centromere and telomere. Specifically, regions that presented less than 100 SNPs per megabase or 100 SNPs per centimorgan were identified and excised and, subsequently, shared segments that were shorter than 3 cM were removed.

The amount of sharing for the analyzed data set was visualized with the ShareViz software, developed in [Gusev *et al.*, 2012]. As described in the previous section, individuals were represented as nodes, grouped into populations of origin. The thickness of the edges between nodes represent the total amount of sharing (in centimorgans) between each pair of individuals. For presenting populations geographically, planar quasi-isometric embedding (ISOMAP [Tenenbaum *et al.*, 2000]) was used, where distances between populations were defined as inverse of the populations' pairwise average.

To compute the average total sharing between populations I and J, the following expression was used:

$$W_{IJ} = \frac{\sum_{i \in I} \sum_{j \in J} W_{ij}}{nm} \tag{2.6}$$

where W_{ij} is the total sharing between individuals i and j from populations I and J, respectively, and n and m are the number of individuals in populations I and J. The average lengths of the shared segments across populations were computed through the arithmetic mean of the shared segments for each pair of populations.

IBD between Jewish individuals exhibited high frequencies of shared segments (Table 2.2). The median pair of individuals within a community shared a total of 50 cM IBD (quartiles: 23.0 cM and 92.6 cM). Such levels are expected to be shared by 4th or 5th cousins in a completely outbred population. However, the typical shared segments in these communities were shorter than expected between 5th cousins (8.33 cM length), suggesting multiple lineages of more remote relatedness between most pairs of Jewish individuals.

Within the different Jewish communities, three distinct patterns were observed. The Greek and Turkish Jews had relatively modest levels of IBD, similar to that observed in the French HGDP samples. The Italian, Syrian, Iranian, and Iraqi Jews demonstrated the high levels of IBD that would be expected for extremely inbred populations. Unlike the other populations, the Ashkenazi Jews exhibited increased sharing of segments at the shorter end of the range (i.e., 5 cM length), but decreased sharing at the longer end (i.e., 10 cM) (Figure

	N[a]	IRN	IRQ	SYR	ASH	ITJ	GRK	TUR	N_Italian	Sardinian	French	Basque	Adygei	Russian	Palestinian	Druze	Bedouin
IRN	29	41.95	4.91	1.00	0.75	0.61	0.56	0.75	0.68	0.67	0.50	0.58	0.60	0.47	0.53	0.66	0.57
IRQ	40	4.91	33.36	3.14	0.83	0.86	0.77	1.04	0.74	0.68	0.62	0.66	0.50	0.52	0.64	0.64	0.61
SYR	25	1.00	3.14	17.26	1.93	1.57	1.57	2.05	0.86	0.97	1.00	0.85	0.66	0.62	0.60	0.75	0.56
ASH	34	0.75	0.83	1.93	11.62	3.09	2.15	2.95	1.01	1.10	1.01	1.15	0.74	0.91	0.58	0.78	0.58
ITJ	37	0.61	0.86	1.57	3.09	28.45	2.48	2.41	0.98	0.85	0.95	0.86	0.75	0.82	0.66	0.67	0.57
GRK	42	0.56	0.77	1.57	2.15	2.48	6.01	2.56	0.91	0.96	0.89	0.93	0.81	0.64	0.61	0.74	0.54
TUR	34	0.75	1.04	2.05	2.95	2.41	2.56	4.46	0.90	0.95	0.94	0.90	0.70	0.81	0.71	0.75	0.59
N_Italian	21	0.68	0.74	0.86	1.01	0.98	0.91	0.90	2.37	1.39	1.36	1.43	0.84	1.24	0.51	0.66	0.61
Sardinian	28	0.67	0.68	0.97	1.10	0.85	0.96	0.95	1.39	10.84	1.35	1.47	0.65	0.93	0.67	0.71	0.55
French	29	0.50	0.62	1.00	1.01	0.95	0.89	0.94	1.36	1.35	1.63	2.08	0.83	1.46	0.55	0.59	0.53
Basque	24	0.58	0.66	0.85	1.15	0.86	0.93	0.90	1.43	1.47	2.08	15.97	1.07	1.21	0.59	0.67	0.54
Adygei	17	0.60	0.50	0.66	0.74	0.75	0.81	0.70	0.84	0.65	0.83	1.07	6.29	0.91	0.56	0.80	0.39
Russian	25	0.47	0.52	0.62	0.91	0.82	0.64	0.81	1.24	0.93	1.46	1.21	0.91	5.80	0.48	0.57	0.37
Palestinian	51	0.53	0.64	0.60	0.58	0.66	0.61	0.71	0.51	0.67	0.55	0.59	0.56	0.48	25.50	0.62	1.01
Druze	47	0.66	0.64	0.75	0.78	0.67	0.74	0.75	0.66	0.71	0.59	0.67	0.80	0.57	0.62	49.59	0.65
Bedouin	48	0.57	0.61	0.56	0.58	0.57	0.54	0.59	0.61	0.55	0.53	0.54	0.39	0.37	1.01	0.65	25.36

Table 2.2: IBD sharing within and across Jewish communities and other populations from the HGDP and PopRes datasets.

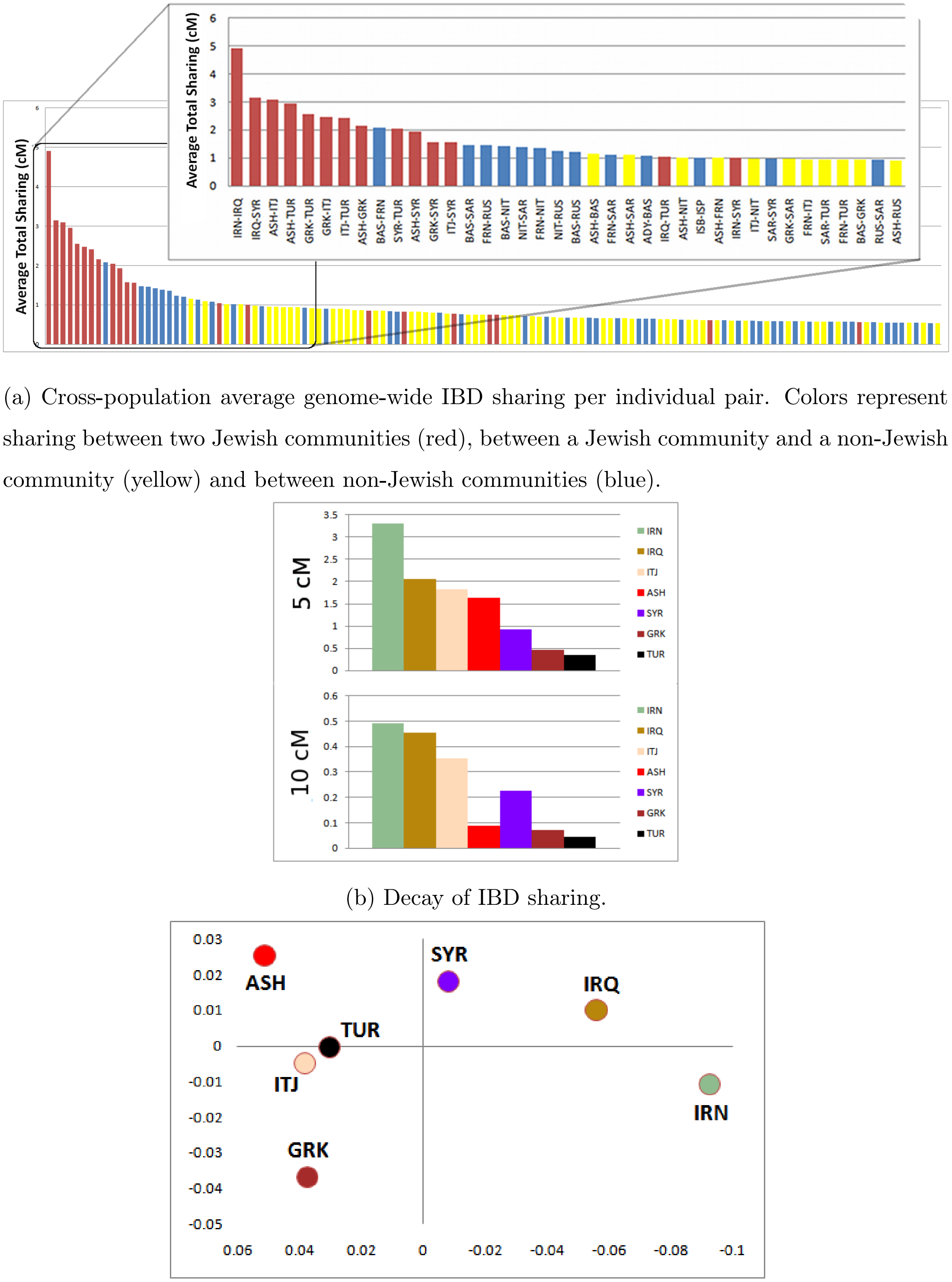
[a] number of samples

2.6b).

As expected, the vast majority of long shared segments (89% of 15 cM segments, 78% of 10 cM segments) were shared within communities. However, the genetic connections between the Jewish populations became evident from the frequent IBD across these Jewish groups (63% of all shared segments). The web of relatedness between the 27, 966 pairs of individuals in this study was intricate, even if restricted only to the 2, 166 pairs sharing a total 50 cM or more, a level of sharing among third cousins (Figure 2.7). When population averages were examined, this network of IBD was consistent with the geographic distances between populations, with planar embedding representing 93% of the initial information content (Figure 2.6c). The notable exception was that of Turkish and Italian Jews who were nearest neighbors in terms of IBD, but more distant on the geographical map, potentially reflecting their shared Sephardic ancestry. Jewish populations shared more and longer segments with one another than with non-Jewish populations, highlighting the commonality of Jewish origin. Among pairs of populations ordered by total sharing, 12 out of the top 20 were pairs of Jewish populations, and none of the top 30 paired a Jewish population with a non-Jewish one (Figure 2.6a).

2.2.2 IBD sharing is enriched for Sephardic ancestry in modern Latino populations

Modern day Latin America resulted from the encounter of Europeans with the indigenous peoples of the Americas in 1492, followed by waves of migration from Europe and Africa. As a result, the genomic structure of present day Latin Americans was determined both by the genetic structure of the founding populations and the numbers of migrants from these different populations. In ([Velez *et al.*, 2012]), we analyzed DNA collected from two well-established communities in Colorado (Hispanos, 33 unrelated individuals) and Ecuador (Lojanos, 20 unrelated individuals) with a measurable prevalence of the $BRCA1$ $c.185delAG$ and the GHR $c.E180$ mutations, respectively, using Affymetrix Genome-wide Human SNP 6.0 arrays to identify their ancestry. These mutations are found at relatively high frequency in Sephardic Jewish individuals, suggesting they may have been brought to these



(a) Cross-population average genome-wide IBD sharing per individual pair. Colors represent sharing between two Jewish communities (red), between a Jewish community and a non-Jewish community (yellow) and between non-Jewish communities (blue).

(b) Decay of IBD sharing.

(c) Isomap embedding of IBD sharing.

Figure 2.6: Summary of IBD sharing for Jewish communities.

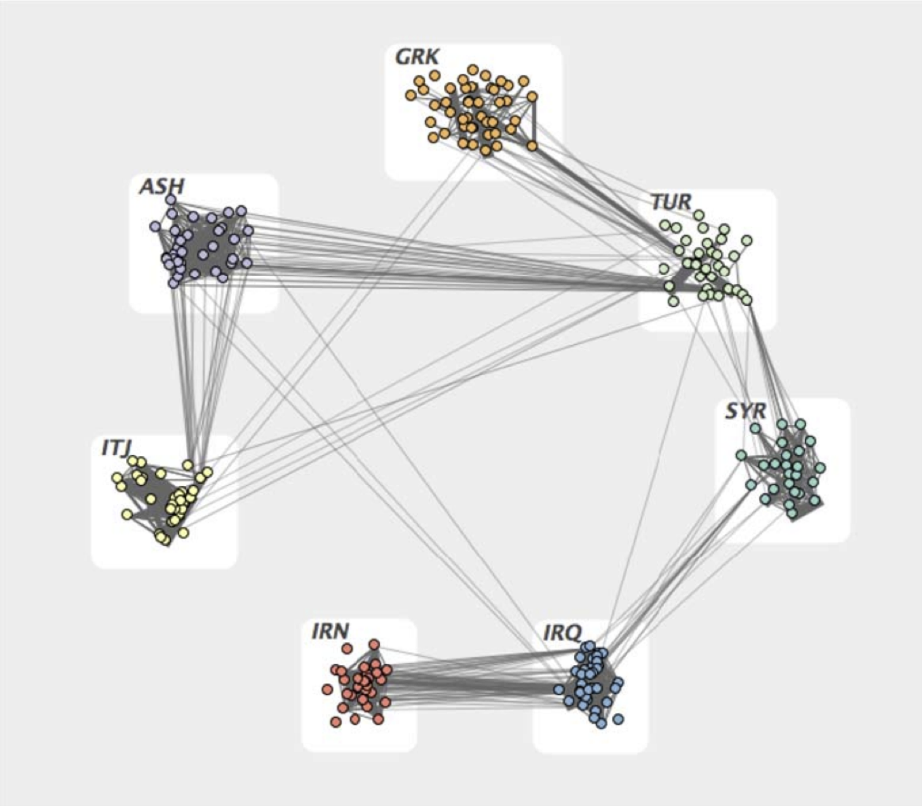


Figure 2.7: IBD sharing graph for the Jewish Hapmap groups.

communities through Jewish migration from the realms that comprise modern Spain and Portugal during the Age of Discovery. In this work, several analyses identified enrichment for Sephardic Jewish ancestry. We here report a summary of IBD sharing analysis performed in this dataset.

For this analysis, the Hispano and Lojano datasets were combined with (1) 237 samples from the Jewish HapMap Project (Affymetrix 6.0), including Iranian, Iraqi, Syrian, Italian, Turkish, Greek and Ashkenazi Jews [Atzmon *et al.*, 2010], described in the previous section; (2) 4 US Hispanic/ Latino populations (27 Dominicans, 26 Colombians, and 20 Ecuadorians, as well as 27 Puerto Ricans) from Illumina 610 K arrays [Bryc *et al.*, 2010]; (3) 50 US Mexican samples from HapMap3 (Affymetrix 6.0) [Altshuler *et al.*, 2010]. We phased the genotype data for each group using the Beagle software package [Browning and Browning, 2007], then detected IBD segments using GERMLINE [Gusev *et al.*, 2009] in Genotype Extension mode (preferred to the haplotype mode due to heterogeneous sample size and demographic background of the analyzed groups). The identified segments were used to exclude close relatives (sharing at least 800 cM and at least ten segments of length $\geq$ 10 cM) from the analysis, obtain statistics on the average total IBD sharing within and across groups and identify cross-population regions of increased sharing. The total sharing between an average pair of individuals from two different populations was computed summing the length (in cM) of all IBD segments detected across the two populations and normalizing by the number of possible pairs of individuals (the product of the cardinality for the two groups). We normalized by $\binom{n}{2}$ possible pairs when computing the average total sharing within a population of sample size n.

Identity-by-descent showed elevated cross-population sharing between Hispano, Lojano and Mexican samples. The frequency of identity-by-descent (IBD) between unrelated individuals in a population is indicative of effective population size [Wright, 1931]. We therefore analyzed the average genome-wide levels of IBD sharing within Latino ethnic groups. IBD sharing within Hispano and Lojano samples was higher than within other populations in this study, suggesting correspondingly higher levels of endogamy (Table 2.3a). We further analyzed rates of IBD sharing across different groups to investigate shared ancestry. Ele-

ASH	IRN	IRQ	SYR	ITL	GRK	TUR	HSP	LSN	MEX	TSI	CEU	CHB	YRI
77.0	170.0	152.2	89.8	130.6	50.2	42.2	113.3	131.0	71.3	49.6	59.9	75.2	8.9

(a) IBD sharing within Jewish communities and other populations from the HapMap dataset.

	ASH	IRN	IRQ	SYR	ITL	GRK	TUR	HSP	LSN	MEX	TSI	CEU	CHB
IRN	11.92												
IRQ	16.02	31.73											
SYR	17.96	15.43	31.59										
ITL	24.51	15.20	24.59	25.74									
GRK	21.35	15.05	24.99	27.00	33.68								
TUR	22.93	15.61	26.24	28.72	33.55	33.69							
HSP	12.16	10.16	17.41	16.69	18.81	18.96	19.87						
LSN	8.36	7.73	12.41	11.62	13.01	12.74	13.57	43.75					
MEX	10.34	9.25	14.80	14.74	15.95	15.95	17.15	52.07	53.69				
TSI	19.08	17.57	28.17	27.22	30.15	30.13	31.10	26.37	17.80	22.71			
CEU	19.69	16.37	25.97	25.42	29.45	29.28	30.83	30.05	20.55	25.75	45.31		
CHB	1.88	1.87	3.18	2.42	2.52	2.65	2.69	10.43	10.03	12.15	3.56	3.88	
YRI	0.01	0.01	0.03	0.02	0.03	0.02	0.02	0.05	0.02	0.08	0.03	0.02	0.02

(b) IBD sharing across Jewish communities and other populations from the HapMap dataset.

Table 2.3: IBD sharing within and across Jewish communities and other populations from the HapMap dataset.

vated cross-population sharing between Hispano, Lojano and Mexican samples (Table 2.3b) was consistent with shared recent ancestry. When investigating potential shared ancestry between these groups and other populations, we observed that multiple populations shared segments IBD with Latinos (Table 2.3b). More specifically, highest rates of such Latino-IBD sharing were observed in European and Tuscan samples followed by Sephardic and Mizrahi (Iranian, Iraqi and Syrian) Jewish communities. Lower rates of IBD were observed versus Ashkenazi samples in the Lojano samples, and to the Chinese group in Hispanos and Mexicans. Negligible IBD sharing with Yoruba samples was observed for all populations.

Besides detecting IBD sharing, we used the Xplorigin software package [Bonnen *et al.*, 2009] to investigate the proportion of European, Native American and Jewish ancestry of Hispano and Lojano samples in comparison to another Hispanic/Latino cohort from Mexico. Xplorigin builds a database of short haplotype frequencies for three reference populations, which are assumed to be the source of admixture for a studied group of samples. The haplotype frequencies are probabilistically used to assign locus-specific ancestry proportions to the analyzed individuals. Ancestry deconvolution was also applied to investigate the remote origin of regions shared IBD across populations.

We trained the Xplorigin software using 98 randomly selected phased haplotypes from the following groups: European Basque and French from the HGDP dataset; Sephardic Italian, Greek and Turkish from the Jewish HapMap dataset; Native American Pima, Surui and Maya samples from the HGDP dataset. After pruning some markers during computational phasing, the number of makers used for this cross-platform analysis was $150,157$ SNPs. For each of the three reference groups we determined LD blocks and the frequency of haplotypes and transitions between haplotypes using Haploview [Barrett *et al.*, 2005]. The genome was then partitioned into short haplotype blocks, and Xplorigin's hidden Markov model was used to assign the most likely proportion of ancestry from the three reference populations to each observed individual.

We analyzed the proportions of ancestry in correspondence of IBD segments within and across populations. To overcome phase uncertainty for an IBD segment shared by two individuals, we considered the ancestry of both maternal and paternal chromosomes reported by Xplorigin in correspondence of the IBD region. The values reported in Table 2.4 were computed as follows: given a number of IBD segments between individuals of two populations $P1$ and $P2$, we report the average proportion of IBD ancestry of individuals from $P1$ in position $P1 - P2$ of the table, and the average proportion of IBD ancestry of individuals from $P2$ in position $P2 - P1$. The ancestry of IBD sharing within a population (table entries in positions $P1 - P1$) was computed for both individuals of an IBD sharing pair. The reported mean ancestry proportion is computed as the genome-wide average ancestry proportion. To test for significance of the differences between genome-wide ancestry proportion and IBD ancestry proportions we performed random permutations of the IBD segments. We randomly shuffled IBD segments between populations $P1$ and $P2$, testing the ancestry proportions for the permuted set of IBD segments. The deviation from the genome-wide averages in correspondence of IBD segments was never observed for $1,000$ random permutations of each table entry.

Ancestry deconvolution showed sharing compatible with a history of Latino admixture with Europeans, Native Americans and Sephardic Jews. Many Latino populations are well known to include genetic ancestry components from Native Americans, Europeans and Africans, all admixed within the last 20 generations. Comparing potential European, Sephardic Jewish and Native American ancestry, we observed proportions compatible with a history of Latino admixture from these three ethnicities (Table 2.4). The Hispano samples showed increased European ancestry, whereas the Lojanos and Mexicans showed increased Native American ancestry. We further considered the IBD-shared segments among Latino samples, to explore correlation between the occurrence of such segments and admixture source population. Interestingly, these segments across all examined Latino populations were substantially enriched for Native American ancestry. As such segments indicate a recent common ancestor of the samples who share them. This indicates a small number of recent Native American founders, relative to other source populations. When considering IBD sharing in each Latino group separately, we further observed IBD sharing is also enriched for Sephardic ancestry ($p < 0.001$) within the Lojano community. The relative enrichments in Sephardic versus European ancestry in IBD-shared segments proved robust

	MEAN	MEX	HSP	LSN
MEX	0.309	0.206	0.239	0.205
HSP	0.362	0.276	0.344	0.265
LSN	0.312	0.22	0.239	0.291

(a) European ancestry.

	MEAN	MEX	HSP	LSN
MEX	0.297	0.177	0.205	0.171
HSP	0.342	0.226	0.327	0.211
LSN	0.305	0.203	0.232	0.342

(b) Sephardic ancestry.

	MEAN	MEX	HSP	LSN
MEX	0.394	0.617	0.557	0.624
HSP	0.296	0.498	0.328	0.524
LSN	0.382	0.576	0.53	0.367

(c) Native American ancestry.

Table 2.4: Enrichment of ancestral components in IBD segments (red colors indicate statistically significant enrichment).

to the choice of a source population, showing similar results when compared to Yoruba, who likely did not contribute significantly in terms of ancestry (see Table 2S in [Velez *et al.*, 2012]).

2.2.3 Jewish communities in North Africa

Methods developed in the previous sections for the analysis of IBD sharing and ancestry deconvolution were adopted to analyze a dataset comprising several North African Jewish communities, together with samples from the previously described Jewish communities and several other North African populations, for a total of 509 Jewish samples from 15 populations and 114 non-Jewish individuals from seven North African populations [Henn *et al.*, 2012] (samples listed in Table 2.5). Details of this and other analysis can be found in [Campbell *et al.*, 2012].

IBD discovery was performed as previously described, although only Jewish samples and non-Jewish populations from the same geographic regions were analyzed to maintain high-density of SNP markers in the cross-platform analysis. Ancestry deconvolution using Xplorigin was run on a subset of the analyzed populations, with respect to their Maghrebi, Middle Eastern, and European ancestry, using 36 non-Jewish Tunisian Berber, 48 Palestinian, and 48 Basque reference haplotypes, respectively.

As in the previously analyzed Jewish communities, North African Jewish populations showed a high degree of endogamy and IBD sharing between Jewish groups. We studied the frequency of IBD haplotypes shared by unrelated individuals within and across the analyzed groups. When IBD within populations was examined, the non-Jewish Tunisian Berbers exhibited the highest level of haplotype sharing, suggesting a small effective population size and high levels of endogamy (Figure 2.8, panel A). With the exception of this Tunisian cohort, the Jewish populations generally showed higher IBD sharing than non-Jewish groups, indicating greater genetic isolation.

The relationships of the Jewish communities were outlined further by the IBD sharing across populations (Figure 2.8, panels B and C), because the Jewish groups generally demonstrated closer relatedness with other Jewish communities than with geographically

Population ID	Female	Male	Total	Population
ALGJ	23	1	24	Algerian Jewish
ASHJ	14	20	34	Ashkenazi Jewish
DJEJ	0	17	17	Djerban Jewish
ETHJ	13	3	16	Ethiopian Jewish
GEOJ	4	9	13	Georgian Jewish
GRKJ	25	29	54	Greek Jewish
IRNJ	22	27	49	Iranian Jewish
IRQJ	25	28	53	Iraqi Jewish
ITAJ	20	19	39	Italian Jewish
LIBJ	31	6	37	Libyan Jewish
MORJ	32	6	38	Moroccan Jewish
SYRJ	15	21	36	Syrian Jewish
TUNJ	24	5	29	Tunisian Jewish
TURJ	24	10	34	Turkish Jewish
YMNJ	36	0	36	Yemini Jewish
ADYG	10	7	17	Adygei
ALGE	9	9	18	Algerian
BASQ	8	16	24	Basque
BEDN	20	27	47	Bedouin
DRUZ	32	13	45	Druze
EGYP	0	19	19	Egyptian
FREN	17	12	29	French
LIBY	1	16	17	Libyan
MORN	0	18	18	N Moroccan
MORS	5	5	10	S Moroccan
MOZA	9	19	28	Mozabite
NITA	7	14	21	N Italian
PALN	34	17	51	Palestinian
RUSS	9	16	25	Russian
SARD	12	16	28	Sardinian
SOCC	0	17	17	Saharan
TUNI	0	15	15	Tunisian
AFRI	1	24	25	Sub-Saharan African
ASIA	10	15	25	Asian

Table 2.5: Analyzed samples from Mediterranean and North African Jewish communities, and other non-Jewish populations.

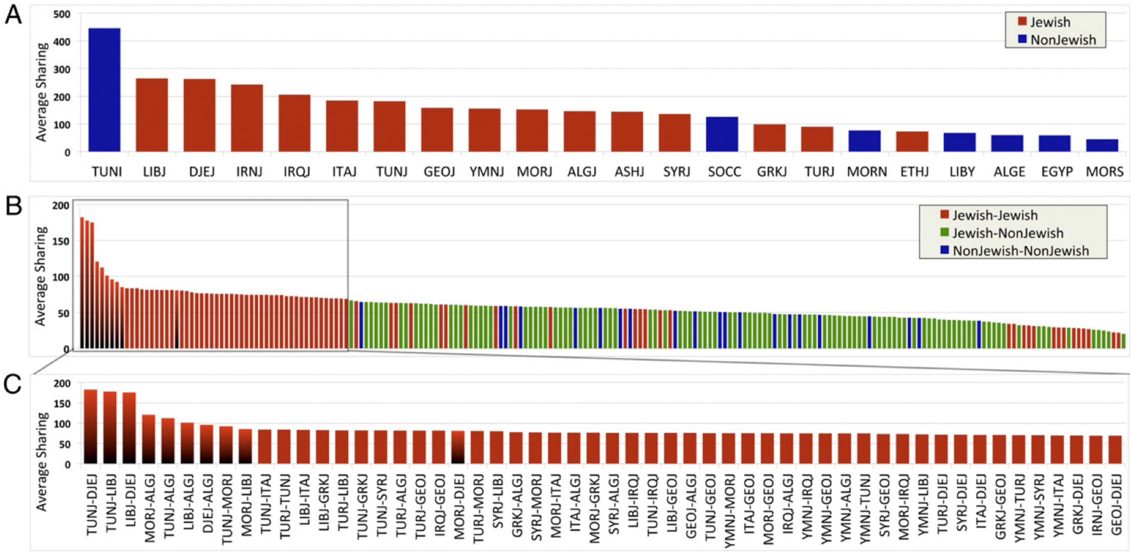


Figure 2.8: IBD sharing in North African communities. (A) Within groups (B) Across groups (C) Top sharing population pairs .

near non-Jewish populations. In particular, North African Jewish communities showed some of the highest levels of cross-population IBD sharing for the average pair of individuals. A strong degree of relatedness was observed across individuals from the Djerban, Tunisian, and Libyan Jewish communities. Noticeable proximity was also found between Jewish Algerian samples and other North African Jewish cohorts such as Moroccan, Tunisian, Libyan, and Djerban Jews, and across individuals from the Tunisian and Moroccan Jewish groups. Among non-Jewish North African groups, Algerians, South Moroccans, and West Saharan samples were found to share, on average, a smaller proportion of their genome IBD to other cohorts.

By using Xplorigin to perform ancestry deconvolution for a subset of the populations, the Maghrebi (Tunisian non-Jewish), European (Basque), and Middle-Eastern (Palestinian) ancestry components of North African Jewish communities were compared with the corresponding non-Jewish groups (Figure 2.9). A stronger signal of European ancestry was found in the genomes of Jewish samples, with a decreased fraction of Maghrebi origins, whereas the Middle Eastern component was comparable across groups. In Jewish groups, geographical

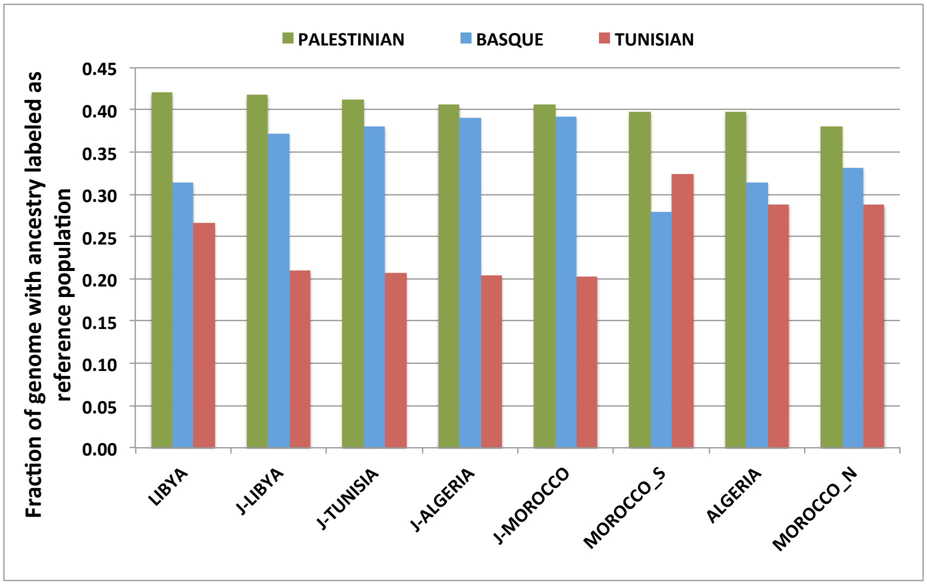


Figure 2.9: Ancestral deconvolution of Jewish and non-Jewish North African Communities.

proximity to the Iberian Peninsula correlated with an increase in European ancestry and a decrease in Middle Eastern ancestry, whereas the Maghrebi component was only mildly reduced. Differences in ancestry proportions were found to be significant ($p < 0.05$), except for the Maghrebi component of non-Jewish Northern Moroccan compared with non-Jewish Algerian samples, and the European component of Jewish Moroccan compared with Jewish Algerian samples.

In addition to genome-wide proportions, this ancestry painting analysis was intersected with regions that harbor long-range IBD haplotypes, as done in the analysis of Sephardic ancestry in Latin American populations. In Jewish populations, the ancestry proportions in corresponding IBD regions highlighted mild, but in some cases significant, deviations from genome-wide averages (Table 2.6), whereas stronger differences were observed in the recent ancestry for the corresponding non-Jewish communities. In these groups, recently co-inherited regions exhibited significantly increased European ancestry, with significantly decreased Maghrebi ancestry, compared with genome-wide averages. This phenomenon was generally stronger for loci shared IBD with individuals from Jewish communities. This increase in European ancestry and corresponding decrease in Maghrebi ancestry may be

	ALGE	LIBY	MORN	MORS	ALGJ	LIBJ	MORJ	TUNJ
ALGE	0.291	0.291	0.292	0.294	0.259	0.268	0.253	0.264
LIBY	0.274	0.249	0.261	0.300	0.241	0.235	0.249	0.247
MORN	0.312	0.295	0.297	0.317	0.263	0.259	0.271	0.275
MORS	0.323	0.349	0.338	0.352	0.345	0.329	0.348	0.312
ALGJ	0.205	0.203	0.207	0.221	0.192	0.194	0.196	0.199
LIBJ	0.200	0.206	0.212	0.215	0.200	0.206	0.199	0.200
MORJ	0.198	0.209	0.202	0.198	0.201	0.198	0.202	0.201
TUNJ	0.218	0.210	0.215	0.222	0.207	0.205	0.203	0.207

(a) Maghrebi ancestry.

	ALGE	LIBY	MORN	MORS	ALGJ	LIBJ	MORJ	TUNJ
ALGE	0.388	0.392	0.382	0.372	0.386	0.387	0.397	0.383
LIBY	0.401	0.425	0.421	0.405	0.417	0.431	0.419	0.417
MORN	0.368	0.364	0.377	0.361	0.377	0.395	0.375	0.377
MORS	0.371	0.362	0.362	0.367	0.350	0.355	0.344	0.350
ALGJ	0.380	0.406	0.385	0.398	0.405	0.393	0.398	0.391
LIBJ	0.394	0.401	0.395	0.395	0.398	0.410	0.402	0.398
MORJ	0.368	0.372	0.362	0.395	0.391	0.368	0.392	0.391
TUNJ	0.396	0.411	0.407	0.414	0.409	0.420	0.420	0.409

(b) Middle Eastern ancestry.

	ALGE	LIBY	MORN	MORS	ALGJ	LIBJ	MORJ	TUNJ
ALGE	0.291	0.335	0.326	0.334	0.355	0.345	0.350	0.352
LIBY	0.325	0.326	0.318	0.294	0.342	0.334	0.332	0.342
MORN	0.321	0.341	0.326	0.322	0.360	0.346	0.354	0.360
MORS	0.306	0.289	0.299	0.281	0.305	0.315	0.308	0.305
ALGJ	0.415	0.391	0.409	0.381	0.403	0.414	0.406	0.410
LIBJ	0.406	0.393	0.393	0.390	0.402	0.384	0.399	0.402
MORJ	0.435	0.418	0.436	0.407	0.408	0.435	0.406	0.408
TUNJ	0.386	0.380	0.378	0.365	0.384	0.375	0.378	0.384

(c) European ancestry.

Table 2.6: Enrichment of ancestral components in IBD segments (green color indicates statistically significant depletion, red indicates enrichment).

interpreted in several ways: (1) This increase may be due to the inherently higher European ancestry of Jewish segments planted into the genomes of non-Jewish populations. (2) Alternatively, the difference in genome-wide ancestries between Jewish and non-Jewish groups alone could explain this observation in the case of recent symmetric gene flow in both directions. However, this second scenario alone is inconsistent with the data, because it would imply a comparable decrease of European ancestry in regions IBD to non-Jewish populations to be observed in Jewish genomes. (3) The observed increase of European ancestry could be similarly explained by European segments newly planted in both populations. This explanation is also unlikely, because it would result in a comparable increase of European ancestry in Jewish genomes, which is instead observed to only mildly increase compared with genome-wide averages. The increase in European ancestry is stronger in IBD regions of length between 3 and 4 cM, compared with regions at least 4 cM long (see Supplementary Table 7 in [Campbell *et al.*, 2012]), which is compatible with European admixture occurring several generations before present, through ancestors that resided in the Iberian Peninsula.

Chapter 3

IBD sharing and demography

The results described in Chapter 2 outlined that IBD sharing in purportedly unrelated individuals does carry information about population-scale features such as demographic history, population stratification and natural selection. This motivated developing a theoretical framework that combines coalescent theory and modeling of IBD sharing in pedigree structures, which enables quantitative approaches to studying hidden relatedness in relation to these population features. In this chapter, we describe this framework and show its application in the inference of recent demographic events in two real populations: Ashkenazi Jewish and Kenyan Maasai, which were both shown to harbor substantial IBD sharing across pairs of unrelated individuals, although such sharing seems to be emerging from distinct demographic backgrounds. The remainder of the text will occasionally refer to supplementary figures and tables. These can be consulted online[1] as supplementary materials of the article [Palamara *et al.*, 2012], where additional details of the presented analysis can also be found.

[1] `http://www.sciencedirect.com/science/MiamiMultiMediaURL/1-s2.0-S0002929712004727/1-s2.0-S0002929712004727-mmc1.pdf/276895/FULL/S0002929712004727/524f0ffb44ca9b5087aa3a4c41eb0202/mmc1.pdf`

3.1 A coalescent-based model for the relationship between demographic events and shared IBD haplotypes

As shown in Chapter 1, coalescent theory [Kingman, 1982b] indicates that, at a specific locus of their genome, two haploid gametes from a Wright-Fisher population of constant (haploid) effective population size N_e have a probability of $1/N_e$ of finding a common ancestor at each generation. The time (in generations before present [gbp]) for these two individual gametes to reach a most recent common ancestor (MRCA) when their lineages are traced back into the past is geometrically distributed and has an expected value of N_e. More generally, if a population is composed of $N(g)$ haploid individuals at generation g, then the chance of finding a common ancestor at that generation is $N(g)^{-1}$, and the time distribution to a common ancestor assumes a more complex form. The relationship between the probability of finding common ancestors and the size of a population is appealing for demographic reconstruction. One can in fact study the distribution of time to a common ancestor at the average genomic locus for many pairs of individuals and can therefore gain information on a population's size across different time scales. In the proposed methodology, we rely on haplotype sharing to obtain a probabilistic estimate of the time to coalescence at any genomic site for any pair of individuals in the population at hand. The extent of a co-inherited IBD haplotype is probabilistically determined by the generation of the MRCA for the two individuals at the considered locus. Unfortunately, individual segments carry little information about specific sites unless the common ancestor is extremely recent (e.g., less than 10 gbp [Huff *et al.*, 2011]). However, because we are interested in genome-wide, population-wide summary statistics, significant information can be gathered from a large number of segments co-inherited by different pairs of individuals from the analyzed population sample. In fact, the number of considered pairs grows quadratically with the sample size, and the number of expected IBD segments increases as shorter segment lengths are considered. Leveraging these principles, we derive analytical results for the distribution of IBD sharing across purportedly unrelated individuals. As detailed below, we express these quantities as a function of historical demography in the population.

3.1.1 IBD and demographic history in Wright-Fisher populations

Formally, consider a random pair of haploid individuals sampled from the studied population and a specific locus along their genome. Note that although we present this analysis in the context of haploid individuals, the following results are easily adapted to the case of diploid individuals by the appropriate multiplication or division by a factor of two. We are interested in modeling the probability that the chosen locus is spanned by a non-recombinant IBD segment of a specific genetic length. We abstract this length as a continuous random variable L and denote its probability density function by $p(l|\theta)$, where θ encodes a parameterization of the population's demographic history. In the simplest case of a constant population size, θ is only parameterized by the constant population size N_e. We assume neutrality throughout; therefore, this is a Wright-Fisher population [Wright, 1931], and we employ the notation $\theta = \theta_{WF} = \langle N_e \rangle$. For more complex scenarios, such as an exponentially expanding population, this parameterization might include the sizes of the ancestral and current populations, N_a and N_c, respectively, and the duration of the exponential expansion G. In such a case, we write $\theta = \theta_{EXP} = \langle N_a, N_c, G \rangle$. In the remainder of this work, we refer to the effective population size in a coalescent model simply as population size. For practical purposes, we focus on closed intervals $R = [u, v]$ of possible values for L and derive a closed-form expression for $p_R(l|\theta) = \int_u^v p(l|\theta)dl$.

We denote time in generations before the present throughout. The time g_{mrca} of the individuals' MRCA at the considered locus is generally unknown. We therefore marginalize it as

$$\int_u^v p(l|\theta)dl = \int_u^v \sum_{g=1}^{\infty} p(l, g_{mrca} = g|\theta)dl. \tag{3.1}$$

When the time to the MRCA is known, the length of the resulting shared segment is only dependent on the number of generations separating the two individuals (i.e., $l \perp\!\!\!\perp \theta|g_{mrca}$). Manipulating this expression, we therefore obtain

$$\int_u^v p(l|\theta)dl = \sum_{g=1}^{\infty} p(g_{mrca} = g|\theta) \int_u^v p(l|g_{mrca} = g)dl. \tag{3.2}$$

The distribution of the distance to the first recombination event encountered as we move either upstream or downstream of a chosen genomic site is exponentially distributed (it has a mean of $2g/100 = g/50$ centiMorgans, or $2g$ Morgans) because this is a haplotype shared by two individuals separated by $2g$ generations. The total length of the shared segment is therefore distributed as the sum of two independent exponential random variables parameterized by their mean of $2g$ Morgans, resulting in an Erlang-2 distribution with the same parameter, which has the form $Erl_2(l; 2t) = l(2t)^2 e^{-2tl}$. We therefore have

$$\int_u^v p(l|\theta)dl = \int_0^\infty \left[p(t_{mcra} = t|\theta_{WF}) \int_u^v Erl_2(l; 2t)dl \right] dt, \tag{3.3}$$

where we also standardly switch to a continuous time axis [Hudson, 1983] by replacing the discrete g_{mrca} with a continuous t_{mrca}, still measured in generations. Note that we are not measuring time in units of N_e generations as it is often done in the coalescent literature [Griffiths, 1991]. To complete the above formulation, we substitute the distribution of the time to MRCA for a specific demographic setting θ. In the coalescent framework, for the simple case of a population of constant size N_e and non-overlapping generations, the probability of finding a common ancestor at $g_{mrca} = g$ is geometric with parameter $p(g_{mrca} = g|\theta) = N_e^{-1}$ (or exponential at the continuous limit). Substituting this expression into Equation 3.3, we obtain the desired relationship between sharing of IBD haplotypes and population size:

$$\begin{aligned} p_R(l|\theta_{WF}) &= \int_0^\infty \left[\frac{e^{-t/N_e}}{N_e} \int_u^v Erl_2(l; 2t)dl \right] dt \\ &= \frac{4N_e^2(v-u)(4N_e uv + u + v)}{(2N_e u + 1)^2(2N_e v + 1)^2} \end{aligned} \tag{3.4}$$

3.1.2 Varying population size

When more complex population dynamics are considered, the probability of coalescence cannot be modeled through a simple geometric distribution. In general, for a population with demographic history θ, we can define a function $N(g, \theta)$ to express the population size at generation g. We can then express the chance of coalescence as

$$p(g_{mrca} = g|\theta) = \frac{1}{N(g,\theta)} \prod_{j=1}^{g-1} \left(1 - \frac{1}{N(j,\theta)}\right). \tag{3.5}$$

Equation 3.5 is very general and might lead to more complex instantiations for Equation 3.3. However, we consider a special and useful case in which the population history converges to $N_a = \lim_{g \to +\infty} N(g,\theta)$. By definition, there exists a finite time G before which $N(g,\theta) = N_a, \forall\{g > G\}$. In practice, we consider G to be the time before the period in history we aim to describe in detail, and we also note that demographic events preceding a sufficiently ancient generation G are unlikely to affect the probability of sharing IBD haplotypes longer than a chosen threshold. We observe that for any such converging history θ, we can always obtain a closed-form expression regardless of the specific form of $N(g,\theta)$ for $g \leq G$. For a population size of $N(g,\theta)$, such that $N(g,\theta) = N_a$ for all $g > G$, Equation 3.5 can in fact be rewritten as

$$\int_u^v p(l|\theta) dl = \phi_1(l,\theta,u,v,1\ldots G) + \phi_2(l,\theta,u,v,G+1\ldots\infty), \tag{3.6}$$

where

$$\phi_1(l,\theta,u,v,1\ldots G) = \sum_{g=1}^{G} \left[\prod_{j=1}^{g-1} \left(1 - \frac{1}{N(j,\theta)}\right) \frac{1}{N(g,\theta)} \int_u^v Erl_2\,(l;2t)\, dl \right], \tag{3.7}$$

and

$$\begin{aligned} \phi_2(l,\theta,u,v,G+1\ldots\infty) = {} & \frac{1}{N_a} \prod_{j=1}^{G} \left(1 - \frac{1}{N(j,\theta)}\right) \\ & \times \sum_{g=G+1}^{\infty} \left(1 - \frac{1}{N_a}\right)^{g-G-1} \int_u^v Erl_2(l;2t) dl. \end{aligned} \tag{3.8}$$

ϕ_1 adds up to a finite number of summands, and continuous time allows a closed-form expression for the infinite summation in ϕ_2. Using the coalescent distribution $e^{-t/N_e}/N_e$, we can integrate this probability between arbitrary time periods, and for segments longer than a threshold u

$$\int_{t_1}^{t_2} \frac{e^{-t/N_e}}{N_e} \int_u^{\infty} l(2t)^2 e^{-2tl} dl\ dt = \\ - \frac{e^{-t\left(\frac{1}{N_e}+2u\right)}\left[2tu(2N_e u+1)+4N_e u+1\right]}{(2N_e u+1)^2}\Bigg|_{t_1}^{t_2}. \tag{3.9}$$

Integrating between generation G and infinity using the ancestral population size N_a, and considering segments between u and v Morgans, this expression becomes

$$\int_{G}^{\infty} \frac{1}{N_a} e^{-t/N_a} \int_u^{v} l(2t)^2 e^{-2tl} dl\ dt = \\ \frac{e^{-G\left(\frac{1}{N_a}+2u\right)}\left[2Gu(2N_a u+1)+4N_a u+1\right]}{(2N_a u+1)^2} \\ - \frac{e^{-G\left(\frac{1}{N_a}+2v\right)}\left[2Gv(2N_a v+1)+4N_a v+1\right]}{(2N_a v+1)^2} \tag{3.10}$$

The function $N(g,\theta)$ can thus be arbitrarily defined to describe different demographic scenarios. Consider, for instance, the case of an ancestral population of size N_a: it exponentially expands during G generations to reach the current size N_c, parameterized by $\theta_{EXP} = \langle N_a, N_c, G\rangle$ as discussed above. The population size can be modeled (under the assumption of continuous time) as

$$N(t, \theta_{EXP}\langle N_a, N_c, G\rangle) = \begin{cases} N_c e^{-rt} & if\ t \leq G, \\ N_a & \text{otherwise.} \end{cases} \tag{3.11}$$

where $r = [\log(N_c) - \log(N_a)]/G$ is the population expansion rate. Note that an integration similar to that of Equation 3.9 may be done for a population that is exponentially growing/shrinking in a specific time range.

3.1.3 Sharing distribution

In the following section, we present explicit expressions for the case of Wright-Fisher populations (i.e., $\theta = \langle N_e\rangle$). Note, however, that these results are general, and analogous calculations can be performed for other demographic models.

Consider a specific site ς and a length range $R = [u, v]$. We are interested in IBD segments whose length lies within that interval, spanning the site ς. We consider the event of such a segment being shared between a randomly chosen pair of individuals from a studied population, and we define an indicator random variable for such an event as

$$\mathrm{I}(\varsigma, R = [u, v]) = \begin{cases} 1 & \text{if } \varsigma \text{ is traversed by a segment of length } u \leq l \leq v, \\ 0 & \text{otherwise.} \end{cases} \tag{3.12}$$

where we omit the dependence on the demographic model θ to simplify the notation. We now use these indicator variables to derive the expected fraction of genome spanned by IBD segments whose length is in this interval. Consider a dense set of sites Γ along the genome. Assume all sites are at equal genetic distance from adjacent sites. We have that

$$\begin{aligned} \mathrm{E}_R[f|\theta] = \mathrm{E}\left[\frac{1}{|\Gamma|}\sum_{\varsigma\in\Gamma}\mathrm{I}(\varsigma, R)\right] &= \frac{1}{|\Gamma|}\sum_{\varsigma\in\Gamma}\mathrm{E}[\mathrm{I}(\varsigma, R)] \\ &= \frac{1}{|\Gamma|}\sum_{\varsigma\in\Gamma}\int_u^v p(l|\theta)dl \\ &= \int_u^v p(l|\theta)dl. \end{aligned} \tag{3.13}$$

For given values of the demographic parameters θ, this predicts the fraction f of the genome shared through segments of length within specific intervals. To obtain the proportion of segments of a given length l, we divide $p(l|\theta)$ by l and multiply by a normalizing constant:

$$p(s = l|\theta) = \frac{p(l|\theta)}{l} \times \frac{1}{\int_0^\infty p(l|\theta)/l \; dl} = \frac{4N_e}{(2lN_e + 1)^3}. \tag{3.14}$$

The probability of finding a segment within the length range $R = [u, v]$ is thus

$$p(s \in R|\theta) = \int_u^v p(s = l|\theta)dl = \frac{1}{(2N_e u + 1)^2} - \frac{1}{(2N_e v + 1)^2}. \tag{3.15}$$

Equations 3.14 and 3.15 allow computing the length distribution of a segment in the range R,

$$p_R(s = l|\theta) = \begin{cases} \frac{p(s=l|\theta)}{p(s\in R|\theta)} & \text{if } s \in R, \\ 0 & \text{otherwise.} \end{cases} \tag{3.16}$$

and the expected length of such a segment,

$$\mathrm{E}_R[s|\theta] = \frac{\int_u^v l \times p(s = l\theta)dl}{p(s \in R|\theta)} = \frac{u + v + 4N_e uv}{2\left[1 + N_e(u + v)\right]}. \tag{3.17}$$

We note that for a typical pair of sharing individuals, the number and length of IBD segments are approximately independent [Huff *et al.*, 2011]. This allows us to express the expected genome-wide sharing between two individuals as the product of the expected number of IBD segments, λ_R, and the expected length of a shared segment in the considered length range, $\mathrm{E}_R[s|\theta]$. For a genome of size γ cM, $\gamma \times \mathrm{E}_R[f|\theta] \approx \mathrm{E}_R[s|\theta] \times \lambda_R$. We can thus compute the expected number of segments found in the considered length range as

$$\lambda_R \approx \gamma \times \frac{\mathrm{E}_R[f|\theta]}{\mathrm{E}_R[s|\theta]} = 2\gamma N_e \left[\frac{1}{(2N_e u + 1)^2} - \frac{1}{(2N_e v + 1)^2}\right]. \tag{3.18}$$

We model the number of shared segments as a Poisson random variable, $p_R(s = n|\theta) \approx Poiss(n, \lambda_R)$; thus, the standard deviation for the segment distribution is $\sigma_R[s|\theta] = \sqrt{\lambda_R}$. If the considered length range is not too wide, the variance of the segment lengths can be neglected, and we can obtain a simple approximation for the standard deviation of the fraction of genome shared through segments in the length range R by scaling $\sigma_R[s|\theta]$ by the expected length of a segment and by dividing it by the genome size:

$$\begin{aligned}\sigma_R[f|\theta] &\approx \frac{\mathrm{E}_R[s|\theta]\sqrt{\lambda_R}}{\gamma} \\ &= \sqrt{\frac{\mathrm{E}_R[f|\theta]\,\mathrm{E}_R[s|\theta]}{\gamma}} \\ &= \frac{2N_e(4N_e uv + u + v)}{(2N_e u + 1)(2N_e v + 1)}\sqrt{\frac{v - u}{\gamma(2N_e u + 2N_e v + 2)}}\end{aligned} \tag{3.19}$$

Finally, the obtained quantities can be used for expressing the full distribution of the portion τ of the genome shared through segments of a desired length again under the assumption of independence between number and length of shared segments. Define l_n to be the sum of n segments of length in the range R:

$$p_R(l_n = x|\theta) = \begin{cases} \delta(x) & \text{if } n = 0, \\ \mathrm{conv}[p_R(s = l|\theta), n] & \text{otherwise.} \end{cases} \tag{3.20}$$

where $\delta(\cdot)$ is the Dirac delta function and $\text{conv}[p_R(s = l|\theta), n]$ is the n-th convolution of $p_R(s = l|\theta)$ (e.g., $\text{conv}[p_R(s = l|\theta), 3] = p_R(s = l|\theta) * p_R(s = l|\theta) * p_R(s = l|\theta)$). The probability of sharing a total of x cM through segments of the desired length is then

$$\begin{aligned} p_R(\tau = x|\theta) &= \sum_{n=0}^{\infty} p_R(s = n, l_n = x|\theta) \\ &= \sum_{n|p_R(s=n|\theta)\neq 0} [p_R(s = n|\theta) p_R(l_n = x|\theta)]. \end{aligned} \tag{3.21}$$

Note that although we have considered the general length range $R = [u, v]$, the interval $R = [u, \infty)$ represents a particular and useful case in which all segments longer than a detectable threshold u are considered. When $v \longrightarrow \infty$, the length distribution simplifies to

$$p_R(s = l|\theta) = \frac{4N_e(2N_e u + 1)^2}{(2lN_e + 1)^3}, \quad \text{for } l \in R, \tag{3.22}$$

and the expected length becomes

$$E[s|\theta] = \frac{1}{2N_e} + 2u. \tag{3.23}$$

Note this becomes $1/(2N_e)$ for $u \longrightarrow 0$. The expected number of segments is therefore

$$\lambda_R \approx \gamma \times \frac{2\gamma N_e}{(2N_e u + 1)^2}. \tag{3.24}$$

When $u \longrightarrow 0$ the expected number of segments is $2\gamma N_e$, recovering a previous result [Hudson and Kaplan, 1985] (the number of segments delimited by recombination events should be $1 + 2\gamma N_e$, but it is reasonable to approximate this as $2\gamma N_e$). Additional quantities can be computed from the full sharing distribution of Equation 3.21, including a more precise expression for the variance of Equation 3.19. Other calculations for higher moments of IBD sharing in the Wright Fisher model can also be found in [Carmi *et al.*, 2013].

3.1.4 Inference

In the case of Wright-Fisher populations, we can obtain an estimate of the population size N_e by comparing the sharing observed in a specific length range to Equation 3.4 and by solving for N_e. The observed sharing in the length range $R = [u, v]$ can be computed from the analyzed data as

$$\hat{p}_R = \frac{\sum_{i|u \leq l_i \leq v} l_i}{\gamma \binom{n}{2}}, \tag{3.25}$$

where l is the length of a detected IBD segment and n represents the number of haploid individuals (see above for discussion of the diploid case). A closed-form solution for N_e can be computed for a given observed value of $\hat{p}_R$. In the particular case of $v \longrightarrow \infty$, where we consider all segments longer than a detectable threshold u, such a solution assumes a simpler form. Equation 3.4 becomes

$$\int_u^v p(l|\theta_{WF})dl = \frac{4N_e u + 1}{(2N_e u + 1)^2}, \tag{3.26}$$

and an estimate of N_e can be computed as

$$\hat{N}_e = \frac{1 - \hat{p}_R + \sqrt{1 - \hat{p}_R}}{2\hat{p}_R u}. \tag{3.27}$$

The number of IBD segments can also be used to derive a similar, improved estimator. Assuming all individual pairs are independent, and identically distributed, a Poisson likelihood for the segment counts can be obtained using the expectation of Equation 3.24. Setting the derivative w.r.t. N_e of the logarithm of such likelihood to zero, we obtain

$$\hat{N}_e = \frac{1 - \eta + \sqrt{1 - 2\eta}}{2\eta u}, \tag{3.28}$$

where $\eta = \frac{2\hat{s}_R u}{\gamma \binom{n}{2}}$ and $\hat{s}_R$ are the total number of segments longer than u cM observed for all $\binom{n}{2}$ pairs of genomes γ cM long.

For much of the analysis reported in this chapter, we minimized the squared deviation between the observed IBD sharing and the theoretical expectation (Equation 3.6) for a tested demographic model. The performance of this approach is comparable to that of estimation based on segment counts, although the latter may produce more accurate results. Both optimization methods were implemented in DoRIS, a publicly available software tool[2]. To compute a distance between observed and predicted sharing, we thus evaluate

[2] http://www.cs.columbia.edu/~pier/doris/

$$\delta_R = [\log(\hat{p}_R) - \log(\mathrm{E}_R[f|\theta])]^2 \tag{3.29}$$

and average this quantity across a collection of intervals $\Pi = \{R_j\}_{1 \leq j \leq |\Pi|}$:

$$\delta_\Pi = \sqrt{\frac{1}{|\Pi|}\sum_{j=1}^{|\Pi|} \delta_{R_j}}. \tag{3.30}$$

The transformation to log space in Equation 3.29 has the effect of making the error contributions along the dynamic range of length intervals more uniform than in linear space. Grid-search minimization of Equation 3.30 can therefore be employed for exploring a large portion of the parameter space. Upon convergence to a grid point of least deviation from the theoretical expectation, a full likelihood-based approach can be used for retrieving the most likely values for the demographic-model parameters in a smaller portion of the parameter space and can thus allow substantial computational savings. An alternative to the minimization of the squared deviation is maximizing a composite likelihood based on Poisson counts of the observed segments. We have observed comparable performance for either approach.

3.1.5 Evaluation on synthetic data

To evaluate the accuracy of the proposed model and of the inference procedure, we simulated a large number of synthetic populations by using the GENOME coalescent simulator [Liang *et al.*, 2007]. We extracted ground-truth information on shared segments to eliminate the noise introduced by methods for IBD discovery. To this extent, the coalescent simulator was modified to output shared segments descending from the same ancestor as observed in the synthetic genealogy (according to definition (b) for IBD segments in Section1.1.4.1). For all the simulations, we generated a total of 500 diploid samples for a single chromosome made of $27,800$ non-recombining blocks with an inter-block recombination rate of 10^4, mimicking the genetic length of chromosome 1 ($\sim$278 cM). We verified that the use of non-recombining blocks of 0.01 cM did not introduce significant biases in our analysis (see Supplementary Figure 1). We simulated 900 synthetic populations that underwent exponential contraction and expansion (see Supplementary Table 1 for the range of demographic parameters). We applied a gradient-driven local-minimization procedure to retrieve the parameter values that minimize Equation 3.30. In order to avoid local minima, we initially performed a grid search in a predefined box volume of the parameter space (see Supplementary Table 1 for the parameters list). We then refined the least-squares solution by using a gradient-based optimization from the best point on the grid.

The accuracy of our inference procedure depends on the length of the analyzed genomic region and on the number of samples for which IBD segments are observed. In particular, it follows from Equation 3.25 that upon fixing $\hat{p}_R$ and $\sum_{i|u\leq l_i\leq v} l_i$, the result is unchanged for several values of γ and n. In terms of accuracy of the proposed evaluation, an equivalent configuration would have been the use of $\sim$140 diploid individuals for the entire genetic length of the autosomal genome ($\sim 3,500$ cM for the HapMap 3 genetic map; see Supplementary Figure 2). The choice of length intervals $R_j = [u_j, v_j]$ also affects the inference results: segments of length between 1 and 2 cM, for instance, might have originated from a wide span of generations in the past, whereas segments of length 10-11 cM tend to have a more deterministic (and more recent) origin. Frequency bins of different sizes can be used for focusing on specific time periods. For all the analyses reported in this paper, we adopted a combination of bins of uniform length and bins of length intervals corresponding to specific percentiles of the Erlang-2 distribution. In particular, we used length values between the 21.4th and the 31.4th percentiles of the Erlang-2 distributions with parameter $\lambda = \frac{k}{50}$ (the maximum likelihood estimate occurs at the 26.4th percentile) for several consecutive integral values of k (i.e., $k = 2, 3, \ldots 43$).

3.1.6 Real data analysis

We applied the proposed inference procedure to genotype samples of 500 AJ individuals from Jerusalem (Israel) and 143 MKK individuals from Kinyawa (Kenya), already analyzed in Chapter 2. The AJ individuals were typed on the Illumina 1M platform and are self-reported unrelated individuals. After quality control, a total of $745,811$ autosomal SNPs were used for the analysis. The MKK samples comprise 56 unrelated trio-phased individuals and 87 unrelated individuals from the HapMap 3 data set previously introduced. As a result of the availability of haplotype phase information, we focused our analysis on the 56 trio-phased samples and used $1,387,466$ markers for the analysis.

The AJ samples were phased with the Beagle software package [Browning and Browning, 2007], whereas trio-phased MKK individuals were downloaded from the HapMap website[3]. IBD sharing was estimated with the GERMLINE software package [Gusev *et al.*, 2009]. We tweaked the parameters of the GERMLINE algorithm to improve the quality of IBD detection for the specific data set by using the following procedure. Using GERMLINE's default *haplotype extension* parameters, we extracted IBD segments from the real data and then used the analytical inference procedure to retrieve demographic parameters. We simulated a synthetic population by using the inferred demography and extracted ground-truth IBD segments. We ran GERMLINE on the synthetic genotypes several times and changed the *err_hom*, *err_het*, *bits* to find a set of parameters that minimized the deviation of the genotype-inferred IBD sharing density from that obtained from ground-truth data. We then used these parameters to extract IBD segments from the real data again and iterated the procedure until convergence. The GERMLINE parameters to which we converged were *-min_m 1 -err_hom 0 -err_het 2 -bits 25 -h_extend* for the Beagle-phased AJ data and *-min_m 1 -err_hom 2 -err_het 2 -bits 60 -h_extend* for the trio-phased MKK data.

3.1.6.1 Demographic Model Selection in the AJ Population

We tested increasingly flexible models to infer the demographic history of the AJ population. In order to control for potential over fitting, we evaluated the parameters obtained for different models by using a likelihood approach. To this extent, after optimizing the model parameters by using the least-squares approach, we used rejection sampling to retrieve parameters corresponding to a local maximum likelihood for each model. We then used the Akaike information criterion (AIC, [Akaike, 1974]) to compare models while controlling for their different degrees of freedom (see the algorithm reported in Supplementary Table 2).

Three models were used for the inference in the AJ population (see Figure 3.1 and an

[3] http://hapmap.ncbi.nlm.nih.gov

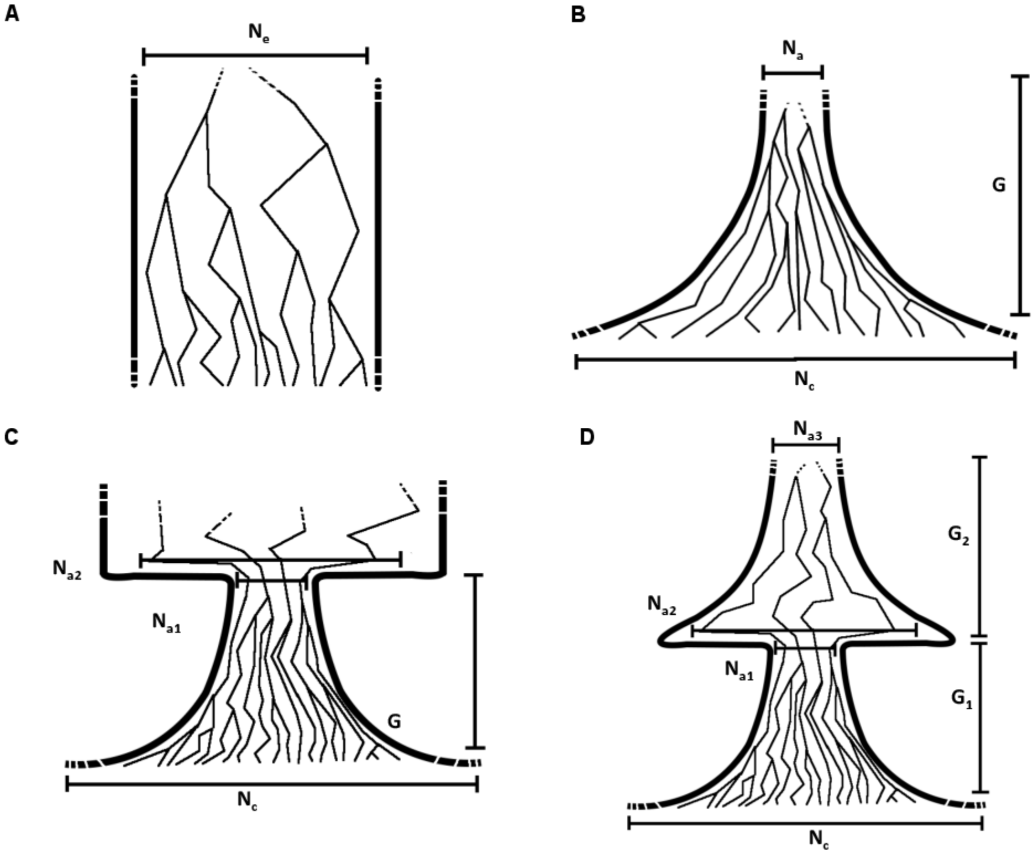


Figure 3.1: Demographic Models. (A) Population of constant size. (B) Exponential expansion (contraction for $N_a > N_c$). (C) A founder event followed by exponential expansion. (D) Two subsequent exponential expansions separated by a founder event.

additional description in the Results): (1) a model of exponential expansion ($\mathcal{M}_E$), (2) a model including a founder event followed by exponential expansion ($\mathcal{M}_{FE}$), and (3) a model of two exponential-expansion periods separated by a founder event ($\mathcal{M}_{EFE}$). The $\mathcal{M}_E$ model did not provide enough flexibility to fit the IBD-sharing summary extracted for the AJ population, resulting in a poor fit (particularly for shorter segments) and unrealistically large values for the recent population size. We therefore excluded this model from further analysis. For models $\mathcal{M}_{FE}$ and $\mathcal{M}_{EFE}$, we used the following rejection-sampling approach to maximize the model likelihood around the least-squares solution obtained in the previous step. (1) For each model, for each model parameter, we generated a list of neighboring points by allowing each parameter to vary by ±3% of its current value. (2) For each point on such a local grid, we sampled several random data sets of sharing individuals by using the corresponding demographic parameters (details in Supplementary Table 3). We created each data set by sampling random sharing values for independent individual pairs from the distribution of Equation 3.21. (3) For each analyzed set of parameter values, we computed a likelihood as the fraction of data points for which the deviation between AJ and sampled sharing was smaller than a tolerance threshold δ ($\delta \approx 0.089$ for $\mathcal{M}_{FE}$ and $\delta \approx 0.037$ for $\mathcal{M}_{EFE}$). (4) We updated the current point to the most likely point in the analyzed neighborhood, if any, and iterated steps 1-3 until no point with a higher likelihood was found. (5) We applied the AIC to compare models.

For both models, only one iteration of the above local maximization was required. The most likely parameter values in the grid matched those obtained with the least-squares approach, except for the current population size, which increased by 3% for model $\mathcal{M}_{FE}$ and decreased by 3% for model $\mathcal{M}_{EFE}$. When comparing the two models, we used a tolerance threshold of $\delta \approx 0.037$ and obtained an AIC value of 19.21 for the $\mathcal{M}_{EFE}$ model, which allows five parameters to vary (such δ results in a likelihood of 0.01 for the $\mathcal{M}_{EFE}$ model). Using the same acceptance threshold, we thus required a log likelihood of at least -5.6 (a likelihood of $\sim 3.7 \times 10^3$) for model $\mathcal{M}_{FE}$, which has four parameters, to be selected. None of the 105 sampled points were accepted with such a threshold, leading us to choose the $\mathcal{M}_{EFE}$ model. The likelihoods of additional parameter values estimated for the $\mathcal{M}_{EFE}$

model with the use of a wider grid are reported in Supplementary Table 4.

Note that when sampling from Equation 3.21, we assumed independence of the analyzed sharing length intervals R_i and of the pairs within a data set, potentially underestimating the variance of randomly sampled summaries of IBD. To account for the presence of small correlations, we thus performed full coalescent simulations according to the most likely set of parameters of each model by only sampling a synthetic chromosome 1 for 500 diploid individuals. We repeated the rejection-based comparison by using 104 such points for each model and obtained an equivalent result.

3.1.6.2 Accounting for Phase Errors

The inference procedure described in the previous sections assumes that high-quality IBD information is available. When real data sets are analyzed, several sources of noise, such as computational phasing errors, might distort summary statistics of haplotype sharing. In the absence of reliable probabilistic measures for the quality of shared segments, modeling this potential bias is complicated. To account for this additional noise, we refined the inferred AJ demographic model by using simulations that mimic SNP ascertainment, inaccurate phasing, and IBD discovery in the analyzed data sets. We expected the distortion of IBD summary statistics in the AJ data set to not be substantial (Supplementary Figure 3). The preliminary inference based on the assumption of high-quality IBD information therefore provides an efficient means for exploring large portions of the parameter space and for performing model comparison. This can be followed by such simulation-based refinement, which requires considerable computation.

After finding the most likely parameters and selecting model $\mathcal{M}_{EFE}$ for the AJ data as previously described, we refined the obtained solution by using a local-search approach. We iteratively varied one demographic parameter at a time and kept a tested value if it resulted in a decreased deviation from the AJ data summary. Note that in order to account for the stochastic variation observed across multiple independent simulations of the same demographic history, we would need to generate several synthetic data sets for each tested set of demographic parameters. However, we did not repeat such simulations multiple times as a result of computational constraints.

For all coalescent simulations in real-data inference, we used the GENOME software package. The simulated chromosomes have the same genetic length as their real-data equivalent and a mutation rate of 1.1×10^8 per site per generation [Roach *et al.*, 2010]. To reduce the computational burden, we used non-recombining block units of 10 kb for MKK simulations and 20 kb units for AJ simulations, resulting in an IBD length resolution of 0.01 and 0.02 cM, respectively. Synthetic markers were randomly ascertained to match the same density of the real data. We matched the spectrum of the real data sets by randomly selecting the same proportion of variants for each frequency bin and used a bin size of 2%. No missing genotypes were allowed in simulated data because occasional missing genotypes in the real data were imputed during Beagle phasing or excluded from the analysis if not reliably imputed. All simulations were carried out for the entire autosomal genome.

3.2 Results of evaluation and real data analysis

3.2.1 Simulated data

The described methods were implemented in *DoRIS*, a freely available software tool[4]. We tested the accuracy of the proposed model through extensive simulation of synthetic populations with known demographic history. For each simulated population, we analyzed a region of length equivalent to chromosome 1 for 500 diploid samples (see Section 3.1). All the derived theoretical quantities were found in good agreement with the values obtained from simulation (see Supplementary Figure 4 for an evaluation summary and Figure 3.2 for examples of total haplotype-sharing distributions). We noted that for populations of constant size, as expected, a smaller population size causes a larger fraction of the genome to be shared through IBD segments for the average pair in the population (Figure 3.3). Furthermore, the frequency of segments at different length intervals is informative of population size at different time scales. Consider the case of an exponential expansion (Figure 3.1.A) with the following parameterization: N_a is the size of the ancestral population when

[4] `http://www.cs.columbia.edu/~pier/doris/`

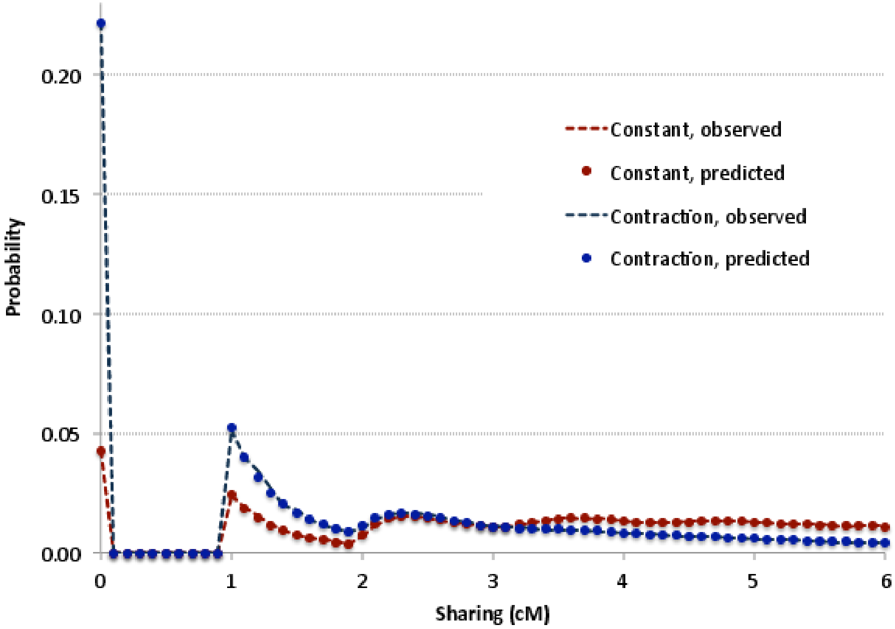


Figure 3.2: Analytical (dots) and empirical (dashed lines) distribution for total IBD (Equation 3.21) for a constant population of $2,000$ diploid individuals (red, $R = [1, 4]$) and an exponentially contracting population ($A = 50,000$, $C = 500$, $G = 20$, $R = [1, \infty)$, in blue).

exponential expansion began, N_c denotes the population size at the current generation, and G represents the number of generations during which the exponential expansion took place. A small ancestral population size N_a causes a higher rate of remote coalescent events and a consequently larger fraction of the genome to be spanned by short segments of IBD. Similarly, a small value of N_c increases the chance of coalescence in the more recent generations, causing a larger fraction of the genome to be spanned by long segments. For fixed Na and N_c, variations of the duration of expansion G affect the expansion rate and have a noticeable effect on the slope of the sharing distribution, i.e., the genome fraction spanned by mid-length segments.

We used the relationship of Equation 4 to infer the size of a Wright-Fisher population by using a realistic chromosome 1 simulated for several populations, each with its own constant size N_e ranging from 500-40,000 individuals. In the analysis of IBD information for 500 diploid samples in each such synthetic population, the predicted value was highly correlated with the true size of the synthetic populations ($r = 0.9994$; Figure 3.4). Across all tested values of N_e, the ratio between true and estimated population size had a median of 1.00 and a 95% confidence interval (CI) of 0.97-1.03.

3.2.2 IBD and heterozygosity in an expanding population

To outline IBD's particular sensitivity to recent demographic variation, we examined the effects of variable population size on demographic inference conducted either through the proposed approach based on IBD haplotypes or through a classical approach based on heterozygosity. We focused on the scenario in which a population of 3,000 ancestral individuals suddenly expands to a size of 25,000 individuals G generations before the present (Figure 3.4.C). We varied G from 10-400 generations and simulated the ascertainment of IBD haplotypes by extracting information on shared haplotypes along a realistic chromosome 1 for 500 diploid samples. For both IBD-based and heterozygosity-based reconstructions, we assumed and inferred a constant population size N_e. We used the relationship of Equation 3.4 for the IBD model and Watterson's estimator of Equation 1.16 for the heterozygosity-based approach (the heterozygosity θ was estimated from the synthetic sequences, and μ matched

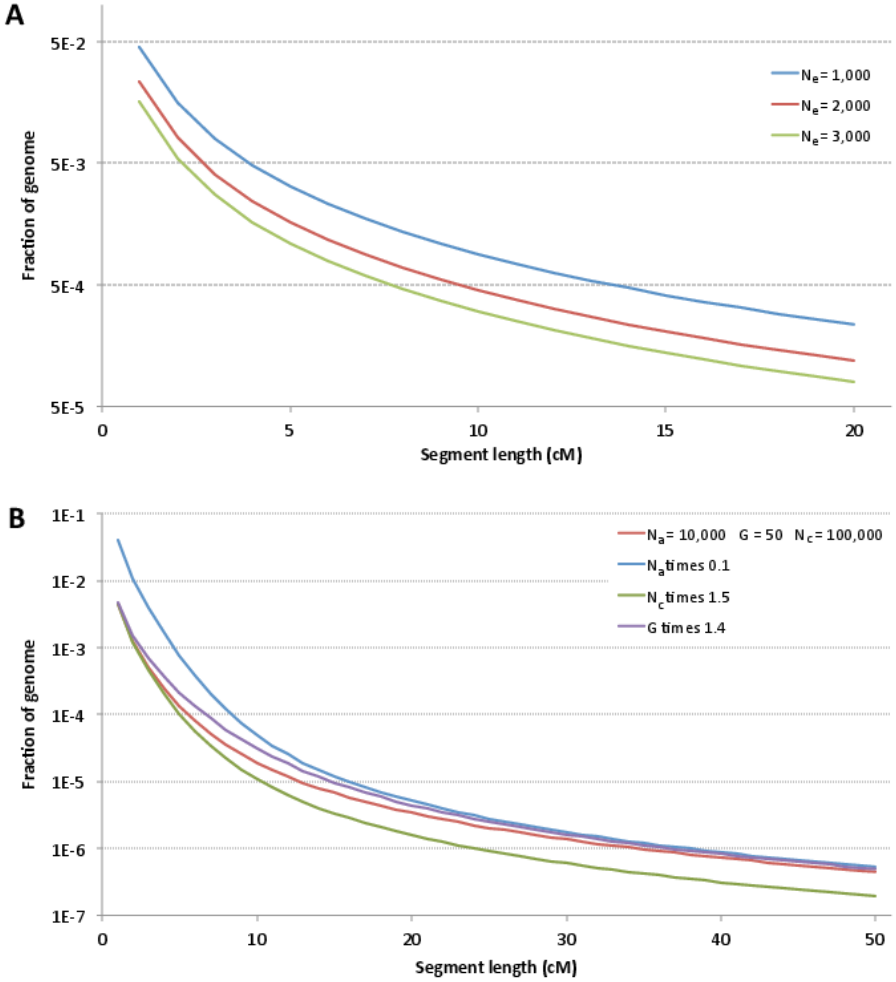


Figure 3.3: Effects of demographic parameters on IBD length distributions. Wright-Fisher models (A) and exponential population expansion (B).

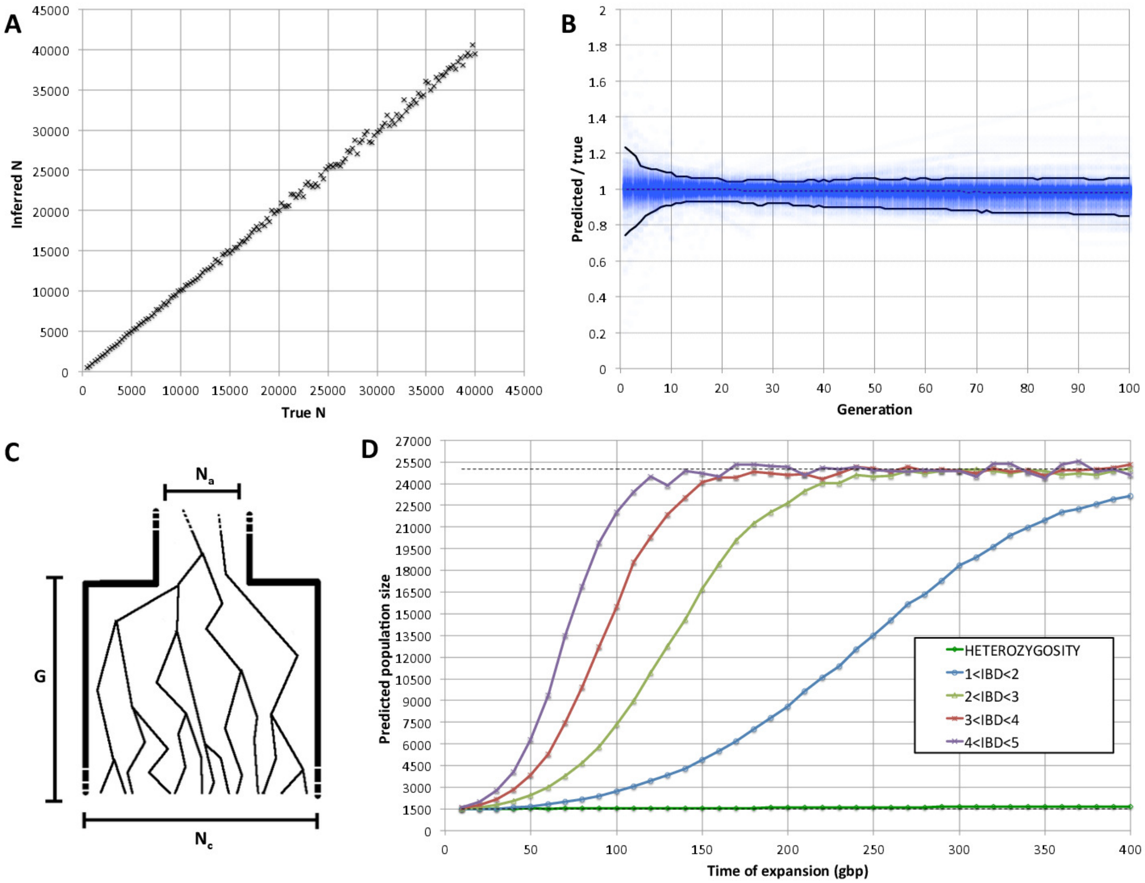


Figure 3.4: Performance for constant-size populations (A), expanding and contracting populations (B), and a suddenly expanding population (C) studied with a constant-size model (D).

the simulated mutation rate). An estimate of N_e was obtained for each data set across all simulated times of expansion (Figure 3.4.D). As expected, the obtained estimate of N_e tended to lie in the range between the ancestral and the current size of the population. Long, recently originated segments provide a better prediction of the current population size, especially for remote expansions. In contrast, the high frequency of shorter segments of more remote origins biases the inference toward a smaller population size when these segments are taken into account. For example, the effects of a small ancestral population size can be observed on segments between 4 and 5 cM in length only for expansions that occurred fewer than 120 generations ago; in contrast, when segments between 1 and 2 cM in length are analyzed, traces of a smaller ancestral population are still notable, even for expansions that occurred as far back as 400 generations ago. When comparing these results to population-size estimates obtained with heterozygosity from full synthetic genomic sequence, we observed the heterozygosity-based estimates of N_e to be strongly biased toward the small size of the ancestral population. Although they present less instability than do the IBD-based estimates, the inferred values approached the ancestral population size, even for expansions that occurred 400 generations before the present. This analysis outlines the unique sensitivity of long-range IBD sharing to recent demographic variation.

3.2.3 Evaluation of the inference in populations of varying size

We tested the accuracy of our inference procedure for the cases of either an exponential increase or decrease in population size (expansion or contraction, respectively). We simulated 450 synthetic populations that underwent an exponential expansion and 450 that underwent exponential contraction. We analyzed the IBD sharing of 500 diploid samples from each simulated population along a 278 cM chromosome. We evaluated the accuracy of the inferred demography by using the ratio between true and predicted sizes of each analyzed population (Figure 3.4.B) for all generations between 1 and 100. We found our inferred population size to be within 10% of the true value 95% of the time. The population size of recent generations was harder to infer because of the scarcity of long IBD segments in very large populations (this scarcity is due to a low chance of recent coalescent events).

Note that the reconstruction accuracy is influenced by sample size and length of the analyzed region (see Section 3.1). The rates of expansion and contraction also substantially affect the ability to recover the correct population size; faster expansion and contraction rates incur more noisy estimates (the testing reported in Figure 3.4.B included extreme and possibly unrealistically large rates of expansion and contraction). This was evident when we classified the synthetic populations as either strong or mild contraction or expansion events and separately assessed the inference accuracy for each of these classes (see Supplementary Figure 5).

3.2.4 Two periods of expansion in the Ashkenazi Jewish population

We analyzed the demographic history of the AJ population by applying our method to a real data set of 500 individuals (segment-length distributions in Figure 3.5). We initially tested several models by using the proposed procedure. After inferring the most likely parameters for the chosen model, we used simulations to refine the analytical solution and account for potential errors in IBD detection (see Supplementary Table 2 for an algorithmic summary of the analysis).

As a first step, we fitted a simple model of exponential growth (Figure 3.1.B). If only long ($\geq$ 5 cM) segments are considered, the parameters of this model can be optimized to provide a good match for the observed sharing. This supports the occurrence of an expansion event in the recent history of this population, as reported in our previous analysis using a simpler simulation-based approach [Gusev *et al.*, 2012]. However, exponential growth alone is unable to provide a good fit for the observed frequency of shorter segments, suggesting additional demographic dynamics during more ancient AJ history. The decay in the frequency of medium-length segments, between 2 and 5 cM, was weaker than that observed for longer ones, suggesting a founder event—a reduction of the ancestral population size and subsequent rapid expansion. Indeed, a refined model that allows such an event to predate exponential expansion (Figure 3.1.C) provides a good fit for the frequency of all segments of length $\geq$ 2 cM. We note that such a severe founder event was also reported in a previous analysis based on lower throughput data [Slatkin, 2004;

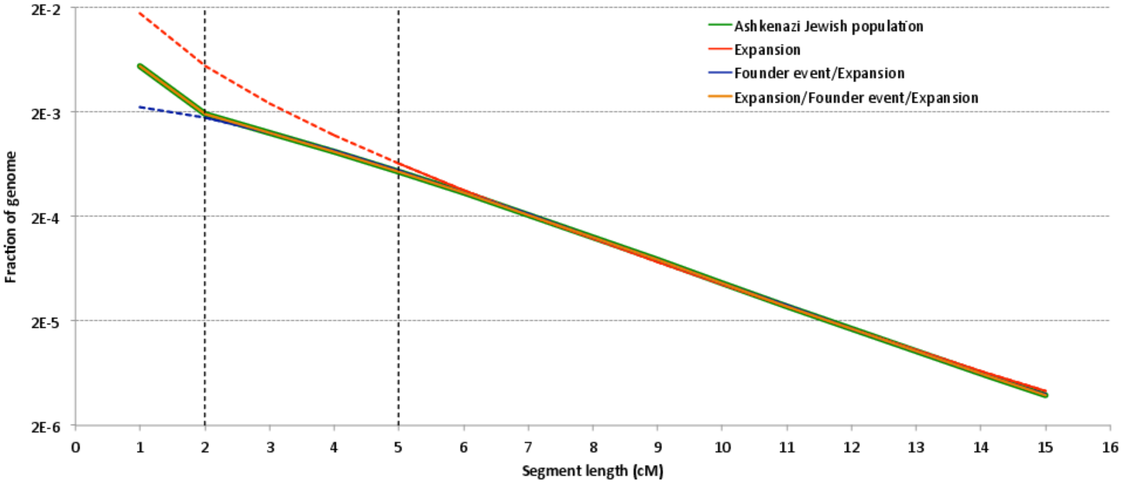


Figure 3.5: AJ inference. Observed distribution of haplotype sharing (green line); exponential expansion for only long (> 5cM) segments (red line, best fit: $N_c = 97,700,000, G = 26, N_a = 1,300$); founder event-expansion (purple line, best-fit: $N_c = 12,800,000; G = 35; N_{a1} = 230; N_{a2} = 70,600$); exponential expansion-founder event-exponential expansion (orange line, best-fit: $N_c = 42,000,000; G_1 = 33; N_{a1} = 230; N_{a2} = 37,800; N_{a3} = 1,800; G_2 = 167$).

Atzmon *et al.*, 2010] and is consistent with historical reports of this population [Finkelstein, 1960]. However, this model does not adequately explain why a further change in the slope of the sharing spectrum was observed for short segments between 1 and 2 cM of length. Such a steep increase in the frequency of short segments can again support the occurrence of an exponential growth preceding the observed founder event. We therefore optimized parameters for a model that allows two subsequent exponential-expansion periods separated by a founder event (Figure 3.1.D). We focused our analysis on generations $1-200$ (i.e., setting $G_1+G_2=200$ in Figure 3.1.D). The considered model allows N_{a3} founders to exponentially expand to a population of N_{a2} individuals during G_2 generations. After a founder event, N_{a1} individuals are randomly selected and exponentially expand to reach a current population of N_c individuals during the remaining G_1 generations. Using this model, we were able to obtain a good fit for the entire IBD frequency spectrum, corresponding to the parameter values $N_{a3}\sim 1,800$, $N_{a2}\sim 37,800$, $N_{a1}\sim 230$,and $G_1=33$ (therefore, $G_2=167$) and $N_c\sim 42,000,000$. Model comparison based on the AIC supports this model over simpler demographic scenarios. We note that the most recent expansion period was inferred to have a considerably high rate ($r\sim 0.37$, defined in Equation 3.11). More complex models (e.g., inferring the value of G_2 and allowing for a founder event predating the remote expansion) did not significantly improve on the reported demography.

When real data is analyzed, the quality of computational phasing and IBD detection might affect the reconstruction accuracy. Inaccuracies in the recovery of long-range IBD haplotypes are reflected in the inferred current size of the AJ population, which is extremely large. This is most likely due to long IBD segments being shortened to smaller segments because of switch errors during computational phasing, in addition to greater uncertainty associated with the inference of recent large population sizes (Figure 3.4.B and Supplementary Figure 5). We therefore refined inferred parameters to take into account such potential bias by using realistic coalescent simulations that also reproduce noise due to computational phasing and IBD discovery. We obtained an improved fit for a population composed of $\sim$2,300 ancestors 200 generations before the present; this population exponentially expanded to reach $\sim$45,000 individuals 34 generations ago. After a severe founder event,

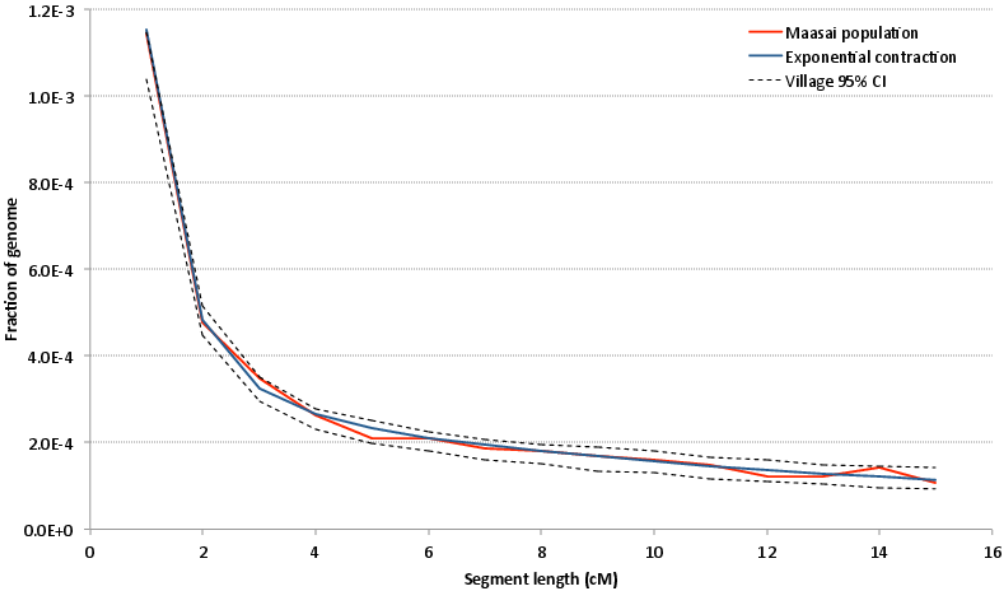


Figure 3.6: Demographic analysis of the Maasai. Observed distribution of haplotype sharing (red); single-population expansion model (blue); several small demes that interact through high migration rates (dashed CI obtained through random resampling of 200 synthetic data sets).

the population was reduced to ∼270 individuals, which then expanded rapidly during 33 generations (rate $r \sim 0.29$) and reached a modern population of ∼$4,300,000$ individuals.

3.2.5 IBD in the Maasai: the village model

We additionally investigated the demographic profile of 56 samples of self-reported unrelated MKK individuals from the HapMap 3 data set. We detected high levels of segmental sharing across individuals, consistent with recent analysis of hidden relatedness in this sample (see [Pemberton *et al.*, 2010; Gusev *et al.*, 2012], Chapter 2). Genome-wide IBD sharing was elevated among all individual pairs, suggesting high rates of recent common ancestry across the entire group rather than the presence of occasional cryptic relatives due to errors during sample collection (Supplementary Figure 6). Optimizing a model of exponential expansion and contraction (Figure 3.1.B), we obtained a good fit to the observed IBD frequency spectrum (Figure 3.6), suggesting that an ancestral population of ∼23,500 individuals decreased to ∼500 current individuals during the course of 23 generations ($r \sim -0.17$). We note that this result might not be driven by an actual gradual population contraction in the MKK individuals, but it most likely reflects the societal structure of this semi-nomadic population. Although little demographic evidence has been reported, the MKK population is in fact believed to have a slow but steady annual population growth [Coast, 2001]. We hypothesized that a high level of migration across small-sized MKK villages (Manyatta) provides a potential explanation for the observed IBD patterns in this population. In such a model, a small genetic pool for recent generations gradually becomes larger as a result of migration across villages as one moves back into the past. To validate the plausibility of this hypothesis, we simulated a demographic scenario in which multiple small villages interact through high migration rates. This setting is similar to Wright's island model [Wright, 1943], and we shall refer to it as the village model in this case (Supplementary Figure 7). We extracted IBD information for one of the simulated villages and attempted to infer its demographic history by using a single-population model of exponential expansion and contraction (Figure 3.1.B). Indeed, the single-population model provides a good fit for this synthetic sample, and the severity of the gradual contraction of the population was observed to be proportional to the simulated migration rate. We thus used the village model to analyze the MKK demography and relied on coalescent simulations to retrieve its parameters: migration rate, size, and number of villages that provide a good fit for the empirical distribution of IBD segments. We observed a compatible fit for this model, in which 44 villages of 485 individuals each intermix with a migration rate of 0.13 individuals per generation (Figure 3.6).

Note that, although our simulations involved several villages of constant size, adequate choices of migration rates would result in the signature of a drastic contraction even among expanding villages (and, therefore, overall expanding population). From a methodological point of view, we further note that LD might also provide information for inferring such a "village effect". However, although current strategies for IBD detection allow finding shared haplotypes in the presence of computational phasing errors, LD analysis over long genomic intervals is substantially affected by noisy phase information (Supplementary Figure 8).

Chapter 4

Reconstructing recent migration events

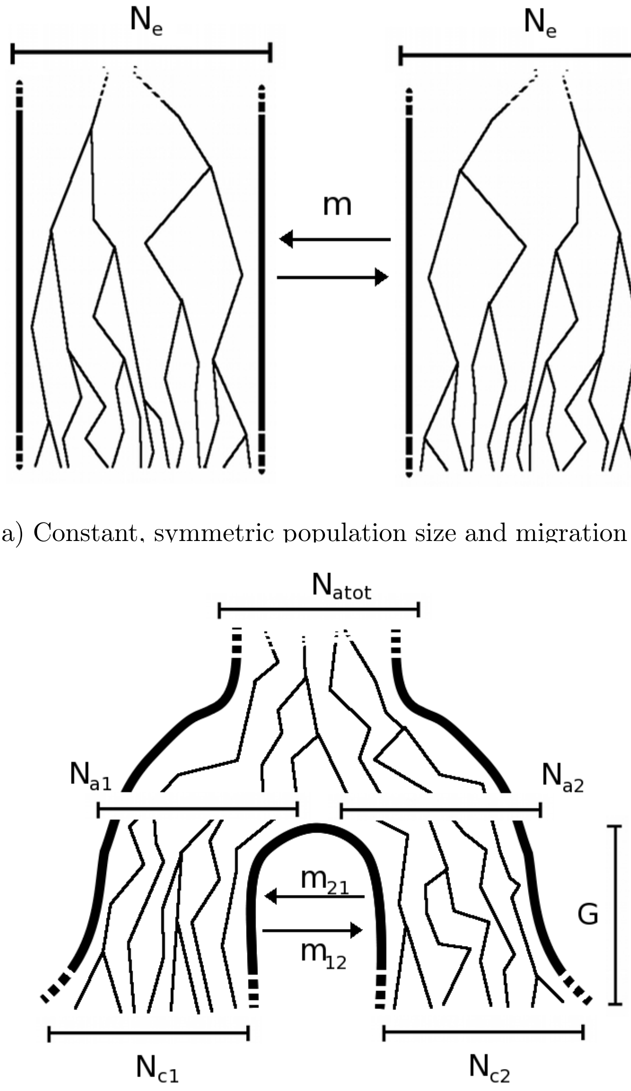
In Chapter 3 we introduced a model that allows expressing several relevant quantities of IBD sharing across purportedly unrelated individuals from a single population. We here extend this analysis to the case of individuals sampled from a number of different demes, of which we want to investigate recent demographic events such as population size fluctuation and migration. Details of this analysis can be found in [Palamara and Pe'er, 2013]. Note that in the remainder we measure genetic length in Morgans (M).

4.1 IBD distributions in the presence of migration

We begin discussing the case of multiple populations referring to a simple scenario, where two populations of constant size N_e exchange individuals at a fixed rate m per individual, per generation (see model in Figure 4.1a). We encode this migration rate using the matrix

$$\mathbf{Q} = \begin{pmatrix} -m & m \\ m & -m \end{pmatrix}$$

We consider two individuals, i and j, each sampled from either population. We trace the ancestors of these individuals at one genomic site, and encode their state (in terms of population their ancestors belong to), using a vector of dimensionality 2. If individual i is sampled



(a) Constant, symmetric population size and migration rate

(b) Population split followed by size changes and migration

Figure 4.1: Two demographic models that involve two populations and migration between them. In model 4.1a the populations have the same constant size N_e, and exchange individuals at the same rate m. In model 4.1b, a population of constant ancestral size N_{atot} splits G generations in the past, resulting in two populations whose sizes independently fluctuate from N_{a1} and N_{a2} individuals to N_{c1} and N_{c2} individuals during G generations. During this period, the populations interact with asymmetric migration rates m_{12} and m_{21}.

from population 1 and individual j from population 2, for example, the state at generation 0 is known and we write it as $\mathbf{v}_i(0) = (1, 0)$, $\mathbf{v}_j(0) = (0, 1)$. If both are sampled from population 1, $\mathbf{v}_i(0) = (1, 0)$, $\mathbf{v}_j(0) = (1, 0)$. After t generations (measured in continuous time), the probability that the ancestor of individual i at this genomic location belongs to either population is given by

$$\mathbf{v}_i(t) = (1, 0)e^{t\mathbf{Q}} = \left(\frac{e^{-2mt}}{2}(1 + e^{2mt}), \frac{e^{-2mt}}{2}(e^{2mt} - 1)\right) \tag{4.1}$$

if individual i was sampled from population 1, or, symmetrically

$$\mathbf{v}_i(t) = (0, 1)e^{t\mathbf{Q}} = \left(\frac{e^{-2mt}}{2}(e^{2mt} - 1), \frac{e^{-2mt}}{2}(1 + e^{2mt})\right) \tag{4.2}$$

if it was sampled from population 2. We are interested in expressing the probability that individuals i and j coalesce at time t. This requires both individuals to be in the same population, in which case coalescence happens at rate $1/N_e$. Formally $p(t|m, N_e) = \mathbf{v}_i(t)\mathbf{v}_j(t)^\intercal/N_e$, which in this setting becomes

$$p(t|m, N_e) \approx \frac{1 + e^{-4mt}}{2N_e} \tag{4.3}$$

if $\mathbf{v}_i(0) = \mathbf{v}_j(0)$, and

$$p(t|m, N_e) \approx \frac{1 - e^{-4mt}}{2N_e} \tag{4.4}$$

otherwise. Note that a Taylor approximation was made in equations 4.3 and 4.4. A more detailed derivation is reported in the Appendix of this chapter. To compute $\int_u^v p(l|\boldsymbol{\theta})\,\mathrm{d}l$, we plug the coalescence probability in Equation 3.2 (or its continuous version). Also, for simplicity we take $R = [u, \infty)$, obtaining

$$\int_u^\infty p(l|\boldsymbol{\theta})\,\mathrm{d}l = \frac{1}{2N_e u} + \frac{m + u}{2N_e(2m + u)^2} \tag{4.5}$$

if $\mathbf{v}_i(0) = \mathbf{v}_j(0)$, and

$$\int_u^\infty p(l|\boldsymbol{\theta})\,\mathrm{d}l = \frac{m(4m + 3u)}{2N_e u(2m + u)^2} \tag{4.6}$$

otherwise. Recall that $\int_u^v p(l|\boldsymbol{\theta})\,\mathrm{d}l = f_R$, which is the expected fraction of genome shared through segments of length between u and v by an individual pair. To infer $\hat{N}$ and $\hat{m}$, we therefore consider the observed average fraction of genome shared through IBD segments longer than a threshold u, for all pairs of individuals sampled from the same population or from different populations (which we call $\hat{f}_s$ and $\hat{f}_d$, respectively, now omitting the dependence on the length range). We then solve the system obtained by equating $\hat{f}_s$ and $\hat{f}_d$ to the quantities in (4.5) and (4.6), to obtain the estimators

$$\begin{aligned}\hat{N}_e &= \frac{1}{(\hat{f}_d + \hat{f}_s)u} \\ \hat{m} &= \frac{u\left(3\hat{f}_s - 5\hat{f}_d - \sqrt{2\hat{f}_d\hat{f}_s - 7\hat{f}_d^2 + 9\hat{f}_s^2}\right)}{8(\hat{f}_d - \hat{f}_s)}\end{aligned} \tag{4.7}$$

Note that, although computations are not explicitly reported here, a similar derivation may be obtained for analysis based on the counts of shared segments, as described in Chapter 3.

A simple generalization of the above scenario consists in allowing the two considered populations to differ in their effective population sizes, N_{e1} and N_{e2}. In this scenario it is still possible to obtain a closed form expression for $\int_u^v p(l|\boldsymbol{\theta})\,\mathrm{d}l$, and a closed form estimator for $\hat{N}_{e1}$, $\hat{N}_{e2}$, $\hat{m}$, which are reported in the Appendix.

4.1.1 The general case

Although the previously discussed case of constant population sizes and migration rates has a simple formulation and can be used to gain initial insight into the recent demography of a study cohort, such population dynamics are oversimplified and generally unrealistic. Luckily, given a few reasonable assumptions, population sizes and migration rates can be allowed to arbitrarily fluctuate in time, still permitting a closed form computation of $\int_u^v p(l|\boldsymbol{\theta})\,\mathrm{d}l$.

Consider two populations whose sizes at generation g are expressed as $N_1(g)$ and $N_2(g)$. The rate at which these two populations exchange individuals can be encoded in a discrete migration matrix

$$\mathbf{M}(g) = \begin{pmatrix} 1 - m_{12}(g) & m_{12}(g) \\ m_{21}(g) & 1 - m_{21}(g) \end{pmatrix} \tag{4.8}$$

where $m_{12}(g)$ represents the probability of an individual migrating from population 1 to population 2 at generation g (backwards in time). After g generations, the probability that the ancestor of individual i at a genomic location belongs to either population is given by the vector $\mathbf{v_i}(0)\prod_{k=0}^{g}\mathbf{M}(k)$. Define the matrix $\mathbf{N}(g)$ to be diagonal with $1/N_1(g)$ and $1/N_2(g)$ as its diagonal elements. The probability of coalescence from generation $g-1$ to generation g is then

$$c_g = \left[\mathbf{v_i}(0)\prod_{k=0}^{g}\mathbf{M}(k)\right]\mathbf{N}(g)\left[\mathbf{v_j}(0)\prod_{k=0}^{g}\mathbf{M}(k)\right]^{\mathsf{T}} \tag{4.9}$$

and the probability of the two individuals to coalesce g generations before present is

$$p(g|\mathbf{M}(g),\mathbf{N}(g)) = c_g\prod_{k=1}^{g-1}(1-c_k) \tag{4.10}$$

Equation 4.10 can be used to compute

$$\int_u^v p(l|\mathbf{M}(g),\mathbf{N}(g))\,\mathrm{d}l = \sum_{g=1}^{\infty}\left[c_g\prod_{k=1}^{g-1}(1-c_k)\int_u^v p(l|g)\,\mathrm{d}l\right] \tag{4.11}$$

Note that Equation 4.11 is very general, and we can allow additional demographic changes to take place. For instance, by setting $N_2(g) = 0$, $m_{12}(g) = 0$ and $m_{21}(g) = 1$ for all $g > G$, we encode a population split that occurred G generations ago. In practice, a pair of populations will have split a number of generations back in time, and it is therefore convenient to consider models of the kind depicted in Figure 4.1b. In this model a population of constant size N_{atot} splits G generations in the past, forming two populations of size N_{a1} and N_{a2}. The size of these two groups then fluctuates in time, to reach a present size of N_{c1} and N_{c2}. During their separation, the populations exchange individuals at a rate of m_{12} and m_{21} per generation, per individual. Of course, other models can be defined, allowing variable migration rates, and different population size dynamics.

For mathematical convenience, it is safe to assume the ancestral population size becomes constant a number of generations in the past. Models where the ancestral population size (N_{atot} in Figure 4.1b) is constant from generation G to infinity allow for a closed form computation of Equation 4.11, no matter which demographic dynamics take place from generation 0 to G (Equations 3.9 and 3.10). Furthermore extremely remote demographic events have negligible impact on shared haplotypes of currently detectable lengths (e.g. > 1 cM).

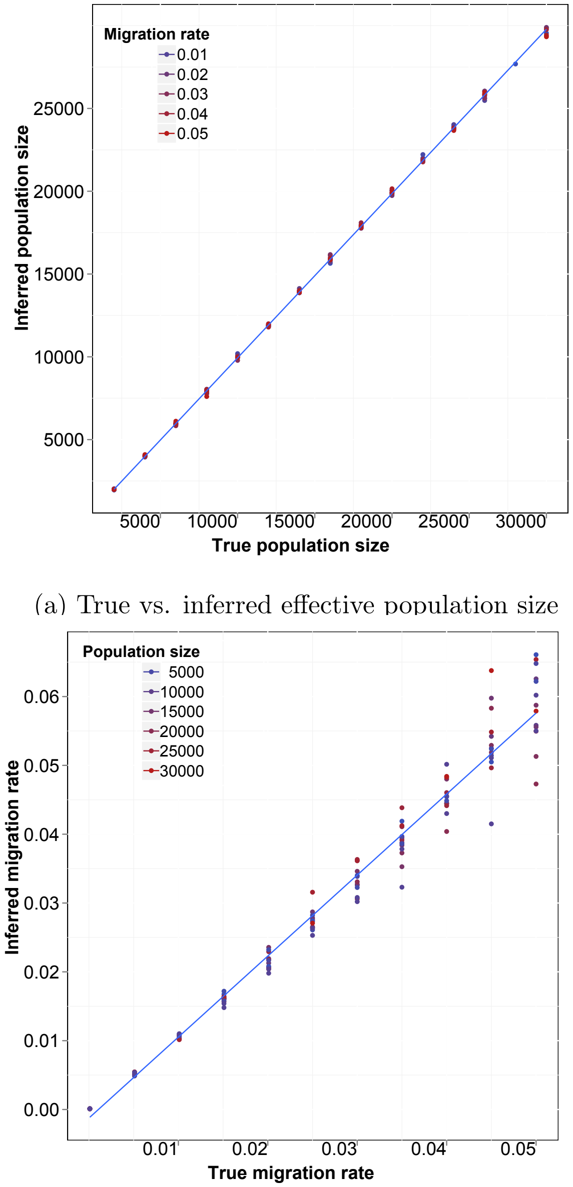
4.1.2 Simulations, ancestry deconvolution and real data

We tested our framework using extensive simulation of realistic chromosomes under several demographic models, using the GENOME coalescent simulator ([Liang *et al.*, 2007]). For computational convenience, we set the size of the simulator's non-recombinant segments between 0.01 and 0.025 cM, always using a recombination rate of 1cM/Mb. A modified version of the simulator was used to extract ground truth IBD haplotypes from the simulated genealogies, defined as segments co-inherited by pairs of individuals from their most recent common ancestor (see definition (b) in Section 1.1.4.1). For some of the simulations we inferred shared haplotypes using the GERMLINE software package ([Gusev *et al.*, 2009]) on phased genotype data, which was obtained setting GENOME's mutation rate to 1.1×10^{-8} per base pair ([Roach *et al.*, 2010]). Genotypes were post-processed to mimic the information content of array data. To this extent, we computed the allele frequency spectrum of European individuals from the HapMap 3 dataset ([Frazer *et al.*, 2007]), using frequency bins of 2%. We then randomly selected the same proportion of alleles from the simulated genotypes. We obtained an average density of ~230 single nucleotide polymorphisms/Mb.

In order to compare the proposed IBD-based approach for migration inference to the approach of [Gravel, 2012], which is based on ancestry deconvolution, we simulated synthetic datasets under several demographic models, and extracted genotype data as previously described. We then ran the PCAdmix software ([Brisbin *et al.*, 2012]) with windows of size 0.3cM and the genetic map used in the simulations. The output of PCAdmix was used to infer migration rates via the Tracts software package ([Gravel, 2012]). IBD information was computed in the same datasets running the GERMLINE software, and the output was used to infer migration rates using the DoRIS software package, which implements the proposed framework. Perfectly phased haplotypes were used in input for both PCAdmix and GERMLINE. Only migration rates were inferred, while all other demographic parameters were set to the true simulated values for both Tracts and DoRIS.

To demonstrate the use of the DoRIS framework on real data, we analyzed 56 trio-phased samples from the HapMap 3 dataset. Phased genotypes were downloaded from the HapMap 3 web page at http://hapmap.ncbi.nlm.nih.gov. IBD haplotypes were extracted using GERMLINE, as previously described in [Palamara *et al.*, 2012].

4.2 Results of evaluation and real data analysis



(a) True vs. inferred effective population size

(b) True vs. inferred migration rate

Figure 4.2: True vs. inferred parameters for the model in Figure 4.1a.

4.2.1 Constant size and symmetric migration rates

In order to test the accuracy of demographic inference based on the proposed model, we initially simulated a number of populations of constant size N_e, which exchange individuals at a constant, symmetric migration rate m, as depicted in the model of Figure 4.1a. We simulated 15 possible sizes of synthetic populations, ranging from $2,000$ to $30,000$ haploid individuals, with increments of $2,000$. For each population size, we simulated 11 possible migration values, uniformly chosen between 10^{-4} and 5×10^{-2}. For a total of 165 datasets, we simulated a chromosome of 300 centimorgans for 500 haploid individuals from each subpopulation, and computed IBD sharing within and across populations. The simulations used non-recombining blocks of 0.02 cM. This resolution may introduce small biases in the analysis, which we found to be negligible in our previous work. We then used Equation 4.7 to estimate $\hat{m}$ and $\hat{N}_e$, with results shown in Figure 4.2. To test the model's accuracy, for this analysis we only considered ground-truth IBD segments extracted from the synthetic genealogies.

We obtained a good correspondence between the true population size and the size inferred via the estimator of Equation 4.7, with almost perfect correlation shown in Figure 4.2a. Inferred migration rates were also very close to the simulated rates, although a moderate upward bias and higher estimation variance for large migration rates was observed in this case (Figure 4.2b). In addition to using the effective population size estimator of Equation 4.7, we used the estimator of Equation 3.27, which assumes constant population sizes and no migration. As expected, the inferred recent effective population size was in this case inflated by the presence of migration, as shown in Figure 4.3. When migration rates are increased, the inferred population size quickly approaches the total population size (in this case $2N_e$).

4.2.2 Dynamic size and asymmetric migration rates

We then tested our model's performance in the more complex demographic scenario depicted in Figure 4.1b, where a population splits into two subpopulations which grow at different exponential rates, interacting with asymmetric migration rates. We simulated a chromosome of ~275 cM for 500 haploid individuals per subpopulation. Simulated non-recombinant

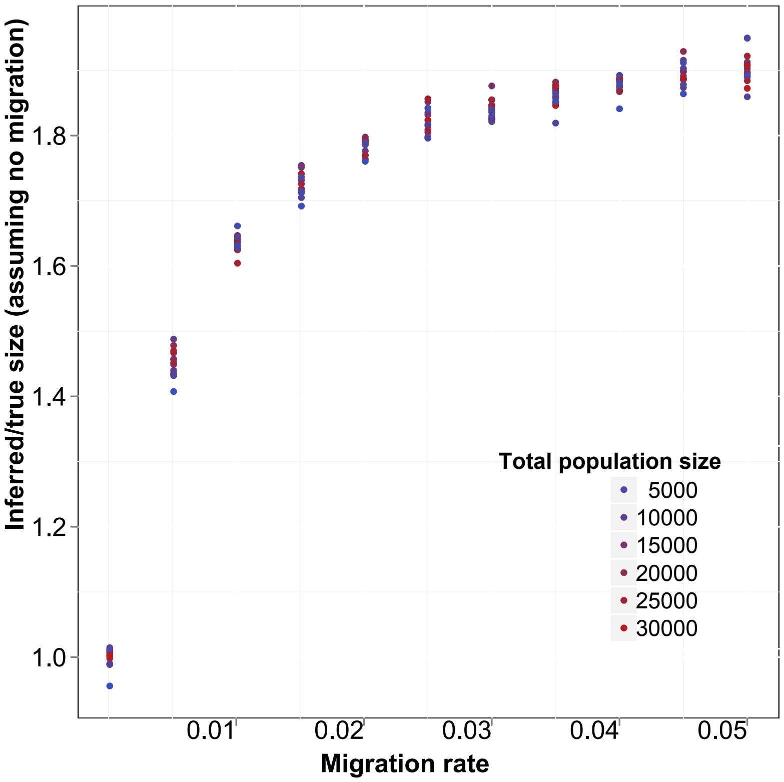
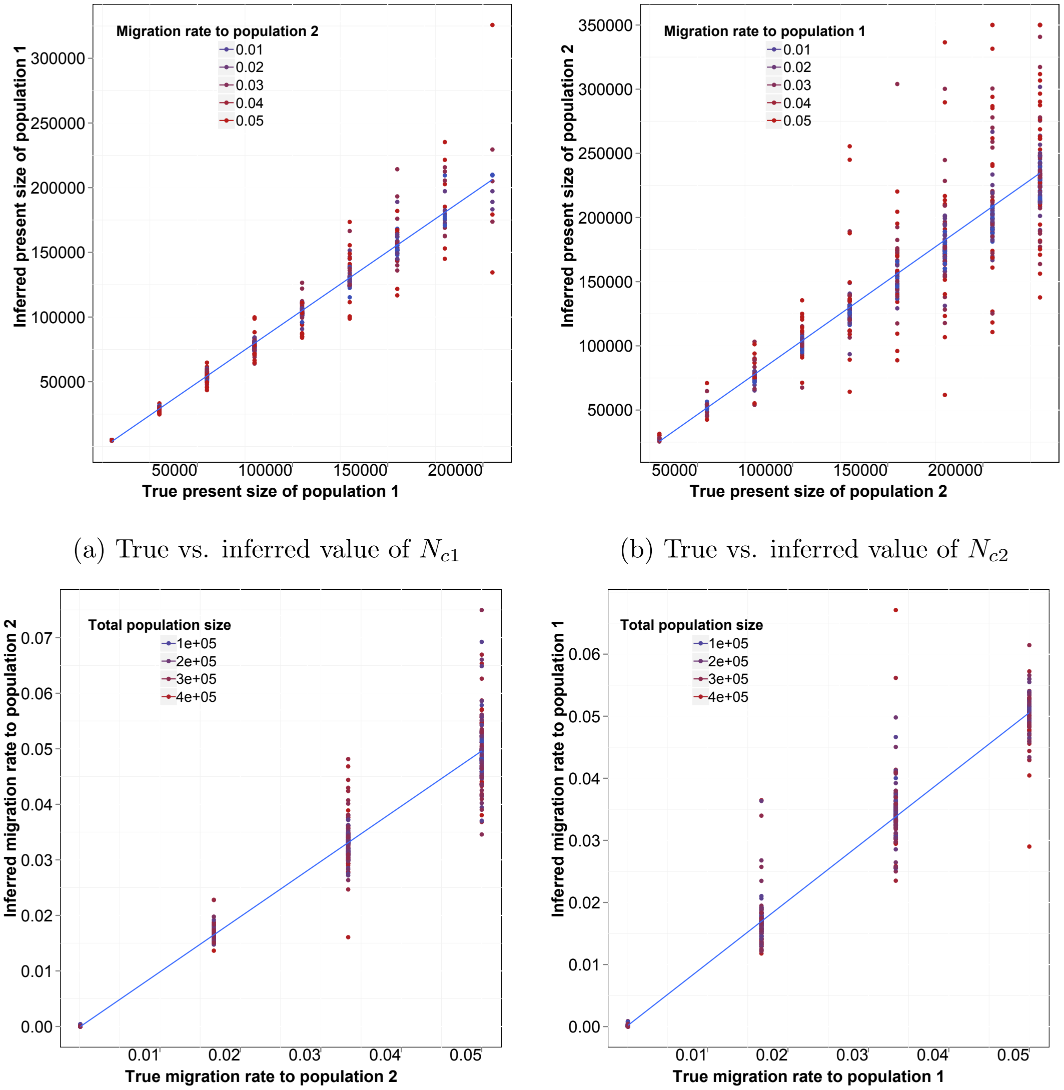


Figure 4.3: Inference of recent effective population size using Equation 3.27, which neglects migration. The ratio between inferred and true population size (y axis) increases as the migration rate (x axis) is increased, approaching the sum of population sizes for both populations (twice the true size).

blocks had size 0.025 cM. In all simulated scenarios, we kept N_{atot} fixed to $10,000$ haploid individuals, while N_{a1} and N_{a2} were kept fixed at $5,000$ individuals. For N_{c1} and N_{c2} we simulated all possible combinations of sizes between $5,000$ and $205,000$ haploid individuals, with increments of $15,000$ (excluding cases where $N_{c1} = N_{c2}$). Note that on average the simulated values of N_{c1} were smaller, resulting in higher inference accuracy compared to N_{c2}. For each pair of population sizes we simulated values of m_{12} and m_{21} using all combinations of the migration rates 0.0001, 0.0167, 0.0334, and 0.5.

A total of 540 synthetic populations were tested. For each synthetic population we extracted the average fraction of genome shared through haplotypes of different length in-



(a) True vs. inferred value of N_{c1}

(b) True vs. inferred value of N_{c2}

(c) True vs. inferred value of m_{12}

(d) True vs. inferred value of m_{21}

Figure 4.4: Results of the evaluation of our method on synthetic populations with demographic history depicted in the model of Figure 4.1b. Higher variance in the method's accuracy is observed due to limited sample sizes and increased population sizes. Higher migration rates further decrease the rate of coalescent events in the recent generations (Figure 4.4b), resulting in additional uncertainty. However no significant bias is observed in the inference.

tervals by pairs of individuals within each population or across populations. As in our previous work, we used a combination of intervals of uniform length and length intervals corresponding to quantiles of the Erlang-2 distribution, which is used in $p(l|t)$. Inference performance was tested via minimization of the root mean squared deviation between observed and predicted average fraction of shared genome. Note that a likelihood based approach (e.g. considering the number of shared segments) could be used based on the quantities derived in the previous chapter. We scanned several possible values for one parameter at a time, performing a line search while fixing the remaining model parameters to the true simulated value. The results of this analysis are reported in Figure 4.4.

As expected, due to the large recent effective population sizes we simulated, the variance of the inference accuracy was higher in this scenario, suggesting that more than a single chromosome for 500 diploid individuals may be required for the analysis of these demographies. A single chromosome of ∼250 centimorgans sampled in 500 diploid individuals is in fact equivalent for the purpose of this inference to the analysis of all the autosomal chromosomes for ∼150 diploid samples (see [Palamara *et al.*, 2012]). Larger population sizes result in lower signal to noise ratio for the estimation of the expected fraction of genome shared via IBD segments, and increasing sample size or analyzing additional chromosomes is expected to reduce the variance in the inference performance. Lower accuracy was observed in the inference of N_{c2} since, as previously mentioned, this simulated subpopulation was on average larger. Inferred population sizes were more accurate in the presence of low migration rates (represented by colors in figures 4.4a and 4.4b), as high migration further reduces the chance of early coalescent events, exacerbating the effects of large population sizes. Overall, no significant bias was observed in the recovered parameter values, suggesting our model provides a good match for the empirical distributions.

4.2.3 Applicability of the model to genotype data

While the previous analysis was mainly concerned with testing the model's accuracy, and relied on ground-truth IBD sharing extracted from the simulated genealogies, it is interesting to ask whether this approach can be used on genotype data. To this extent, we simulated genotypes for the demographic model of Figure 4.1a. We set the population sizes to $4,000$ or $12,000$ diploid individuals per population, and extracted 300 diploid sampled from each group. The migration rate was symmetric, and set to 0.04 per individual, per generation. Chromosomes of 150 cM were simulated using non-recombinant blocks of size 0.01 cM, and the synthetic genotypes were post-processed to reproduce the density and allele frequency spectrum of realistic SNP array data. In addition to extracting the ground truth IBD information as previously described, we inferred IBD haplotypes from the simulated genotypes using the GERMLINE software. The results suggest that when accurate phase information is available (e.g. for the X Chromosome, or for trio-phased samples), GERMLINE is able to recover the IBD sharing distribution across any pair of samples with high fidelity (Figure 4.5). However, when the samples were computationally phased using the Beagle software ([Browning and Browning, 2007]), GERMLINE had an inconsistent performance, accurately recovering the IBD sharing in the case of $N = 4,000$, while poorly inferring long haplotypes in the case of $N = 12,000$. This suggests that additional care must be taken when analyzing computationally phased data, particularly when analyzing cross-population IBD spectra, were the quality of the inferred IBD haplotypes will likely vary from population to population, as a result of different underlying demographic histories.

4.2.4 MKK analysis, revisited

To demonstrate the applicability of our method to real data, we analyzed the HapMap 3 Masai dataset, which was already studied in our previous work using a simulation-based approach. We here revisit this analysis, using the described analytical framework.

Cryptic relatedness across individuals in this dataset is extremely common, and does not appear to be due to the presence of occasional outliers among the samples. Demographic reports are not supportive of recent population bottlenecks in this group, which is though to be slowly but steadily expanding ([Coast, 2001]). The Masai are a semi-nomadic people, and individuals often reside in small communities (*Manyatta*) of tens to few hundreds of members. To study their demography, we therefore use a model where V villages of constant size N exchange individuals at a constant and symmetric rate m. This model is similar to the one depicted in Figure 4.1a, with symmetric migration rates across several populations. We assumed that all samples were extracted from the same village, and used the model described in Section 4.1.1 for the analysis. We performed a grid search testing migration rates from 0.01 to 0.4, with intervals of 0.01, village sizes from 50 to $4,000$ with steps of 10, and number of villages from 3 to 150 with increments of 1. We also obtained 95% confidence intervals for the inferred values using a bootstrap approach, by creating 400 resamples randomly selecting individuals with replacement, then recomputing the optimal parameters using a gradient-driven procedure, which was initialized using the parameters inferred using the original samples (note, however, that small correlations exist for IBD sharing across individual pairs, and this method may provide optimistic intervals). Using this approach, we obtained the following estimates: $V = 58$ (95% CI: 46 to 75), $N = 400$ (95% CI 360 to 470), and $m = 0.1$ (95% CI 0.09 to 0.12).

4.2.5 Comparison with existing methods

The structure of long-range haplotypes is known to carry relevant information about recent population dynamics, but this genomic feature has only recently become observable thanks to the development of modern high-throughput genomic technologies. As a consequence, methods that rely on a population's haplotypic structure to reconstruct demographic events have only recently arose. A model proposed in [Pool and Nielsen, 2009], and recently expanded in [Gravel, 2012], provides a way to analyze the distribution of migrant tracts and infer the timing and intensity of very recent migration events. In order to analyze the distribution of migrant haplotypes, however, ancestry deconvolution needs to be accurately performed. This typically requires the availability of two suitable reference populations, which are required to be sufficiently diverged from each other. The amount of required divergence depends on the specific method used for the deconvolution, but in general this poses significant constraints in terms of the demographic scenarios that can be analyzed using these methods.

To compare our IBD-based approach to methods based on ancestry deconvolution, we simulated the demographic scenario of Figure 4.6, where two populations split G_s generations in the past, and G_a generations in the past contribute a fraction of genomes to the creation of a group of admixed individuals, with probability m and $1 - m$, through a unique pulse of migration. All three population sizes were fixed to either $N = 5,000$ or $N = 10,000$, m was set to 0.2, and G_a was 25 in all simulations. We varied G_s from 40 to 600, with increments of 20, and extracted genotype data on a single 400 cM chromosome for 250 diploid samples in each of the three extant populations. We used the output of the PCAdmix software as input for the Tracts program ([Gravel, 2012]), and the IBD segments retrieved by GERMLINE as input for the DoRIS software. Note that for the IBD analysis we only used the 250 admixed samples and the 250 samples from the population contributing $\sim m$ haplotypes at generation G_a, while the samples from the third population were ignored. In both cases we inferred the value of m, setting all other parameters to the true simulated values, with results shown in Figure 4.7.

DoRIS performed better on average (mean inferred $m = 0.205$, std 0.025), while providing slightly noisy results, suggesting the need for a larger sample size and/or the analysis of additional chromosomes. The migration rate inferred by Tracts (mean $m = 0.104$, std 0.0233) was strongly biased. We note that in this setting Tracts is essentially used to only report the proportion of ancestry inferred by the deconvolution method, which is the actual source of inaccuracy. Even for populations that diverged 600 generations in the past ($\sim$15,000 years before present assuming a generation of 25 years), the recovered rate was substantially lower than the simulated rate. The case of $N = 5,000$ yielded better estimates, due to the higher drift found in smaller populations, which improved the power of PCAdmix to call migrant tracts. We additionally run the PCAdmix+Tracts analysis on longer time scales, simulating values of G_s from 200 to 6,000, with intervals of 200 generations, using $N = 10,000$. Even for several thousand generations since the split of the reference populations, a small bias was observed (Figure 4.8).

This analysis suggests that while the methods that rely on ancestry deconvolution are a useful tool for the specific case of recently admixed groups arising from strongly diverged populations, they may not be suitable for the analysis of fine-scale migration events, such as those that occurred across populations that split few tens to hundreds of generations

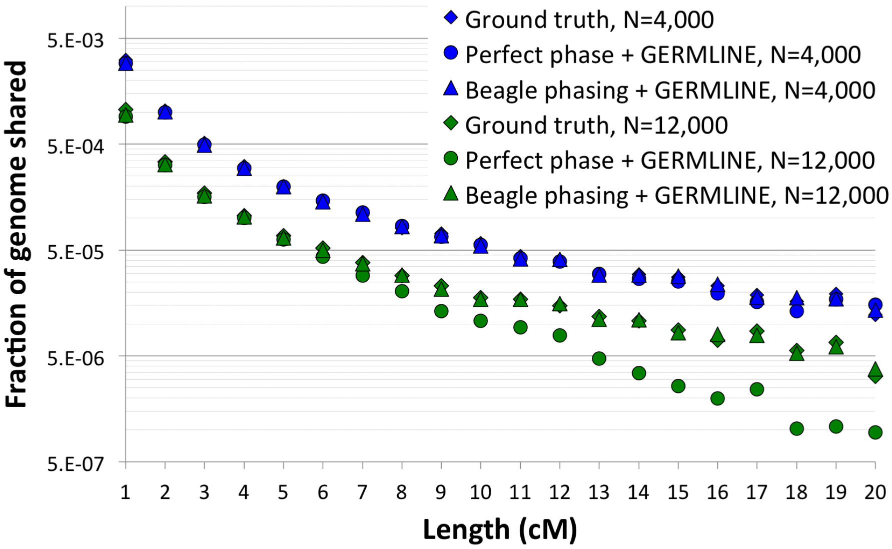


Figure 4.5: We simulated a chromosome of 150 cM for 600 individuals using the model in Figure 4.1a, setting population sizes to 4, 000 and 12, 000 diploid individuals, with a migration rate of 0.04. IBD sharing was extracted directly from the simulated genealogy (diamonds), or inferred trough GERMLINE using perfectly phased (circles) or computationally phased (triangles) chromosomes.

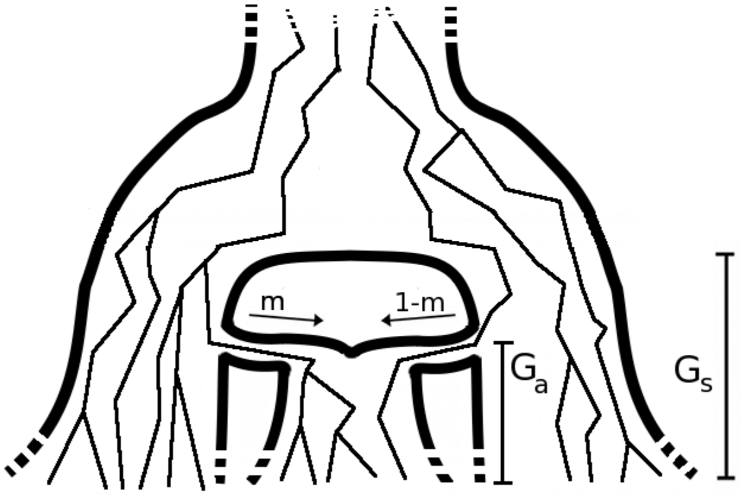


Figure 4.6: The model used to simulate admixed populations.

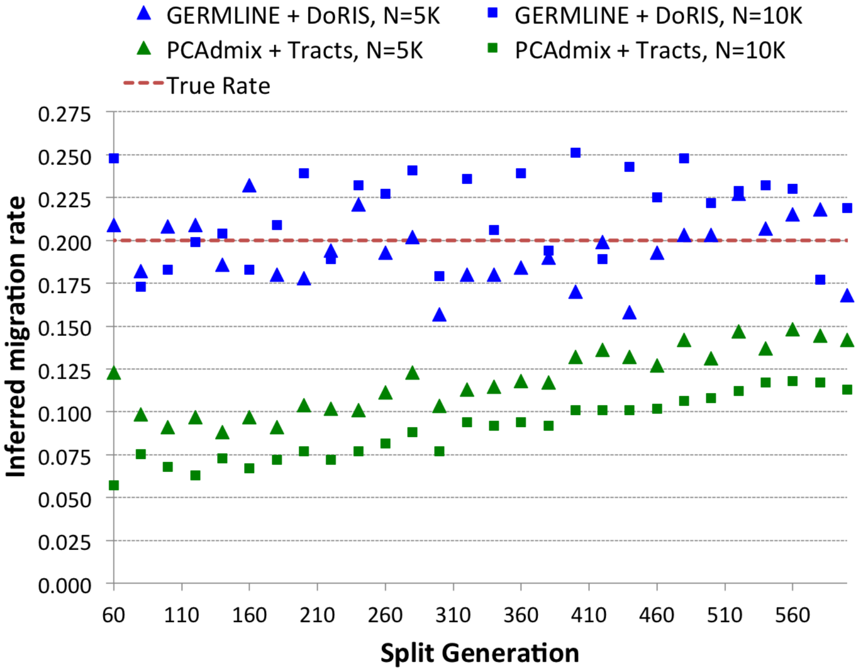


Figure 4.7: We created several simulation genotype datasets using the model in Figure 4.6, varying G_s while keeping $m = 0.2$, $G_a = 25$, and using constant populations of size $5,000$ or $10,000$ diploid individuals. We inferred the value of m using PCAdmix+Tracts, or GERMLINE+DoRIS, here reported as a function of G_s.

in the past. It is however possible that adjusting some of the parameters used for the GENOME simulations and for the PCAdmix software, or using other deconvolution methods, the obtained accuracy may be increased. Furthermore, the development of methods for ancestry deconvolution in sequence data, where rare variants are observable, is expected to substantially increase the power of this analysis, although the effects of limited population divergence are likely to still affect the accuracy of methods that do not explicitly take this aspect into account. An additional difference to be noted between the two considered approaches is that Tracts does not model population size changes in the populations, focusing on relative migration rates, while DoRIS allows recovering both population size fluctuations and migration rates, thus providing insights into the magnitude of migration events. This increased flexibility, however, may complicate the inference, also in light of our observation

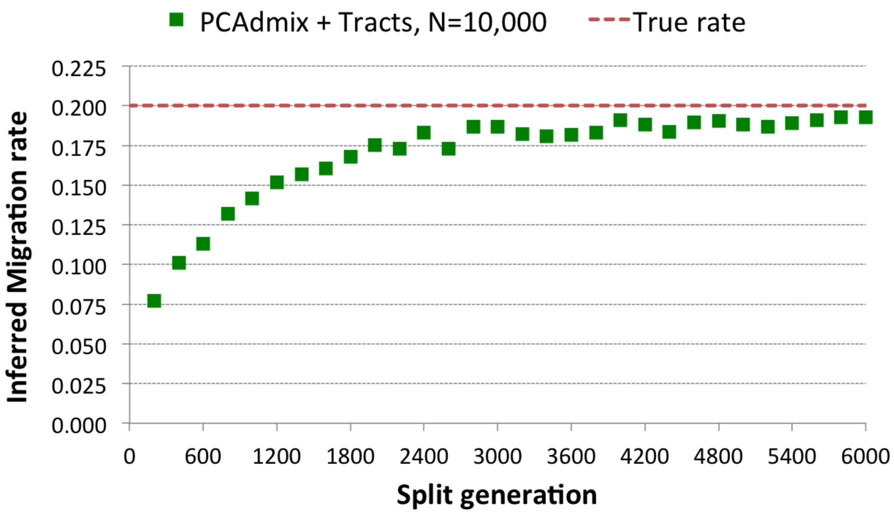
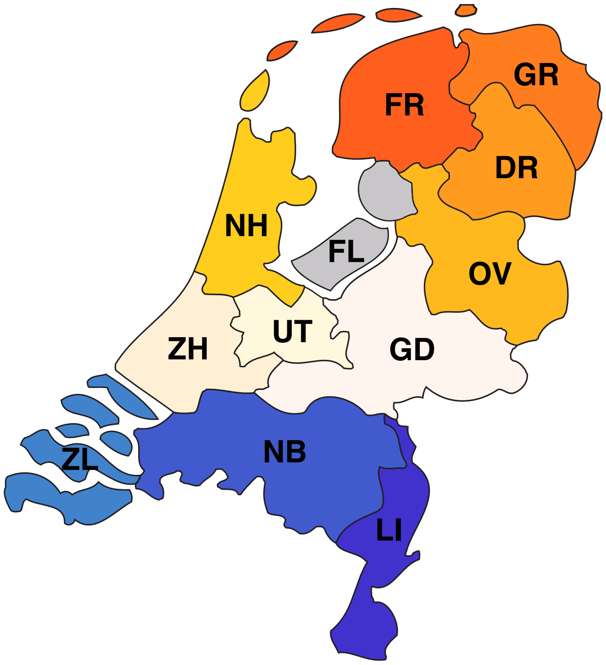


Figure 4.8: We created several datasets using the model in Figure 4.6, varying G_s from 200 to $6,000$, and using $m = 0.2$, $G_a = 25$ with population sizes of $10,000$ diploid individuals. We inferred the value of m using PCAdmix+Tracts from phased genotype data.

that large sample sizes are required for the IBD analysis.

4.3 IBD sharing outlines regional-scale demographic history: the Genome of the Netherlands

The Genome of the Netherlands (GoNL) Project was established with the goal of characterizing genomic variation in the Dutch population [Boomsma *et al.*, 2013]. To this extent, 250 trio-families from 11 provinces of the Netherlands were sequenced with an average coverage of 14-15x. Sequencing data provides a very large number of informative (high frequency) variants, which added to the trio design of the GoNL project results in reliable inference of haplotype phase, therefore enabling accurate detection of IBD sharing. Furthermore, fine-grained information about genomic variation across 11 provinces from a single country provides the opportunity to demonstrate the methods developed in this and the previous chapter for the inference of fine-scale demographic history, using the DoRIS software package. Results presented in this section are described in [The Genome of the Netherlands



Province	Code	N
Friesland	FR	62
Groningen	GR	57
Drenthe	DR	56
Overijssel	OV	57
Noord-Holland	NH	91
Utrecht	UT	48
Gelderland	GD	57
Zuid-Holland	ZH	168
Zeeland	ZL	46
Noord-Brabant	NB	67
Limburg	LI	58

Figure 4.9: A map of the analyzed provinces and the number of collected samples.

Consortium, 2014].

We initially removed all variants with a minimum allele frequency below 1%, and also excluded from downstream analysis all markers with trio-phasing posterior probability [Menelaou and Marchini, 2013] below 1.0, obtaining 3,525,142 SNPs. We used the genetic map provided for the Phase I integrated variant set release (v3) of the 1,000 Genomes Project[1] (build 37, hg19 coordinates). For all markers that were not found in the genetic map, we inferred the genetic distance by linear interpolation assuming uniform recombination rate between the two closest markers found upstream and downstream in the available map. We run GERMLINE using the parameters *-min_m 1 -err_hom 2 -err_het 0 -bits 75 -haploid*, i.e. requiring a minimum IBD segment length of 1 cM, allowing at most 2 mismatches in windows of 75 markers, for perfectly phased haploid chromosomes. We excluded regions with unusual density of IBD sharing, which may be caused by false positive/negative segments due to low density of markers, deviations from neutrality or presence of common

[1] `http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html`

structural variations that affect recombination. To this extent, we restricted our analysis to regions with IBD density within 5 standard deviations from the mean genome-wide sharing. We further required the analyzed regions to be at least 45 cM long, obtaining a total of 26 regions spanning $2,160.26$ cM (Table 4.1), IBD sharing density per site per pair 3.07×10^{-3}, std 1.35×10^{-3}).

When we analyzed long IBD segments (at least 7 cM), recently co-inherited within each province, differences in the inferred effective population sizes (obtained using the methods of Chapter 3) were substantial (Figure 4.10), reflecting recent differentiation as a result of heterogeneous growth and migration events that occurred in the past 20-25 generations (expected time to recent common ancestor based on segment length in a model of exponential expansion: ∼500 years ago). In particular, Zuid-Holland exhibits an effective population size of $> 100,000$ individuals (eight times that of Overijssel), suggestive of recent expansion and possibly gene flow from other provinces. The sharing of such long IBD segments (also across provinces) supports localized recent common ancestry, with all provinces sharing, on average, the largest number of long IBD segments with other individuals from the same geographic region (Figure 4.11).

We then considered the fraction of genome shared within each province through short IBD segments (1 to 2 cM), and thus inferred the ancestral effective population size per province (Figure 4.12). Although these effective population sizes are rather homogeneous across the 11 provinces, consistent with common genetic origins, we observed a South to North gradient of decreasing ancestral population size accompanied by increased homozygosity in the northern provinces (correlation between province latitude and IBD sharing $r = 0.923$, $p = 5.12 \times 10^{-5}$). Such gradient has been previously described for average inbreeding coefficients and similar metrics of genome-wide similarity across Dutch individuals [Lao *et al.*, 2013; Abdellaoui *et al.*, 2013], and interpreted as the signature of remote northwards migration during early waves of European colonization, although more complex scenarios involving recent demographic events could not be ruled out. Indeed, a model of serial founder migrations from the South to the North of the country may produce the observed pattern of increasing homozygosity towards the North, as the results of shrinking

From Chromosome	From (genetic)	To (genetic)	From (physical)	To (physical)
1	97.5	150.5	66,874,699	118,837,888
2	36.5	115	17,246,473	85,384,179
2	209	257.5	193,010,478	235,351,139
3	1	190	678347	176030190
4	101	217.5	85315581	189,657,996
5	38	148	22,657,926	141,420,437
6	54	111	33,954,192	103,983,460
6	150.5	197.3	139,903,959	170,245,872
7	1	63	962,247	38,722,532
7	66.5	172	41,688,961	152,254,508
8	78	166	55,170,178	139,553,601
9	83	158	72,512,292	132,515,730
10	44	181.5	19,570,732	134,866,854
11	5	160.9	2,047,054	134,587,122
12	16.5	90	6,476,123	75,656,510
12	98.5	158.5	82,586,486	128,401,829
13	1.4	128.8	20,518,406	114,094,544
14	1.1	54	20,545,390	59,184,876
14	58	113.5	63,846,103	104,808,535
15	70	149.9	50,284,344	101,969,749
17	1.1	84.5	163,278	55,936,970
18	37.5	84	11,962,813	59,189,703
19	26	105.9	7,857,579	158,513,172
20	19	83	5,649,902	52,818,462
21	1.9	63.7	15,636,220	47,031,048
22	21	74.1	23,874,416	50,493,062

Table 4.1: Regions that passed quality control and were analyzed in the GoNL dataset.

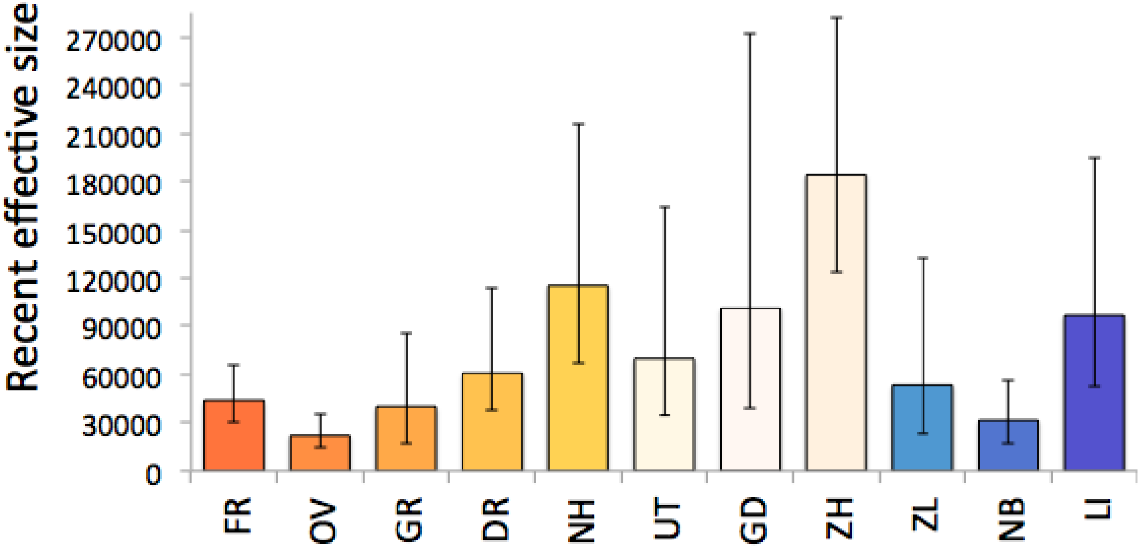
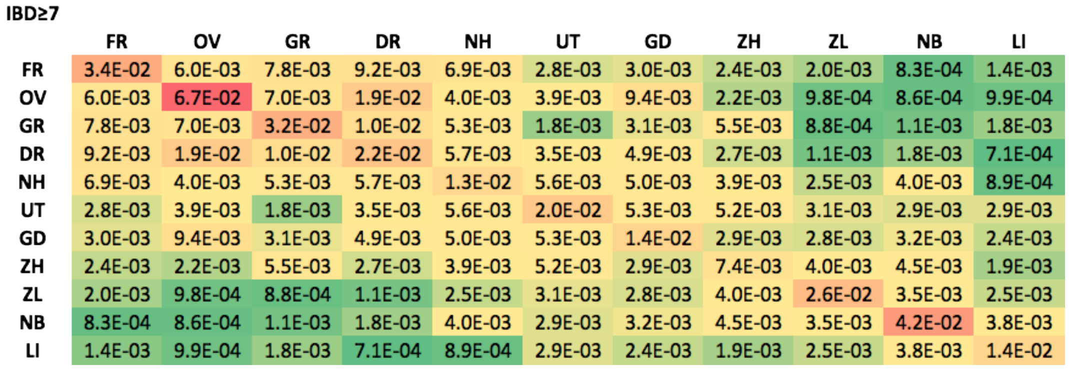


Figure 4.10: Reconstructed recent population sizes.



IBD≥7

	FR	OV	GR	DR	NH	UT	GD	ZH	ZL	NB	LI
FR	3.4E-02	6.0E-03	7.8E-03	9.2E-03	6.9E-03	2.8E-03	3.0E-03	2.4E-03	2.0E-03	8.3E-04	1.4E-03
OV	6.0E-03	6.7E-02	7.0E-03	1.9E-02	4.0E-03	3.9E-03	9.4E-03	2.2E-03	9.8E-04	8.6E-04	9.9E-04
GR	7.8E-03	7.0E-03	3.2E-02	1.0E-02	5.3E-03	1.8E-03	3.1E-03	5.5E-03	8.8E-04	1.1E-03	1.8E-03
DR	9.2E-03	1.9E-02	1.0E-02	2.2E-02	5.7E-03	3.5E-03	4.9E-03	2.7E-03	1.1E-03	1.8E-03	7.1E-04
NH	6.9E-03	4.0E-03	5.3E-03	5.7E-03	1.3E-02	5.6E-03	5.0E-03	3.9E-03	2.5E-03	4.0E-03	8.9E-04
UT	2.8E-03	3.9E-03	1.8E-03	3.5E-03	5.6E-03	2.0E-02	5.3E-03	5.2E-03	3.1E-03	2.9E-03	2.9E-03
GD	3.0E-03	9.4E-03	3.1E-03	4.9E-03	5.0E-03	5.3E-03	1.4E-02	2.9E-03	2.8E-03	3.2E-03	2.4E-03
ZH	2.4E-03	2.2E-03	5.5E-03	2.7E-03	3.9E-03	5.2E-03	2.9E-03	7.4E-03	4.0E-03	4.5E-03	1.9E-03
ZL	2.0E-03	9.8E-04	8.8E-04	1.1E-03	2.5E-03	3.1E-03	2.8E-03	4.0E-03	2.6E-02	3.5E-03	2.5E-03
NB	8.3E-04	8.6E-04	1.1E-03	1.8E-03	4.0E-03	2.9E-03	3.2E-03	4.5E-03	3.5E-03	4.2E-02	3.8E-03
LI	1.4E-03	9.9E-04	1.8E-03	7.1E-04	8.9E-04	2.9E-03	2.4E-03	1.9E-03	2.5E-03	3.8E-03	1.4E-02

Figure 4.11: Sharing of segments of at least 7 cM within and across provinces.

effective population sizes of smaller groups that migrate away from larger populations.

In addition to the observed gradient of increasing haplotypic homozygosity within provinces, however, we observed that all GoNL samples, regardless of modern-day geographic location, share on average more IBD segments with other individuals from the North of the country than with individuals from the same province (Figure 4.13, off-diagonal elements, correlation between average IBD sharing and average latitude of provinces $r = 0.934$, $p < 10^{-5}$). This

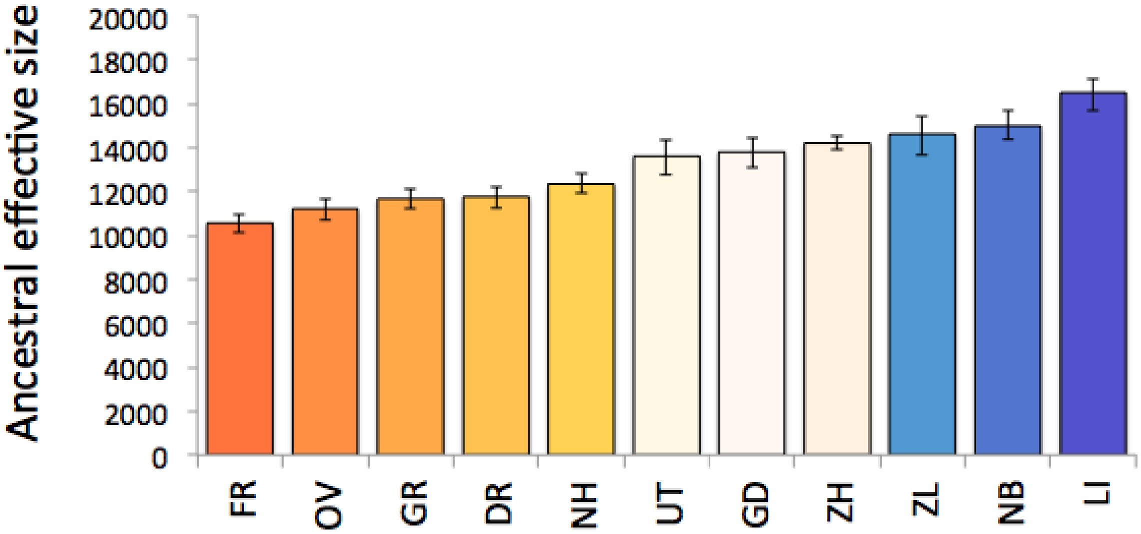
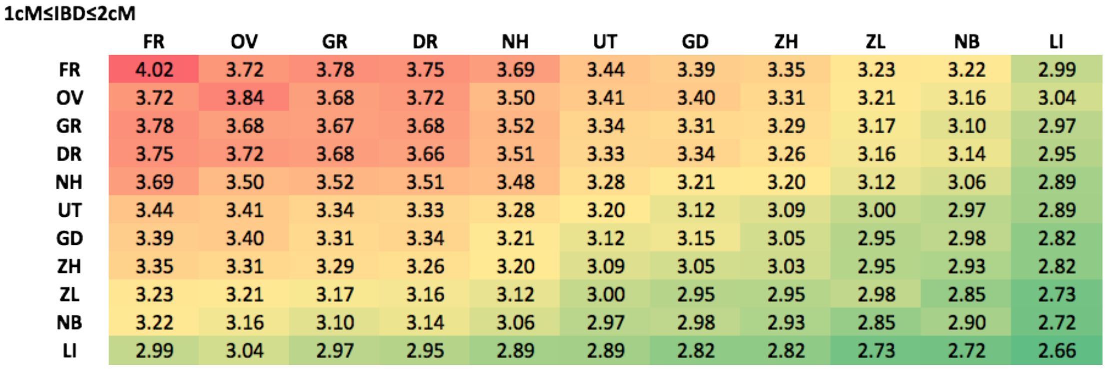


Figure 4.12: Reconstructed ancestral population sizes.



1cM≤IBD≤2cM

	FR	OV	GR	DR	NH	UT	GD	ZH	ZL	NB	LI
FR	4.02	3.72	3.78	3.75	3.69	3.44	3.39	3.35	3.23	3.22	2.99
OV	3.72	3.84	3.68	3.72	3.50	3.41	3.40	3.31	3.21	3.16	3.04
GR	3.78	3.68	3.67	3.68	3.52	3.34	3.31	3.29	3.17	3.10	2.97
DR	3.75	3.72	3.68	3.66	3.51	3.33	3.34	3.26	3.16	3.14	2.95
NH	3.69	3.50	3.52	3.51	3.48	3.28	3.21	3.20	3.12	3.06	2.89
UT	3.44	3.41	3.34	3.33	3.28	3.20	3.12	3.09	3.00	2.97	2.89
GD	3.39	3.40	3.31	3.34	3.21	3.12	3.15	3.05	2.95	2.98	2.82
ZH	3.35	3.31	3.29	3.26	3.20	3.09	3.05	3.03	2.95	2.93	2.82
ZL	3.23	3.21	3.17	3.16	3.12	3.00	2.95	2.95	2.98	2.85	2.73
NB	3.22	3.16	3.10	3.14	3.06	2.97	2.98	2.93	2.85	2.90	2.72
LI	2.99	3.04	2.97	2.95	2.89	2.89	2.82	2.82	2.73	2.72	2.66

Figure 4.13: Sharing of segments between 1 and 2 cM within and across provinces.

counterintuitive observation was confirmed when we grouped the 11 provinces into three clusters, North, Center and South, based on hierarchical clustering [Ward Jr, 1963] of the cross-province IBD matrix for segments of length 1 cM or more, and considered the average fraction of genome shared through IBD segments of at least 1 cM. Again, higher sharing between South and North than within the South was observed (and, similarly, more sharing across North and Center than within the Center). Consider the model of serial migrations shown in 4.14, and recall that to compute the fraction of genome shared through IBD segments of length at least u, we calculate $E[f|\theta] = \int_0^\infty \left[p(t_{mcra} = t|\theta) \int_u^\infty Erl_2(l; 2t)dl\right] dt$, where $p(t_{mcra} = t|\theta)$ represents the coalescent distribution for two individuals in a demographic model defined by θ, and $\int_u^\infty Erl_2(l; 2t)dl$ represents the probability of a genomic site being spanned by a segment of length at least u, given coalescence occurs at time t. The latter probability decreases monotonically for $u > 0$. If South and North do not exchange individuals after the initial split, the coalescence rate between present and the split time is zero, therefore the observation of high sharing across regions compared to within regions cannot be observed in the absence of subsequent migration.

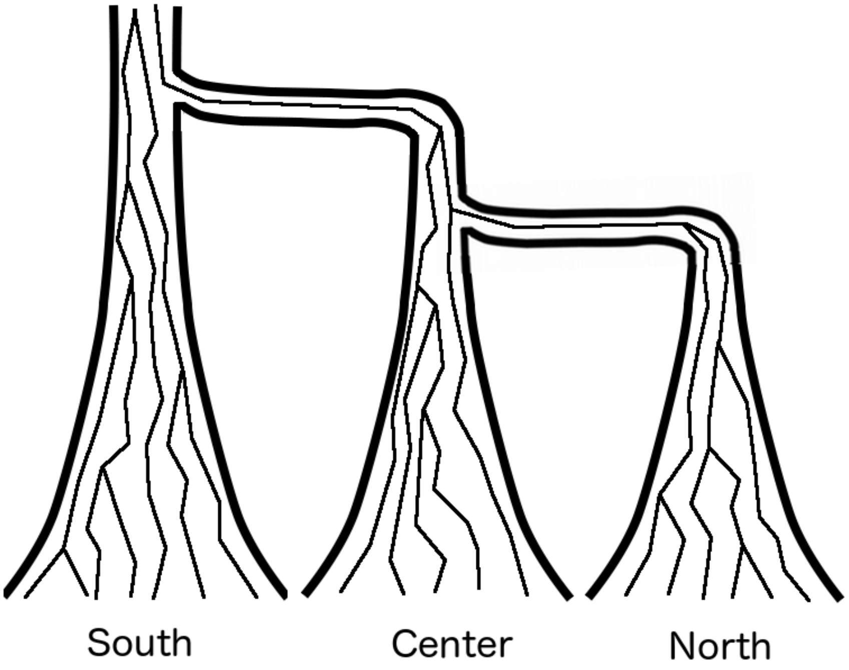
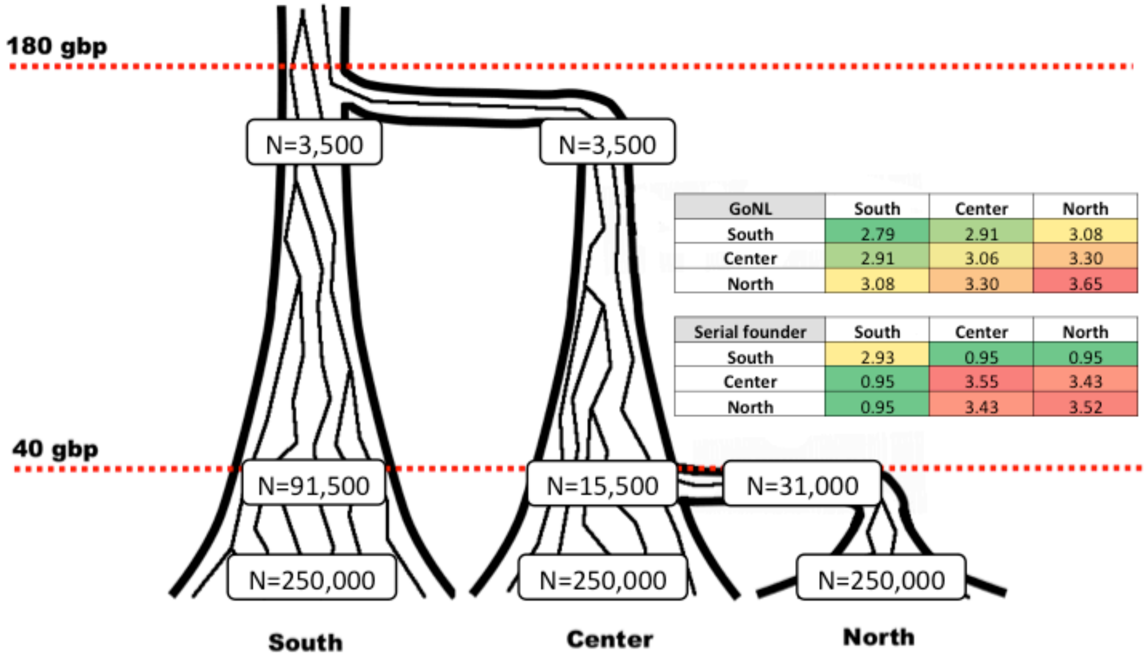


Figure 4.14: A model of serial migrations from the south to the north.

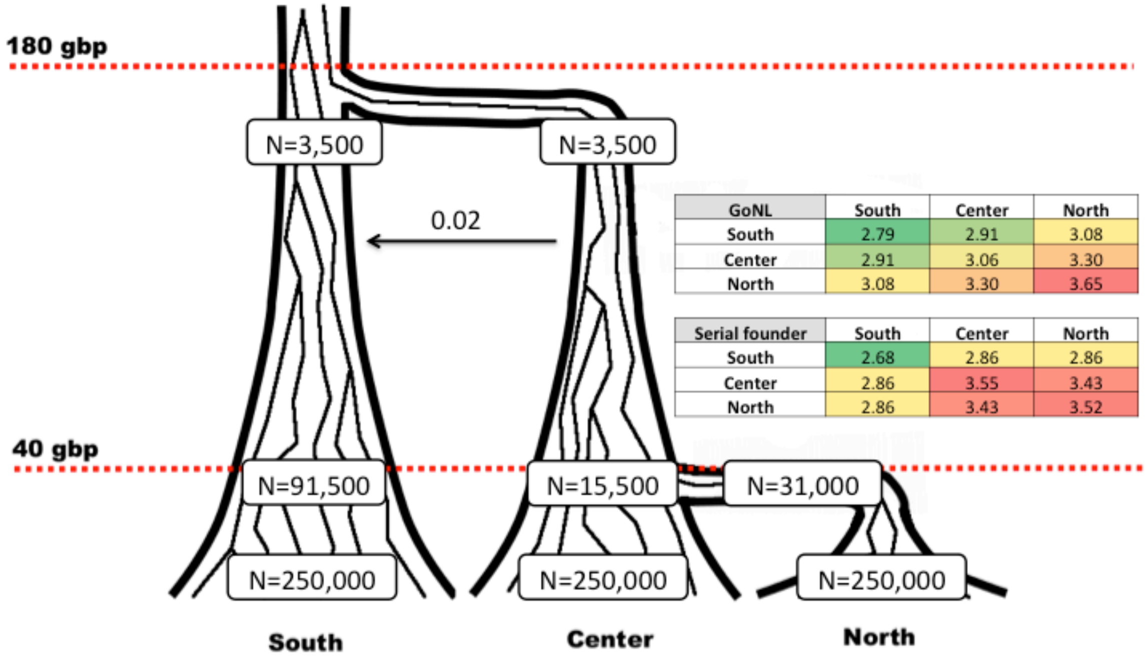
However, we note that the model of simple northwards serial migrations shown in Figure 4.14 may be enriched to explain the pattern observed in the GoNL data. Figure 4.15 shows the same simple model of subsequent population subdivisions. Increased sharing from the South to the Center and the North may be achieved by including migration from the Center to the South (forward in time) in the period preceding the formation of the Northern group



GoNL	South	Center	North
South	2.79	2.91	3.08
Center	2.91	3.06	3.30
North	3.08	3.30	3.65

Serial founder	South	Center	North
South	2.93	0.95	0.95
Center	0.95	3.55	3.43
North	0.95	3.43	3.52

Figure 4.15: Simple model of serial migrations and comparison of IBD patterns in GoNL and expected in this model.



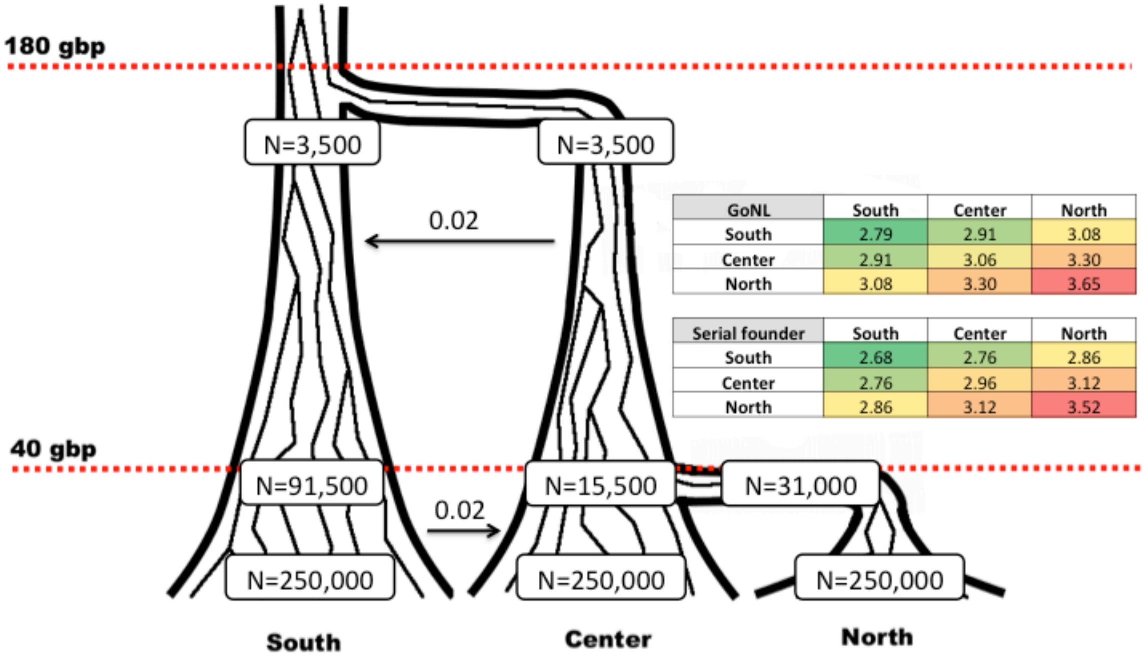
GoNL	South	Center	North
South	2.79	2.91	3.08
Center	2.91	3.06	3.30
North	3.08	3.30	3.65

Serial founder	South	Center	North
South	2.68	2.86	2.86
Center	2.86	3.55	3.43
North	2.86	3.43	3.52

Figure 4.16: Model of serial migrations that includes remote migrations to the South, and comparison of IBD patterns in GoNL and expected in this model.

(Figure 4.16). Finally, the increased sharing between Southern and Northern individuals may be obtained if individuals are allowed to migrate from the South back to the Center in the period following the creation of the Northern group, as shown in Figure 4.17.

We note that while the inclusion of these two migration rates in the model recapitulates the observed pattern of IBD sharing, comprehensive demographic analysis should include further testing of several models, and formal comparison across them. Further note that in this model a migration rate of 0.02 per individual, per generation, implies the ancestry of an individual has a relatively high 1/50 chance of moving between regions at each generation. During the course of 140 generations, this results in $1 - (1 - 0.02)^{140} = 0.94$ probability of migration for an ancestral lineage. For the more recent period, the chance of moving to the



GoNL	South	Center	North
South	2.79	2.91	3.08
Center	2.91	3.06	3.30
North	3.08	3.30	3.65

Serial founder	South	Center	North
South	2.68	2.76	2.86
Center	2.76	2.96	3.12
North	2.86	3.12	3.52

Figure 4.17: Model of serial migrations that includes remote migrations to the South and subsequent migrations to the Center. Comparison of IBD patterns in GoNL and expected in this model.

Center in this model is $1 - (1 - 0.02)^{40} = 0.55$. This model therefore implies substantial ancestral contribution from the genetic pool currently represented in the north of the country, rather than a unidirectional colonization from the South.

While a conclusive analysis would imply substantial additional hypothesis testing, possibly involving the inclusion of non-Dutch samples from neighboring regions, the compatibility of the presented model is in line with the complex fine-grained migratory history expected in the Netherlands. The Dutch territory was in fact shaped by frequent sea level changes and flooding events, which caused many parts of the coastal regions to vanish and reappear, likely resulting in several pulses of colonization and admixture across demes. A notable event is the occurrence of St. Lucia's flood in $1,287$ C.E., which separated the provinces of Noord Holland and Friesland. Interestingly, these two provinces exhibit high sharing for segments longer than 5cM, suggesting remote genetic links, which may go back to migration across these geographic regions before the occurrence of flooding.

Finally, we outline that the development of algorithms that detect short IBD segments in potentially heterogeneous groups is a subject of current research, and improvements in this direction will facilitate studying cross-population IBD sharing and strengthen the conclusions of these analyses. While the accuracy of currently available IBD detection methods in cross-population analysis is not fully understood (e.g. see Figure 4.5), the observed enrichment for IBD sharing with Northern provinces of the GoNL dataset was also observed in independent analysis performed using Beagle's FastIBD method for detecting haplotype sharing (correlation of within-province sharing and latitude: $r = 0.784$, $p = 4.27 \times 10^{-3}$, correlation between cross-province sharing and average latitude $r = 0.882$, $p < 10^{-5}$).

4.4 Appendix

4.4.1 Estimators for different N_{e1} and N_{e1}

When the population sizes of N_{e1} and N_{e2} are allowed to vary, the derivation of Section 4.1 leads to the following closed form estimators

$$\begin{aligned}\hat{N}_{e1} = \{&9\hat{f}_2^3 + 31\hat{f}_1^3 + 128\hat{f}_d^3 + 4\hat{f}_1\hat{f}_d(k - 18\hat{f}_d) + \\ &-3\hat{f}_1^2(18\hat{f}_d + k) + \hat{f}_2^2(49\hat{f}_1 - 10\hat{f}_d + 3k) + \\ &+\hat{f}_2[71\hat{f}_1^2 - 64\hat{f}_1\hat{f}_d - 4\hat{f}_d(22\hat{f}_d + k)]\}\times \\ \times\frac{1}{2u}&[\hat{f}_1(\hat{f}_2 + \hat{f}_1)^2(9\hat{f}_2 + 11\hat{f}_1) + 8\hat{f}_2\hat{f}_1(\hat{f}_2 + \hat{f}_1)\hat{f}_d + \\ &-4(4\hat{f}_2^2 + 19\hat{f}_2\hat{f}_1 + 13\hat{f}_1^2)\hat{f}_d^2 - 16\hat{f}_2\hat{f}_d^3 + 64\hat{f}_d^4]^{-1}\end{aligned} \tag{4.12}$$

$$\begin{aligned}\hat{N}_{e2} = \{&31\hat{f}_2^3 + 9\hat{f}_1^3 + 128\hat{f}_d^3 - 4\hat{f}_1\hat{f}_d(22\hat{f}_d + k) + \\ &+\hat{f}_1^2(3k - 10\hat{f}_d) + \hat{f}_2[49\hat{f}_1^2 - 64\hat{f}_1\hat{f}_d + \\ &+4\hat{f}_d(k - 18\hat{f}_d)] + \hat{f}_2^2[71\hat{f}_1 - 3(18\hat{f}_d + k)]\}\times \\ \times\frac{1}{2u}&[\hat{f}_2(\hat{f}_2 + \hat{f}_1)^2(11\hat{f}_2 + 9\hat{f}_1) + 8\hat{f}_2\hat{f}_1(\hat{f}_2 + \hat{f}_1)\hat{f}_d + \\ &-4(13\hat{f}_2^2 + 19\hat{f}_2\hat{f}_1 + 4\hat{f}_1^2)\hat{f}_d^2 - 16\hat{f}_1\hat{f}_d^3 + 64\hat{f}_d^4]^{-1}\end{aligned} \tag{4.13}$$

$$\hat{m} = \frac{u(k - 3\hat{f}_2 - 3\hat{f}_1 + 10\hat{f}_d)}{8(\hat{f}_2 + \hat{f}_1 - 2\hat{f}_d)} \tag{4.14}$$

where $\hat{f}_1, \hat{f}_2, \hat{f}_d$ are observed within and across populations, and

$$k = \sqrt{[9(\hat{f}_1 + \hat{f}_2) - 14\hat{f}_d](\hat{f}_1 + \hat{f}_2 + 2\hat{f}_d)} \tag{4.15}$$

4.4.2 Probability of coalescence in a two-population model with migration

The probability that the two ancestral lineages are found in the same population can be obtained from the terms specified in equations 4.1 and 4.2. Specifically, the chance of the two ancestral lineages being in the same population is given by

$$p(P_i(t) = P_j(t)) = \mathbf{v}_i(t)\mathbf{v}_j(t)^T \approx \frac{1 + e^{-4mt}}{2} \tag{4.16}$$

If both individuals are sampled from the same population, or

$$p(P_i(t) = P_j(t)) \approx \frac{1 - e^{-4mt}}{2} \tag{4.17}$$

Otherwise. We will focus on the first case, as the derivation in the latter case is analogous. In order to obtain the full coalescent distribution for the ancestral lineages of the considered two individuals sampled from one of the extant populations, we need to consider the chance that their ancestral lineages coalesce while being in the same populations. To do this, we consider the expected time that these lineages spend in the same population after t generation (expressed in continuous time). Such quantity can be computed as

$$\begin{aligned} E[f|T, m, N_e] &= \frac{\int_0^T p(P_i(t) = P_j(t))dt}{T} \\ &= \frac{\int_0^T 1 - e^{-4mt}dt}{2T} \\ &= \frac{1 - e^{-4mT} + 4mT}{8mT} \end{aligned} \tag{4.18}$$

A constant population of effective size N_e has coalescent distribution $p(t|N_e) = \frac{1}{N}e^{-t/N_e}$. Since the two populations have the same effective population size, N_e, we can obtain the coalescent distribution in the case of two population model by scaling the time in the distribution of single population model by the expected fraction of time spent in the same deme, which was computed above:

$$\begin{aligned} p(T|m, N_e) &= \frac{1}{N}e^{-E[f|T,m,N_e]T/N_e} \\ &= \frac{1}{N}e^{-\frac{1-e^{-4mT}+4mT}{8mN_e}} \end{aligned} \tag{4.19}$$

With cumulative distribution

$$P(T|m, N_e) = 1 - e^{-\frac{1-e^{-4mT}+4mT}{8mN_e}} \tag{4.20}$$

Note, however, that we are not interested in accurately describing this distribution for large values of T. In fact, we later introduce the factor $p(l|t)$, which quickly goes to 0 for values of u that are large enough to be practically considered, e.g. above 0.5 cM. For the purpose of this section, given reasonably large u and N_e (e.g. above $1,000$ effective individuals), the coalescent distribution can be approximated using a Taylor expansion of the kind $e^x \approx 1 + x$. Applying this approximation to the cumulative distribution we just derived, we get

$$\begin{aligned} P(T|m, N_e) &= 1 - e^{-\frac{1-e^{-4mT}+4mT}{8mN_e}} \\ &\approx 1 - 1 + \frac{1 - e^{-4mT} + 4mT}{8mN_e} \\ &= \frac{1 - e^{-4mT} + 4mT}{8mN_e} \end{aligned} \tag{4.21}$$

Whose derivative gives the approximate pdf

$$p(t|m, N_e) \approx \frac{1 + e^{-4mt}}{2N_e} \tag{4.22}$$

Chapter 5

Mutation events and haplotype sharing

In chapters 3 and 4, we have introduced a coalescent-based framework that allows computing several quantities related to haplotype sharing in purportedly unrelated samples as a function of past demographic events. The presented framework was derived using coalescent theory to model the occurrence of recombination events at the boundary of IBD segments. We have thus far entirely neglected the occurrence of mutation in the described ancestral processes, but as we shall describe in this chapter, the joint consideration of IBD sharing and mutations may support additional analyses. As mentioned in the Introduction, "identical-by-descent" segments may not be strictly identical. They may in fact harbor mutations that occurred along ancestral lineages connecting a pair of individuals to their common ancestor. We here describe the distribution of these mutations on IBD segments, and discuss applications that make use of this information. We limit derivations to the case of Wright-Fisher populations. The extension to arbitrary coalescent distributions is similar to previous chapters.

5.1 The number of mutations on an IBD segment

Suppose a common ancestor that lived t generations (in continuous time) in the past transmits an IBD segment of genetic length l Morgans to a pair of modern day individuals. We assume fixed recombination and mutation rates per nucleotide are provided, so that multiplying genetic length into a constant r returns the number of nucleotides n the segment spans (i.e. $r = \rho^{-1}$). The probability that the IBD segment has length l Morgans and harbors k mutated sites is

$$p(k, l|t, r, \mu) = p(k|l, t, r, \mu) \times p(l|t), \tag{5.1}$$

where μ is the mutation rate per generation, per nucleotide, per individual. As previously described, the distribution of $p(l|t)$ is exponential, with mean $1/(2t)$. Under the infinite sites assumption [Kimura, 1969], the probability that a single site mutates during $2t$ transmission events may be modeled as $1 - e^{-2t\mu}$. Since the value of μ is in the order of 10^{-8} [Roach *et al.*, 2010], the product $2t\mu$ is also expected to be small, and we can use the approximation $1 - e^{-x} \approx x$ for $x \longrightarrow 0$, to rewrite this probability as $2t\mu$. Because each site is modeled as independently mutating, the total number of mutated sites is binomial, with probability $2t\mu$ and $n = lr$ independent attempts. In addition, because n is large and $2t\mu$ is small, we can model the distribution for the number of mutations on the segment as Poisson with mean $2lrt\mu$. It is interesting to note that the number of mutations that are expected to be found on an IBD segment does not depend on how long ago the common ancestor that transmitted the segment has lived. Because the mutation process is independent from the recombination process, the expected number of mutations $\mathrm{E}[k|t]$ for an IBD segment transmitted from an ancestor t generations ago is obtained by taking the number of mutations that is expected for a segment of length $\mathrm{E}[l|t] = 1/(2t)$. Namely

$$\mathrm{E}[k|t] = 2t\mu r \times \frac{1}{2t} = \mu r, \tag{5.2}$$

where the generation cancels out. This occurs because segments transmitted from a common ancestor decrease in length at the same rate as mutations accumulate. Substituting the Poisson and the exponential distributions into Equation 5.1, we obtain

$$p(k, l|t, r, \mu) = \frac{(2lrt\mu)^k e^{-2lrt\mu}}{k!} \times (2te^{-2tl}) \tag{5.3}$$

If the IBD segment is allowed to be of any length, we can integrate l out, to obtain the distribution for the number of segments found on a segment from generation t

$$\begin{aligned} p(k|t,r,\mu) &= \int_0^\infty p(k,l|t,r,\mu)dl \\ &= \int_0^\infty \frac{(2lrt\mu)^k e^{-2lrt\mu}}{k!} \times (2te^{-2tl})dl \\ &= (\mu r)^k (1+\mu r)^{-(k+1)} \ , \end{aligned} \tag{5.4}$$

which has expectation and variance given by

$$\begin{aligned} \mathrm{E}[k|t,r,\mu] &= \sum_{k=0}^{\infty} k(\mu r)^k (1+\mu r)^{-(k+1)} = \mu r \\ \mathrm{Var}[k|t,r,\mu] &= \sum_{k=0}^{\infty} (k-\mu r)^2 (\mu r)^k (1+\mu r)^{-(k+1)} = \mu r(1+\mu r) \end{aligned} \tag{5.5}$$

Note that this is a Negative Binomial distribution

$$\mathrm{NB}(k,r,\theta) = \frac{\Gamma(r+k)}{k!\,\Gamma(r)}\,\theta^k(1-\theta)^r \ , \tag{5.6}$$

where $r = 1$ and $\theta = (\mu r)/(1+\mu r)$ (or, equivalently, a Geometric distribution with $p = 1-(\mu r)/(1+\mu r)$, since $r = 1$). This distribution occurs as a result of the Gamma-Poisson mixture described in Equation 5.5, where the rate of the Poisson distribution is a random variable itself. In particular, it is an exponential random variable, i.e. it is Gamma distributed with shape parameter $k = 1$ and scale parameter $1/(2t)$. Again, note that the distribution for the number of mutations is independent from when the transmitting common ancestor has lived, therefore this result extends to arbitrary demographic histories, as the coalescent distribution becomes irrelevant.

One can now ask what is the distribution for the number of mutations when the IBD segment can only be longer than a detectable length threshold u. Recall that an exponentially distributed random variable L enjoys the *memorylessness* property, i.e. $p(L > s+t \mid L > s) = p(L > t)$ for all $s, t \geq 0$. This implies that if a segment has length distributed as a truncated exponential random variable, we can express the distribution of its length as a constant segment of length u, plus a remaining part that is itself exponentially distributed with parameter $1/(2t)$, i.e. $L = u + L_\tau$, where the tip of the segment, L_τ, has exponential distribution $p(l_\tau|t) = 2te^{-2tl}$.

We have already computed the distribution for the number of mutations on L_τ in Equation 5.4, and noted it does not depend on t. The distribution of the number of mutations on the fixed part of length u, however, does depend on t, and can be modeled as a Poisson with mean $\lambda = 2t\mu ru$, as previously motivated. The distribution for the number of mutations on the entire segment L, therefore, can be expressed as the discrete convolution of this Poisson distribution and the Negative Binomial (or Geometric) distribution of Equation 5.4, with expectation and variance given by

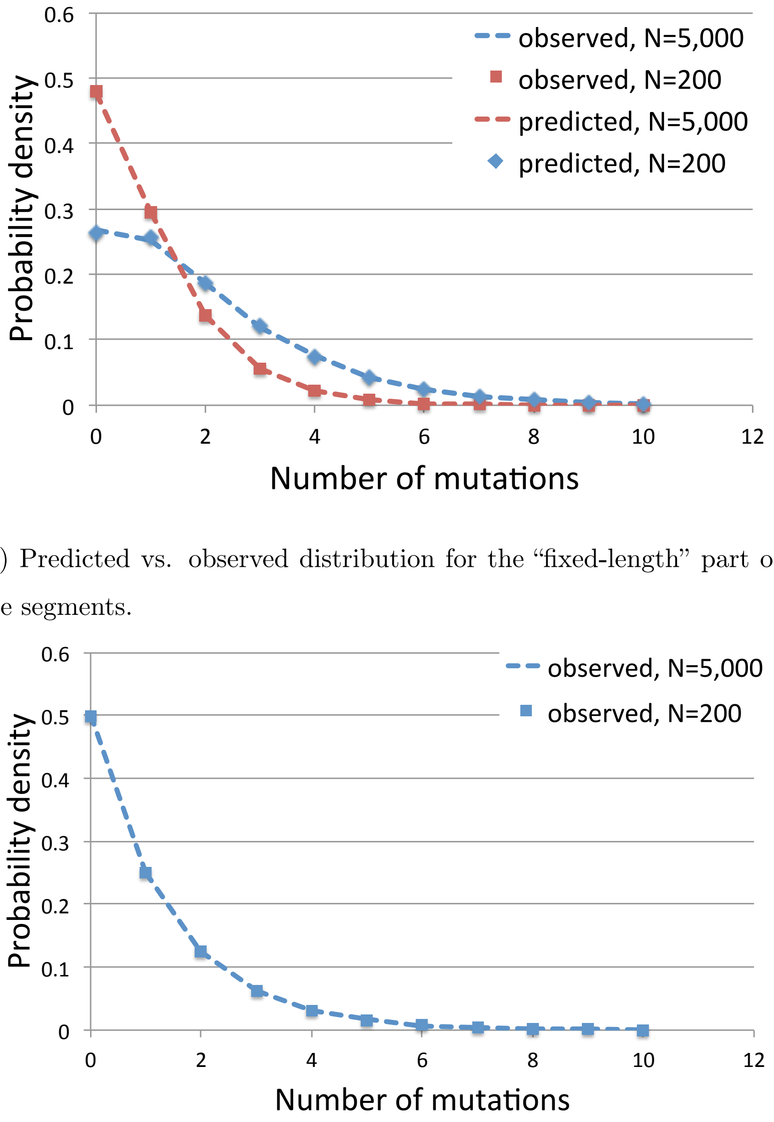
$$\begin{aligned} \mathrm{E}[k|t,r,\mu,u] &= \mu r(2tu+1) \\ \mathrm{Var}[k|t,r,\mu,u] &= \mu r(1+\mu r+2tu) \end{aligned} \tag{5.7}$$

We have tested these models on empirical distributions obtained via simulations, and observed a good fit (Figure 5.1).

5.2 Mutations on IBD segments and demographic history

5.2.1 The age of a randomly drawn IBD segment

We now want to move on to computing the distribution for the number of mutations found on an IBD segment of a given minimum length coming from a population of specific demographic history, rather than the exact generation of the common ancestor. We will need to express the distribution for the age of a randomly sampled IBD segment of length l, given the population has constant size N (simplifying the previous notation, where a contant population size was indicated as N_e). This quantity can be computed using the results of Chapter 3. Recall that for a pair of individuals, given a common ancestor that lived t generations in the past, the probability a genomic site is spanned by a segment of length l is the Erlang-2 distribution with parameter $(2t)^{-1}$, multiplied by the chance of coalescence at generation t, which is $p(t|N) = N^{-1}e^{-tN^{-1}}$. We can divide this expression by l, switching our unit from genomic site to whole segment, and then divide by a normalizing factor to obtain the segment's age



(a) Predicted vs. observed distribution for the “fixed-length” part of the segments.

(b) Observed distribution for the “variable-length” part of the segments, which does not depend on N.

Figure 5.1: Comparison of empirical and analytical distributions for the number of mutations in the “fixed-" and “variable-length” parts of IBD segments, using SMC simulations.

distribution as

$$\begin{aligned} p(t|l,N) &= \frac{4t^2 l e^{-2tl} \times l^{-1} \times N^{-1} e^{-tN^{-1}}}{\int_0^\infty 4t^2 l e^{-2tl} \times l^{-1} \times N^{-1} e^{-tN^{-1}} dt} \\ &= \frac{t^2(2lN+1)^3 e^{-t\left(N^{-1}+2l\right)}}{2N^3} \\ &= t\left(N^{-1}+2l\right)^2 e^{-t(N^{-1}+2l)} \times \left(N^{-1}+2l\right)\frac{t}{2} \end{aligned} \tag{5.8}$$

Note that this is closely related to computing the expected number of segments of length l for a population of size N, as the expected number of segments of length l is obtained by summing the contributions of each generation in the considered demographic history, which for a constant population size is given by $\gamma \times p(t|N)p(l|t)/l$ (see Equation 3.18, which is averaged over all segments of length at least u). This approach was used in [Ralph and Coop, 2013] in the more general context of an arbitrary demographic history, although with a different computation for the expected number of IBD segments.

To get the age of a segment of length greater than u we marginalize the length of the IBD segment for the given demographic history, using Equation 3.14, normalized in the interval $[u, \infty)$, obtaining

$$\begin{aligned} p(t|l \geq u, N) &= \int_u^\infty \frac{t^2(2lN+1)^3 e^{-t\left(N^{-1}+2l\right)}}{2N^3} \times \frac{4N_e(2N_e u+1)^2}{(2lN_e+1)^3} dl \\ &= t \times e^{-t(N^{-1}+2u)} \left(N^{-1}+2u\right)^2, \end{aligned} \tag{5.9}$$

Note that this is again the Erlang-2 distribution, with parameter $N^{-1}+2u$, and that $\lim_{u\to 0} p(t|l \geq u, N) = N^{-2} t e^{-tN^{-1}}$.

5.2.2 The number of mutations on an IBD segment with minimum length

To compute the distribution of mutations on a segment of minimum detectable length u in the population, we again separate the contributions of the segment into its two deterministic and stochastic parts, now marginalizing the time t of the transmitting common ancestor for the portion of fixed length u, using Equation 5.9

$$\begin{aligned}
p(k|u,r,\mu,N) &= \int_0^\infty p(k,t|u,r,\mu,N)dt \\
&= \int_0^\infty p(k|t,u,r,\mu,N)p(t|l \geq u,N)dt \\
&= \int_0^\infty \frac{(2t\mu ru)^k e^{-2t\mu ru}}{k!} te^{-t(N^{-1}+2u)} \left(N^{-1}+2u\right)^2 dt \\
&= \frac{2^k(1+k)(\mu Nru)^k(1+2Nu)^2}{[1+2N(u+\mu ru)]^{k+2}}
\end{aligned} \tag{5.10}$$

The mean for the number of mutations in this portion of the segment is

$$\begin{aligned}
\mathrm{E}[k|t,r,\mu,l=u,N] &= \sum_{k=0}^{\infty} k \times \frac{2^k(1+k)(\mu Nru)^k(1+2Nu)^2}{[1+2N(u+\mu ru)]^{k+2}} \\
&= \frac{4\mu Nru}{1+2Nu},
\end{aligned} \tag{5.11}$$

and the variance is

$$\begin{aligned}
\mathrm{Var}[k|t,r,\mu,l=u,N] &= \sum_{k=0}^{\infty} \left(k - \frac{4\mu Nru}{1+2Nu}\right)^2 \times \frac{2^k(1+k)(\mu Nru)^k(1+2Nu)^2}{[1+2N(u+\mu ru)]^{k+2}} \\
&= \frac{4\mu Nru[1+2N(u+\mu ru)]}{(1+2Nu)^2}.
\end{aligned} \tag{5.12}$$

The remaining part of the segment, the "tip", has variable length, but as we have seen the number of mutations it harbors (Equation 5.4) does not depend on the time t of the transmitting common ancestor, and marginalizing it does not affect the resulting distribution. Putting together the "tip" distribution of Equation 5.4 with the distribution for the fixed part of the segment (Equation 5.10), we get

$$\begin{aligned}
p(k|u,r,\mu,N) &= p(k|l=u,r,\mu,N) * p(k|l>0,r,\mu) \\
&= \sum_{n=-\infty}^{\infty} \chi_a(k)\frac{2^k(1+k)(\mu Nru)^k(1+2Nu)^2}{[1+2N(u+\mu ru)]^{k+2}} \times \chi_0(n-k)\frac{(\mu r)^{n-k}}{(1+\mu r)^{n-k+1}} \\
&= (2Nu+1)^2 \left\{ \frac{(\mu r)^n}{(\mu r+1)^{n+1}} - \frac{2^{n+1}Nu(\mu Nru)^n[2N(\mu ru+u)+n+2]}{[2N(\mu ru+u)+1]^{n+2}} \right\} ,
\end{aligned} \tag{5.13}$$

where $\chi_a(x)$ is the step function

$$\chi_a(x) = \begin{cases} 1 & \text{if } x \geq a, \\ 0 & \text{if } x < a. \end{cases} \tag{5.14}$$

The mean and the variance of this distribution can be obtained summing mean and variance of the two components of the final distribution, as

$$\begin{aligned} \mathrm{E}[k|u, r, \mu, N] &= \mu r + \frac{4\mu N r u}{1 + 2Nu} \\ \mathrm{Var}[k|u, r, \mu, N] &= \mu r(1 + \mu r) + \frac{4\mu N r u[1 + 2N(u + \mu r u)]}{(1 + 2Nu)^2}. \end{aligned} \tag{5.15}$$

The obtained distribution for the number of mutations on an IBD segment was found to provide a good fit for empirical distributions obtained from simulations (Figure 5.2). Note that the distribution is overdispersed, as the variance is always larger than the mean.

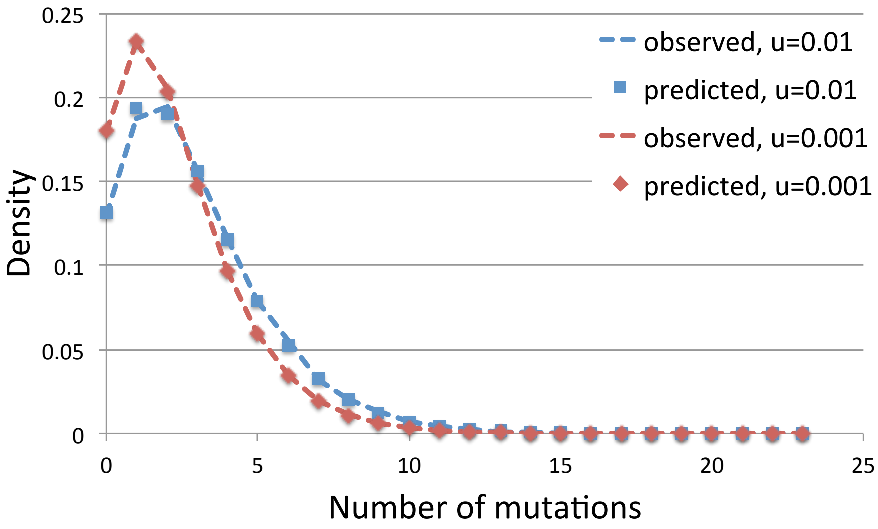


Figure 5.2: Empirical and analytical distributions of mutation counts on IBD segments for $N = 500$, using SMC simulations.

5.3 Using mutations on IBD segments for demographic inference

We now turn to the problem of using knowledge on the number of mutations on IBD segments to infer demographic history. We consider the number of segments s_u that are longer than a Morgan threshold u, and the number k of mutations found on these segments in a population of size N. Again, we can write

$$p(s_u, k|N)dl = p(k, |s_u, N)p(s_u|N) \tag{5.16}$$

We have previously computed an expression for $p(s_u|N)$, which is Poisson distributed with mean described in Equation 3.24, and Equation 5.8 provides the distribution for the number of mutations coming from a single segment sampled from the population. To compute the distribution for an independent number of segments, we can repeatedly take the convolution of Equation 5.11. Note, however, that this is computationally unfeasible for large s_u, and it is also unnecessary, as we can rely on the Central Limit Theorem and use the mean and variance of Equation 5.15 in a Normal distribution. Note, furthermore, that the shape of repeated convolutions of the distribution in Equation 5.11 quickly converges to a Gaussian distribution (Figure 5.3).

After substituting the previously discussed distributions into Equation 5.16, we tested the prediction of this expression against several synthetic datasets (Figure 5.6), and we observed good correspondence between analytical and empirical distributions (note that some of the approximations done in the GENOME simulator software, e.g. the minimum size of recombinant blocks, may distort the empirical distribution).

Compared to the distribution for the number of IBD segments (Figure 5.4), the distributions for the number of mutations on IBD segments overlap substantially across different values of the effective population size N (Figure 5.5), suggesting that knowledge about the distribution of mutations over IBD segments does not result in substantial improvements for demographic inference. Inspecting the expression for the mean number of mutations in an IBD segment in Equation 5.15, it is clear that variations of N, the effective population size, result in minimal variation, as can be observed in Figure 5.7.

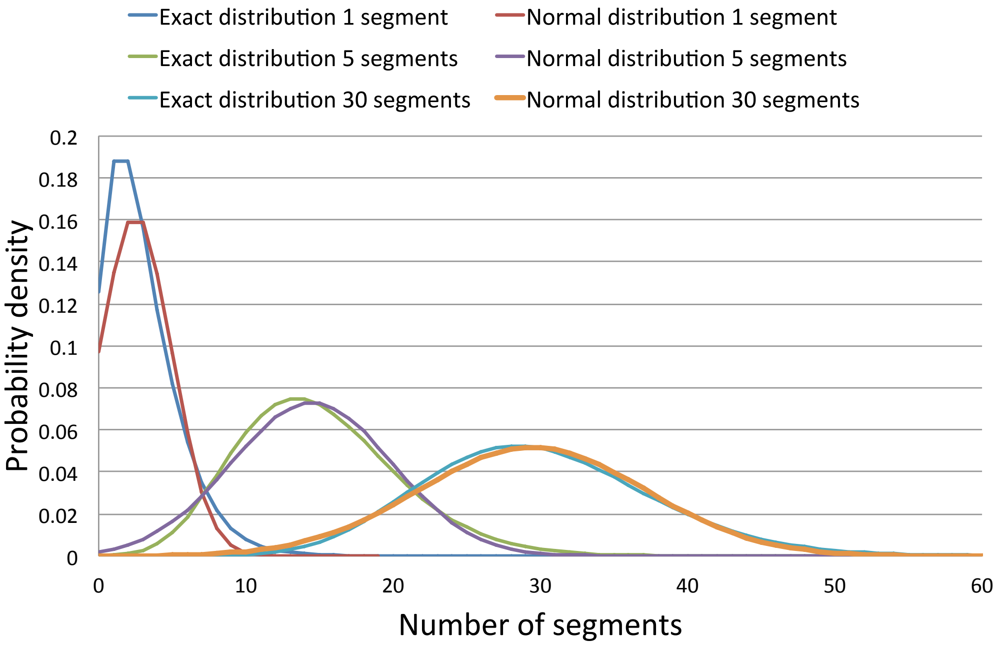


Figure 5.3: Convergence of the distribution for the number of mutations to a Normal distribution.

5.4 Inferring mutation rates using IBD

The mean and variance of Equation 5.15 show a weak connection between demography and mutations in IBD segments. While this does not facilitate using this feature for demographic inference, it does provide support for other analyses. The mean and variance are in fact strongly influenced by the mutation rate (Figure 5.8), suggesting this parameter may be inferred based on the derived distributions of mutation on IBD segments. Furthermore, we note that it is possible to entirely remove the dependence on demographic history through very simple manipulations of the observed IBD segments. We have previously noted that due to the memorylessness property of the exponential distribution, the length of a detected IBD segment longer than a detectable threshold u may be seen as the sum of two parts: $L = u + L_\tau$, the fixed length, determined by the threshold of detectable IBD segments, and a stochastic part, L_τ. As previously described, the latter is not affected by coalescent

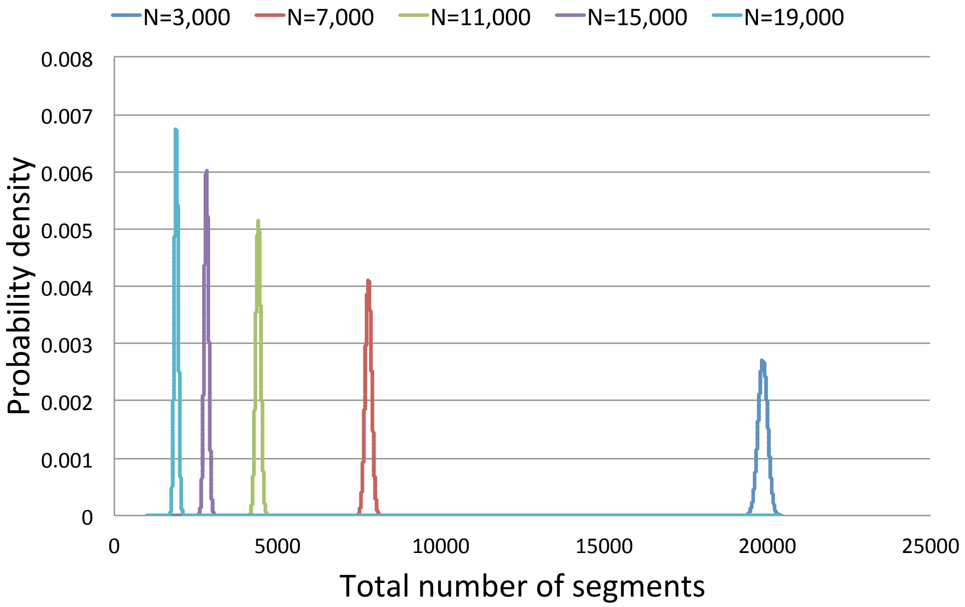


Figure 5.4: The distribution of the number of IBD segments does not overlap across different populations (20 samples, segments of at least 0.5cM in SMC simulations).

times, while the “fixed” portion of the segment is weakly influenced by population size. This property can be exploited to isolate summary statistics that are informative about mutation rates, while not being confounded by the presence of latent demographic history. Recall that the sum of two Poisson random variables is also Poisson distributed, with rate given by the sum of the addends’ rates. The number of mutations on the segment $L = u + L_\tau$ that are found on the L_τ portion are therefore expected to be a fraction L_τ/L of those observed on the entire segment. To estimate the total number of such mutations, removing the small bias introduced by the demographic history, it is therefore sufficient to multiply the number of mutations observed on each segment of the detected set S_u by the factor $(l - u)/l$.

$$\hat{k} = \sum_{s \in S_u} k_s \times [(l_s - u)/l_s] \tag{5.17}$$

As shown in Equation 5.5, the expected number of mutations occurring in the stochastic portion of each segment is simply μr. Because the distribution for the number of mutations

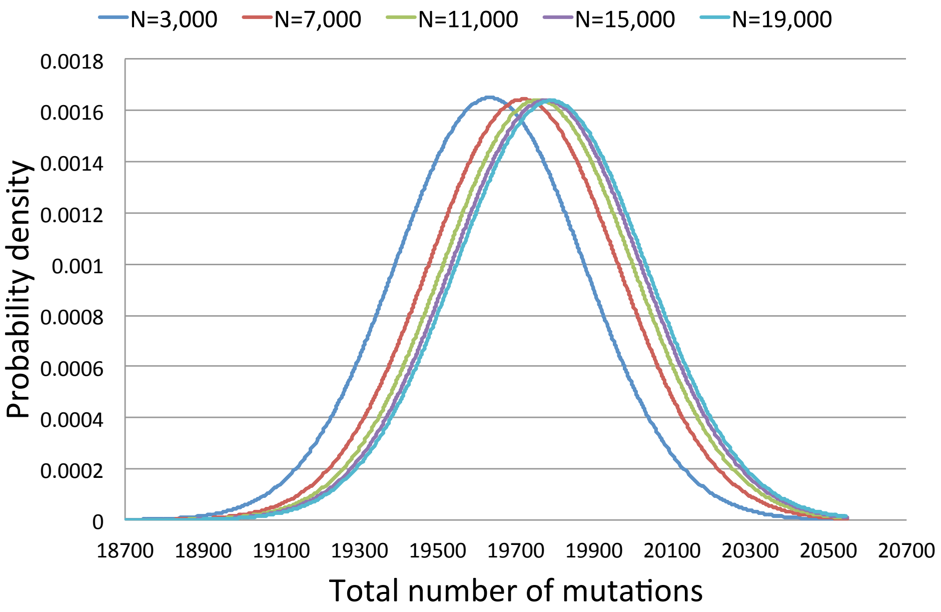
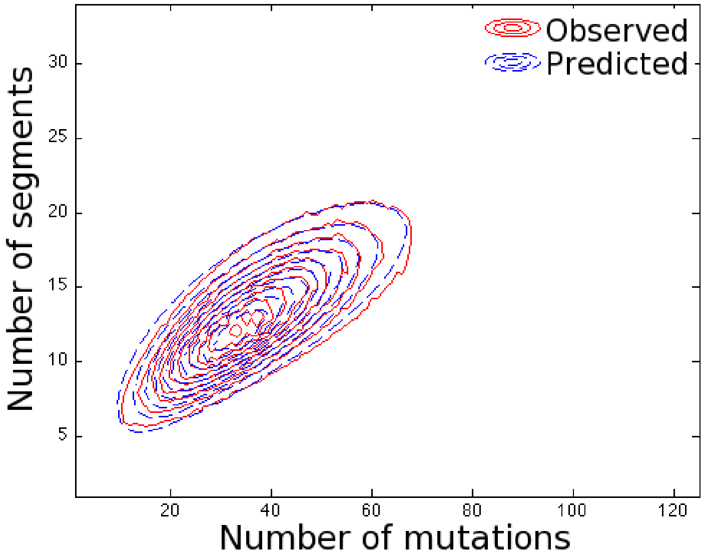


Figure 5.5: Substantial overlap for the distribution of the total number of mutations in different populations (20 samples, segments of at least 0.5cM in SMC simulations).

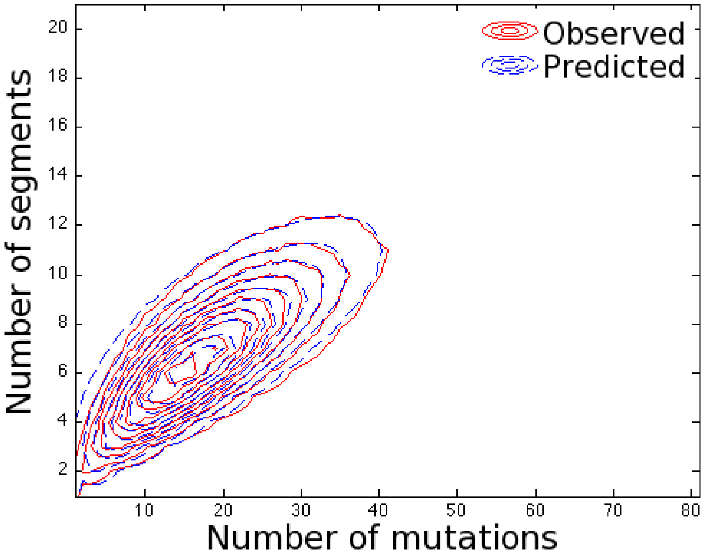
found on a large number of segments $n_s = |S_u|$ is well described by a Normal distribution with mean $n_s \mu r$, we can derive the maximum likelihood estimator

$$\hat{\mu} = \frac{\hat{k}}{r n_s}, \tag{5.18}$$

which is unbiased under the discussed assumptions. If we denote the set of segments coming from the discrete generation g as S_{ug}, then



(a) $N = 2,000$, $\mu = 10^{-8}$, $r = 10^{8}$, $u = 0.01$, genomic region of 10M.



(b) $N = 4,000$, $\mu = 10^{-8}$, $r = 10^{8}$, $u = 0.01$, genomic region of 10M.

Figure 5.6: Comparison of joint distribution for the number of segments and the number of mutations and empirical distribution obtained from GENOME simulations.

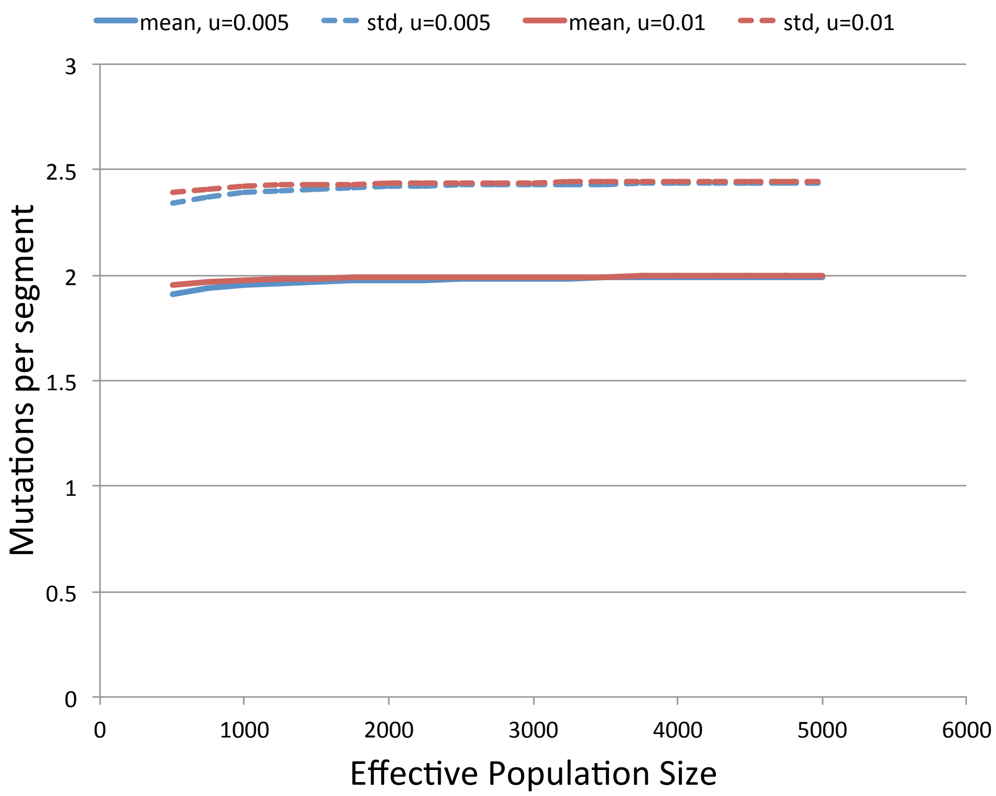


Figure 5.7: Mean and standard deviation (Equation 5.15) for several values of N.

$$\begin{aligned}
\mathrm{E}[\hat{\mu}] &= \mathrm{E}\left[\frac{\hat{k}}{rn_s}\right] = \frac{1}{rn_s}\,\mathrm{E}\left[\hat{k}\right] \\
&= \frac{1}{rn_s}\sum_{g=1}^{\infty}\sum_{s\in S_{ug}} \mathrm{E}\left[\frac{l_s-u}{l_s}\times k_s\right] \\
&= \frac{1}{rn_s}\sum_{g=1}^{\infty}\sum_{s\in S_{ug}} \int_u^{\infty} P(l|g)\,\mathrm{E}\left[\frac{l-u}{l}\times k\middle| l\right] dl \\
&= \frac{1}{rn_s}\sum_{g=1}^{\infty}\sum_{s\in S_{ug}} \int_u^{\infty} 2ge^{-2g(l-u)}\left(\frac{l-u}{l}\times 2g\mu lr\right) dl \\
&= \frac{1}{rn_s}\sum_{g=1}^{\infty}\sum_{s\in S_{ug}} \mu r = \frac{1}{rn_s}\times n_s\mu r = \mu
\end{aligned} \tag{5.19}$$

We tested this estimator on simulated data using the SMC algorithm [McVean and

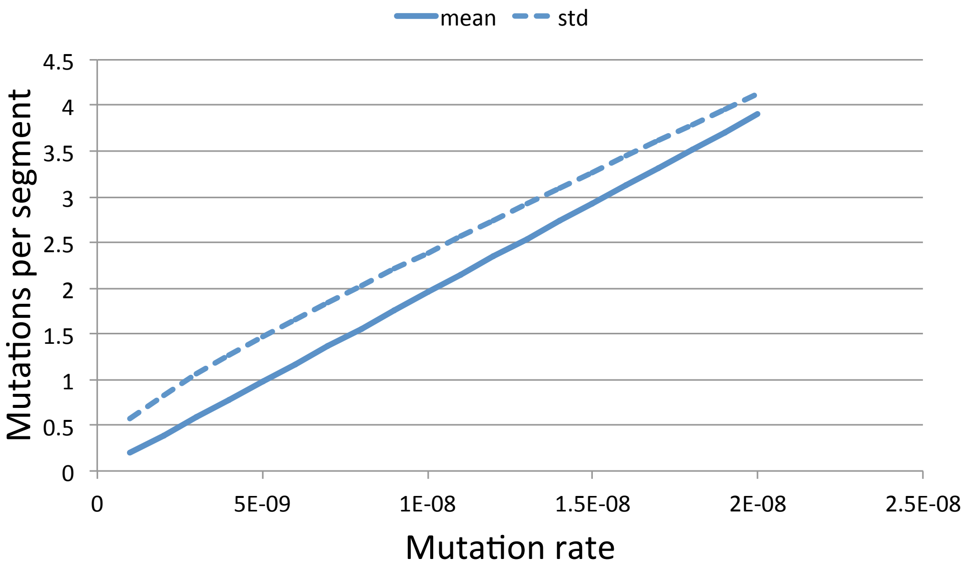
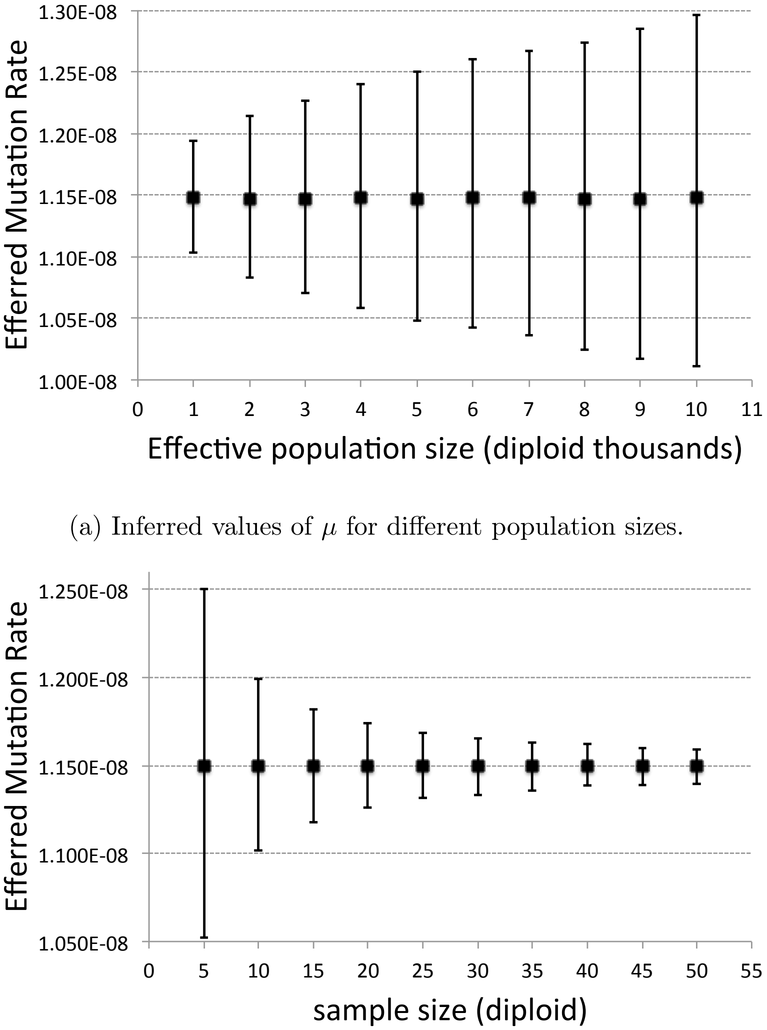


Figure 5.8: Number of mutations per IBD segment as a function of mutation rate.

Cardin, 2005], therefore using definition (c) of Section 1.1.4.1 to generate shared segments, and we observed good performance. By analyzing IBD segments of length at least 1 cM for only 5 diploid samples in a population of $10,000$ haploid individuals, assuming a genome of 35 Morgans and $\mu = 1.15$, $r = 0.83$, we attained tight 95% confidence intervals around the true mutation rate (Figure 5.9a). Variations in the effective population size had a moderate influence on the width of confidence intervals, as larger population sizes result in fewer IBD segments. Because a sample of n individuals contains $\binom{2n}{2}$ haploid chromosomes that may contain IBD segments, there is a quadratic gain of statistical power when more samples are analyzed (Figure 5.9b). A sample of 50 independent individuals provide extremely tight confidence intervals in a population of $N = 10,000$ effective diploid individuals, using the above listed parameters.

Back-of-the-envelope calculations suggest that the statistical power of IBD-based inference of mutation rates is extremely high compared to a trio-based analysis. When mutation rates are estimated based on observed de-novo mutations in transmitted haplotypes of sequenced trio individuals, the number of “effective haplotypes” that are compared is $n/3$, for n sequenced individuals, i.e. the two transmitted haplotypes. In a population of diploid



(a) Inferred values of μ for different population sizes.

(b) Inferred mutation rate as a function of sample size.

Figure 5.9: Inference of mutation rates via IBD sharing in SMC simulations.

effective size $10,000$ ($20,000$ haploid individuals), pairs of samples will share on average $(4Nu+1)/(2Nu+1)^2 \approx 1/(Nu) = 0.5\%$ of their genome through IBD haplotypes longer than 1cM. The number of pairs tested for IBD are however $\binom{2n}{2}$, as previously mentioned, resulting in $\binom{2n}{2} \times 0.05$ effective comparisons across haplotypes, assuming that ancestral lineages do not overlap significantly. In this scenario, for a sample of 60 individuals, about $\binom{120}{2} \times 0.005 \approx 36$ haploid genomes are effectively compared to estimate mutation rates using IBD, against the 40 used in the trio-based approach. However, the number of mutations found on a nucleotide spanned by these IBD segments is higher than the average number of mutations for a nucleotide transmitted by a parent in a trio-based analysis. The number of meioses (generations) separating two IBD individuals from their common ancestor trasmitting a segment of at least u Morgans is on average $2/(N^{-1}+2u) \approx 1/u$ (Equation 5.9). The distance, in meioses, between the two individuals is therefore on average $2/u$, or 200 generations for $u = 1$cM. This results in roughly $\frac{36\times200}{40} = 180$ times the statistical power of a trio-based analysis. If more isolated populations are considered, and shorter segments can be detected, this gain is further increased. For a population of effective size $N = 1,000$, the fraction of genome shared for segments of at least 1 cM is about 5%, with a gain of $\frac{\binom{120}{2}\times0.05\times200}{40} = 1,785$, and if segments of 0.5 cM and longer can be detected, the gain would reach roughly $3,570$.

Such increase in statistical power enables further analyses, such as the inference of locus-specific mutation rates. Note, however, that some of the assumptions that were made in this derivation, such as the uniformity of recombination rates and selective neutrality, will need to be addressed. Furthermore, this analysis relies on accurate IBD detection, and IBD segments are defined as non-recombinant chromosomal regions transmitted from a common ancestor (see definition (c) of Section 1.1.4.1). Refinements of these models to account for mismatches between the described quantities and what is realistically possible to infer using available IBD detection algorithms will improve this analysis. This approach also assumes lack of genotyping errors, however note that because IBD detection is feasible using only high-frequency markers, typically less prone to genotype errors, there is no substantial interference in using shared haplotypes in this scenario. A related application is the fine-tuning of genotype-calling parameters, as prior knowledge on a plausible range for mutation rates enables detecting a component of error in the inferred mutation parameter. Furthermore, note that assuming the mutation rate is known, this approach may be used to study recombination rates.

It is interesting to note that this approach measures mutation rates observing mutation events that potentially occurred several generations in the past. This method can therefore be used to asses variation in historical mutation rates. A simple test involves obtaining estimates based on different cutoffs for the minimum length of the observed IBD segments. Because shorter haplotypes tend to be transmitted from more remote common ancestors, estimates based on smaller values of u reflect mutation rates at more remote time scales in the population. If IBD detection is accurate and a demographic model has been reconstructed (using the methods of Chapter 3 and 4, or others that more suitable for remote time scales e.g. [Li and Durbin, 2011]), it is possible to estimate the time of these variations using the described methods in conjunction with the segment age distributions of Equation 5.9.

5.5 Appendix

5.5.1 Efficient computation of mean and variance for the number of mutations in an arbitrary demography

If a population of arbitrary demographic history θ has population size $N(g, \theta)$ at discrete time g, the coalescent distribution is described by

$$\left(\prod_{j=1}^{g-1} 1 - \frac{1}{N(j,\theta)} \right) \frac{1}{N(g,\theta)} = C(g,\theta). \tag{5.20}$$

The probability of seeing k mutations on an IBD segment in this population is

$$p(k|u,r,\mu,\theta) = \sum_{g=1}^{\infty} \left[C(g,\theta) \, \text{Poiss}(k, 2\mu r u g) \right]. \tag{5.21}$$

For the mean of the constant part:

$$
\begin{aligned}
\mathrm{E}[k|u,r,\mu,\theta] &= \sum_{k=0}^{\infty}\{k\ p(k|u,r,\mu,\theta)\} \\
&= \sum_{g=1}^{\infty}\left\{\sum_{k=0}^{\infty}[k\ C(g,\theta)\,\mathrm{Poiss}(k,2\mu rug)]\right\} \\
&= \sum_{g=1}^{\infty}\left\{C(g,\theta)\sum_{k=0}^{\infty}[k\,\mathrm{Poiss}(k,2\mu rug)]\right\} \\
&= \sum_{g=1}^{\infty}\{2\mu rug\ C(g,\theta)\}\,.
\end{aligned}
\tag{5.22}
$$

For the variance of the constant part (using $\mathrm{E}_K = \mathrm{E}[k|u,r,\mu,\theta]$):

$$
\begin{aligned}
\mathrm{Var}[k|u,r,\mu,\theta] &= \sum_{k=0}^{\infty}\left\{(k-\mathrm{E}_K)^2\ p(k|u,r,\mu,\theta)\right\} \\
&= \sum_{k=0}^{\infty}\left\{(k-\mathrm{E}_K)^2\sum_{g=1}^{\infty}[C(g,\theta)\,\mathrm{Poiss}(k,2\mu rug)]\right\} \\
&= \sum_{g=1}^{\infty}\left\{C(g,\theta)\sum_{k=0}^{\infty}\left[(k-\mathrm{E}_K)^2\,\mathrm{Poiss}(k,2\mu rug)\right]\right\} \\
&= \sum_{g=1}^{\infty}\left\{\left[(2\mu rug-\mathrm{E}_K)^2+2\mu rug\right]C(g,\theta)\right\}.
\end{aligned}
\tag{5.23}
$$

Where the last step is obtained considering that, for any random variable K with discrete probability distribution $f(k)$, having defined $b = \mathrm{E}_K - c$, where E_K is the expectation of the distribution, then

$$\begin{aligned}
\sum_{k=-\infty}^{\infty} \big[(k-c)^2 f(k)\big] &= \sum_{k=-\infty}^{\infty} \big[(k-\mathrm{E}_K + b)^2 f(k)\big] \\
&= \sum_{k=-\infty}^{\infty} \big\{\big[(b^2 - 2b\,\mathrm{E}_K + 2bk) + (k-\mathrm{E}_K)^2\big]\, f(k)\big\} \\
&= \sum_{k=-\infty}^{\infty} \big\{\big[b^2 - 2b\,\mathrm{E}_K + 2bk\big]\ f(k)\big\} + \sum_{k=-\infty}^{\infty} \big\{(k-\mathrm{E}_K)^2 f(k)\big\} \\
&= b^2 - 2b\,\mathrm{E}_K + 2b \sum_{k=-\infty}^{\infty} [k f(k)] + \sum_{k=-\infty}^{\infty} \big\{(k-\mathrm{E}_K)^2 f(k)\big\} \\
&= b^2 + \mathrm{Var}_K = (\mathrm{E}_K - c)^2 + \mathrm{Var}_K \,.
\end{aligned} \tag{5.24}$$

Chapter 6

Conclusions

In this thesis we presented several new methodologies for population genetics analysis based on the sharing of IBD haplotypes across purportedly unrelated individuals from one or multiple populations. Specifically,

- In Chapter 2, by analyzing several world-wide populations from the HapMap 3 dataset and the Jewish Hapmap dataset, we demonstrated that IBD sharing is informative about demographic events, revealing past migrations and population size fluctuations, carries the signature of evolutionary events, demonstrating enrichment of loci under positive selection for commonly shared regions, and reflects recent stratification, in some cases more accurately than standard methodologies.

- In Chapter 3, motivated by the results of Chapter 2, we used coalescent theory (see Chapter 1) to derive a theoretical framework that allows quantitatively describing the relationship between IBD sharing and demography. We demonstrated these methods by inferring the occurrence of recent demographic events in two populations of distinctive recent demographic profiles: Ashkenazi Jews, exhibiting evidence for a recent founder event followed by substantial expansion and isolation, and the Maasai from Kenya, where haplotype sharing is compatible with a societal structure of several small demes interacting through high migration rates.

- In Chapter 4, we extended this framework to enable simultaneous analysis of several populations, inferring both migration and population size fluctuation. We showed this approach can be used for the analysis of recently diverged populations, where state-of-the-art methods based on ancestry deconvolution using a panel of reference ancestral populations are complicated due to the limiting assumption of strongly diverged ancestral groups. We used these models to study recent demographic history in the Netherlands, showing that IBD-based analysis reveals demographic structure even at fine-grained geographic scales.

- Chapter 5 discussed utilizing IBD segments in studies where whole sequence information is available. We derived distributions for the number of mutated sites on shared haplotypes, and shown that while this information does not provide substantial improvements for demographic inference, it can be used for inferring additional parameters, such as mutation rate, increasing the statistical power by orders of magnitude compared to classical family-based methods. This boost in statistical power enables further applications, such as the inference of a map of locus-specific mutation rates, studying recombination rates, and studying historical variations of these quantities.

During the development of this work, several other methodologies related to demographic inference were published. We here provide a brief overview of their advantages and limitations.

Methods for demographic inference available when this thesis work begun (reviewed in [Pool *et al.*, 2010]) often relied on the simplifying but limiting assumption of unlinked genetic markers, or modeled the linkage induced by the lack of historical recombination using measures of local correlation such as linkage disequilibrium. In the following years, sustained by the increasingly dense genomic datasets, several haplotype-based methods were proposed. In [Pool and Nielsen, 2009], subsequently extended by [Gravel, 2012], an approach based on the frequency and length of migrant tracts was proposed for the inference of migration rates. While effectively recovering migration rates in several demographic scenarios, however, these methods do not model population size fluctuations, and are dependent on the possibility of reliably performing ancestry deconvolution to assign chromosomal tracts to a set of reference populations. These populations may not be available and, more importantly, need to be substantially divergent to attain high-quality deconvolution, as shown in our analysis. Although whole sequence datasets and methodological developments may improve the performance of deconvolution methods, this limitation may prevent methods based on migrant tracts from being effectively employed in the reconstruction of fine-scale migration patterns of the recent millennia. Methods based on ancestry deconvolution, however, may in some scenarios be used in concert with methods based on IBD sharing. Knowing whether an IBD tract was co-inherited from a specific population, in fact, may provide information on the directionality of migration, and also offer further insight into deeper time scales, as shown in [Velez *et al.*, 2012] and [Campbell *et al.*, 2012] and discussed in Chapter 2. These methods may be further explored in light of the recently developed analytical model for migrant tracts and the presented model for IBD. While the methods based on IBD detection, presented in this work, provide some advantages over ancestry deconvolution, it should be noted that these involve dealing with increasingly complex demographic models, and because the ancestors of IBD segments tend to be more remote than those of migrant tracts, larger sample sizes may be required for this analysis.

As mentioned in the Introduction (Section 1.1.3.4), a recently developed Markovian approximation of the coalescent process (Sequentially Markovian Coalescent, SMC, [McVean and Cardin, 2005]) resulted in the development of several new population genetics methods, including methods for inferring demographic history. We note that because we relied on definition (c) of Section 1.1.4.1, the methods described in this thesis are also intrinsically linked to the SMC framework, as IBD segments are defined as being delimited by any recombination event. Future developments include dealing with the potential discrepancies of this definition and the output of IBD detection algorithms, potentially incorporating these and other calculations in new methods for IBD discovery. In [Li and Durbin, 2011], a method based on the SMC model studied population size fluctuations affecting human populations following the out-of-Africa migrations. This approach was limited to pairs of phased whole-sequence haploid individuals from different populations, or single individuals from a population, resulting in limited power for the inference of recent demographic events. Recently, these methods have been extended to allow analyzing multiple phased individuals simultaneously, using a composite likelihood approach [Steinrücken *et al.*, 2012; Sheehan *et al.*, 2013], thus gaining insight into more recent demographic events. Using similar techniques, a recently published pre-print relies on approximated extensions of the SMC to the analysis of multiple individuals, providing insight into the ancestral recombination graph of groups of individuals [Rasmussen and Siepel, 2013]. These methods are extremely appealing, as they make use of almost all available genomic information, but may be suffering from computational limitations. Exploiting the Markovian properties of the SMC model typically leads to algorithms that scan all s sites for all pairs of n samples resulting in at least $O(n^2s)$ complexity. When the analysis of thousands of samples across millions of markers is required, these methods scale poorly compared to methods that rely on summary statistics that can be obtained through methods that are sub-quadratic in the number of samples and may not require analyzing all available genomic markers. A method recently proposed in [Harris and Nielsen, 2013] relies on summary statistics of IBS tract length, decreasing the computational burden but still requiring full pair-wise analysis of sequences to extract summary statistics, also requiring $O(n^2s)$ computation. This method models the length of haplotypes shared through very remote common ancestors. At these time scales, very short IBD segments may be joined together, and definition (c) of shared segment used in this work becomes unrealistic (see Section 1.1.4.1). To work with definition (b), where several short contiguous IBD segments from a shared ancestor are transmitted to a pair of individuals, additional modeling was developed. In addition to these methods, work described in [Ralph and Coop, 2013] infers historical demographic changes from length distributions of IBD segments, using the principles described in Chapter 3 within a less parametric approach to the inference of coalescent distributions across different populations, thereby allowing increased flexibility compared to the model-based inference procedure of Chapter 3, but without providing explicit inference of migration and population size changes, described in Chapter 4.

The models that we proposed in this thesis assume selective neutrality. Although the distribution of haplotype sharing is likely to be affected by localized natural selection [Bamshad and Wooding, 2003], the extent to which the human genome has been shaped by selective forces has yet to be quantified [Hernandez *et al.*, 2011]. The proposed model of IBD sharing can be locally used to test deviations from neutrality and can be improved to explicitly handle the presence of selective forces. Further enhancements of the proposed methodology include improved approaches to demographic model optimization and selection, which may lead to automatic clustering of analyzed individuals into subpopulations. Furthermore, as described in Chapter 5 the proposed framework allows studying basic biological parameters such as mutation and recombination rates, potentially providing insights into questions regarding locus- or population-specific differences, and historical variation of these quantities ([Scally and Durbin, 2012; Coop and Przeworski, 2007])

The proposed methodology facilitates tackling questions beyond demographic inference from genotype data; such questions include those that arise when phenotype data are also considered. A problem that has recently received much attention is that of estimating heritability with the use of large samples of unrelated individuals. Haplotype sharing across purportedly unrelated individuals has been used in this context [Zuk *et al.*, 2012; Zaitlen *et al.*, 2013; Price *et al.*, 2011], and the proposed model for IBD sharing across unrelated samples can be used for improving such analysis.

On the applied side, genome-wide association studies have taught us the lesson of needing to know the demographic makeup of a study population. Although linear-trend analysis has been shown to capture population stratification when common genomic variants are considered [Price *et al.*, 2006], methods for association of rare variants are an active field of investigation [Li and Leal, 2008; Madsen and Browning, 2009; Price *et al.*, 2010] in which recent stratification poses new challenges [Mathieson and McVean, 2012]. The reconstruction of a fine-grained picture of population stratification thus gains importance in the context of full sequence data. Stratification might in fact occur at different historical timescales, and statistical indicators designed to account for ancient diversification trends might not reveal signatures of recent demographic events.

The reported analysis of HapMap's MKK samples provides an example of this phenomenon. This sample exhibits high levels of endogamy through ubiquitous shared long-range haplotypes, suggesting a small population size, but it appears to have an outbred profile when the decay of LD is analyzed [McEvoy *et al.*, 2011]. As discussed in Chapter 3, a plausible reason for the observed data might in this case be found in the societal structure of the MKK people. We hypothesize that this "village effect" will be established in other modern populations that are commonly considered outbred on the basis of their ancient-timescale characteristics. Several genetic surveys have in fact outlined surprisingly high levels of runs of homozygosity in a number of outbred populations worldwide [Henn *et al.*, 2011; Henn *et al.*, 2011; Henn *et al.*, 2011; Broman and Weber, 1999; Gibson *et al.*, 2006]. When migration events are included in the model, long runs of homozygous haplotypes in otherwise outbred populations are plausibly interpreted as reflecting a genetic pool of several small demes that slowly but constantly intermix. The ability to reconstruct recent demographic events will enable the analysis of these phenomena. Combined with prior knowledge of a population's history, this analysis will provide a useful tool for describing the fine-grained evolutionary context in which recent genetic variation arose.

Bibliography

[Abdellaoui *et al.*, 2013] Abdel Abdellaoui, Jouke-Jan Hottenga, Peter de Knijff, Michel G Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, Erik A Ehli, Yueshan Hu, Gareth E Davies, et al. Population structure, migration, and diversifying selection in the netherlands. *European Journal of Human Genetics*, 2013.

[Akaike, 1974] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[Altshuler *et al.*, 2010] David M Altshuler, Richard A Gibbs, Leena Peltonen, Emmanouil Dermitzakis, Stephen F Schaffner, Fuli Yu, Penelope E Bonnen, PI De Bakker, Panos Deloukas, Stacey B Gabriel, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

[Atzmon *et al.*, 2010] Gil Atzmon, Li Hao, Itsik Pe'er, Christopher Velez, Alexander Pearlman, Pier Francesco Palamara, Bernice Morrow, Eitan Friedman, Carole Oddoux, Edward Burns, et al. Abraham's children in the genome era: major jewish diaspora populations comprise distinct genetic clusters with shared middle eastern ancestry. *The American Journal of Human Genetics*, 86(6):850–859, 2010.

[Bamshad and Wooding, 2003] Michael Bamshad and Stephen P Wooding. Signatures of natural selection in the human genome. *Nature Reviews Genetics*, 4(2):99–111, 2003.

[Barrett *et al.*, 2005] Jeffrey C Barrett, B Fry, JDMJ Maller, and MJ Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.

[Behar *et al.*, 2006] Doron M Behar, Ene Metspalu, Toomas Kivisild, Alessandro Achilli, Yarin Hadid, Shay Tzur, Luisa Pereira, Antonio Amorim, Lluís Quintana-Murci, Kari Majamaa, et al. The matrilineal ancestry of ashkenazi jewry: portrait of a recent founder event. *The American Journal of Human Genetics*, 78(3):487–497, 2006.

[Bonnen *et al.*, 2009] Penelope E Bonnen, Jennifer K Lowe, David M Altshuler, Jan L Breslow, Markus Stoffel, Jeffrey M Friedman, and Itsik Pe'er. European admixture on the micronesian island of kosrae: lessons from complete genetic information. *European Journal of Human Genetics*, 18(3):309–316, 2009.

[Boomsma *et al.*, 2013] Dorret I Boomsma, Cisca Wijmenga, Eline P Slagboom, Morris A Swertz, Lennart C Karssen, Abdel Abdellaoui, Kai Ye, Victor Guryev, Martijn Vermaat, Freerk van Dijk, et al. The genome of the netherlands: design, and project goals. *European Journal of Human Genetics*, 2013.

[Brisbin *et al.*, 2012] Abra Brisbin, Katarzyna Bryc, Jake Byrnes, Fouad Zakharia, Larsson Omberg, Jeremiah Degenhardt, Andrew Reynolds, Harry Ostrer, Jason G Mezey, and Carlos D Bustamante. Pcadmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology*, 84(4):343–364, 2012.

[Broman and Weber, 1999] Karl W Broman and James L Weber. Long homozygous chromosomal segments in reference families from the centre d'etude du polymorphisme humain. *The American Journal of Human Genetics*, 65(6):1493–1500, 1999.

[Browning and Browning, 2007] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[Browning and Browning, 2010] Sharon R Browning and Brian L Browning. High-resolution detection of identity by descent in unrelated individuals. *The American Journal of Human Genetics*, 86(4):526–539, 2010.

[Browning and Browning, 2011a] Brian L Browning and Sharon R Browning. A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182, 2011.

[Browning and Browning, 2011b] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[Browning and Browning, 2012] Sharon R Browning and Brian L Browning. Identity by descent between distant relatives: detection and applications. *Annual review of genetics*, 46:617–633, 2012.

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

[Bryc *et al.*, 2010] Katarzyna Bryc, Christopher Velez, Tatiana Karafet, Andres Moreno-Estrada, Andy Reynolds, Adam Auton, Michael Hammer, Carlos D Bustamante, and Harry Ostrer. Genome-wide patterns of population structure and admixture among hispanic/latino populations. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):8954–8961, 2010.

[Campbell *et al.*, 2012] Christopher L Campbell, Pier F Palamara, Maya Dubrovsky, Laura R Botigué, Marc Fellous, Gil Atzmon, Carole Oddoux, Alexander Pearlman, Li Hao, Brenna M Henn, et al. North african jewish and non-jewish populations form distinctive, orthogonal clusters. *Proceedings of the National Academy of Sciences*, 109(34):13865–13870, 2012.

[Carmi *et al.*, 2013] Shai Carmi, Pier Francesco Palamara, Vladimir Vacic, Todd Lencz, Ariel Darvasi, and Itsik PeâĂŹer. The variance of identity-by-descent sharing in the wright–fisher model. *Genetics*, 193(3):911–928, 2013.

[Chapman and Thompson, 2003] NH Chapman and EA Thompson. A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical population biology*, 64(2):141–150, 2003.

[Coast, 2001] Ernestina Coast. *Maasai demography.* PhD thesis, University of London, University College London, 2001.

[Coop and Przeworski, 2007] Graham Coop and Molly Przeworski. An evolutionary view of human recombination. *Nature Reviews Genetics*, 8(1):23–34, 2007.

[de Bakker *et al.*, 2006] Paul IW de Bakker, Gil McVean, Pardis C Sabeti, Marcos M Miretti, Todd Green, Jonathan Marchini, Xiayi Ke, Alienke J Monsuur, Pamela Whittaker, Marcos Delgado, et al. A high-resolution hla and snp haplotype map for disease association studies in the extended human mhc. *Nature genetics*, 38(10):1166–1172, 2006.

[Duerr *et al.*, 2006] Richard H Duerr, Kent D Taylor, Steven R Brant, John D Rioux, Mark S Silverberg, Mark J Daly, A Hillary Steinhart, Clara Abraham, Miguel Regueiro, Anne Griffiths, et al. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *science*, 314(5804):1461–1463, 2006.

[Ewens, 2004] Warren J Ewens. *Mathematical population genetics: I. Theoretical introduction*, volume 27. Springer, 2004.

[Finkelstein, 1960] Louis Finkelstein. *The Jews: Their history, culture, and religion*, volume 1. Harper & Row, 1960.

[Fisher, 1930] RA Fisher. The genetical theory of natural selection, 1st edn. clarendon, 1930.

[Frazer *et al.*, 2007] Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, 2007.

[Gibson *et al.*, 2006] Jane Gibson, Newton E Morton, and Andrew Collins. Extended tracts of homozygosity in outbred human populations. *Human molecular genetics*, 15(5):789–795, 2006.

[Gravel, 2012] Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, 2012.

[Griffiths and Marjoram, 1997] Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257, 1997.

[Griffiths, 1991] RC Griffiths. The two-locus ancestral graph. *Lecture Notes-Monograph Series*, pages 100–117, 1991.

[Gusev *et al.*, 2009] Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326, 2009.

[Gusev *et al.*, 2012] Alexander Gusev, Pier Francesco Palamara, Gregory Aponte, Zhong Zhuang, Ariel Darvasi, Peter Gregersen, and Itsik Pe'er. The architecture of long-range haplotypes shared within and across populations. *Molecular biology and evolution*, 29(2):473–486, 2012.

[Hamburg and Collins, 2010] Margaret A Hamburg and Francis S Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[Harris and Nielsen, 2013] Kelley Harris and Rasmus Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS genetics*, 9(6):e1003521, 2013.

[Hartl and Clark, 1997] Daniel L Hartl and Andrew G Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

[Hartl, 1988] Daniel L Hartl. *A primer of population genetics.* Sinauer Associates, Inc., 1988.

[Hein *et al.*, 2004] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene genealogies, variation and evolution: a primer in coalescent theory.* Oxford university press, 2004.

[Henn *et al.*, 2011] Brenna M Henn, Christopher R Gignoux, Matthew Jobin, Julie M Granka, JM Macpherson, Jeffrey M Kidd, Laura Rodríguez-Botigué, Sohini Ramachandran, Lawrence Hon, Abra Brisbin, et al. Hunter-gatherer genomic diversity suggests a southern african origin for modern humans. *Proceedings of the National Academy of Sciences*, 108(13):5154–5162, 2011.

[Henn *et al.*, 2012] Brenna M Henn, Laura R Botigué, Simon Gravel, Wei Wang, Abra Brisbin, Jake K Byrnes, Karima Fadhlaoui-Zid, Pierre A Zalloua, Andres Moreno-Estrada, Jaume Bertranpetit, et al. Genomic ancestry of north africans supports back-to-africa migrations. *PLoS genetics*, 8(1):e1002397, 2012.

[Hernandez *et al.*, 2011] Ryan D Hernandez, Joanna L Kelley, Eyal Elyashiv, S Cord Melton, Adam Auton, Gilean McVean, Guy Sella, Molly Przeworski, et al. Classic selective sweeps were rare in recent human evolution. *Science*, 331(6019):920–924, 2011.

[Hindorff *et al.*, 2009] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[Hudson and Kaplan, 1985] Richard R Hudson and Norman L Kaplan. Statistical properties of the number of recombination events in the history of a sample of dna sequences. *Genetics*, 111(1):147–164, 1985.

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

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

[Huff *et al.*, 2011] Chad D Huff, David J Witherspoon, Tatum S Simonson, Jinchuan Xing, W Scott Watkins, Yuhua Zhang, Therese M Tuohy, Deborah W Neklason, Randall W Burt, Stephen L Guthery, et al. Maximum-likelihood estimation of recent shared ancestry (ersa). *Genome research*, 21(5):768–774, 2011.

[HUGR, 2013] The Hebrew University of Jerusalem HUGR. Hugr, the hebrew university genetic resource, 2013.

[Iafrate *et al.*, 2004] A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome. *Nature genetics*, 36(9):949–951, 2004.

[Kimura, 1969] Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

[Kingman, 1982a] JFC Kingman. Exchangeability and the evolution of large populations. 1982.

[Kingman, 1982b] John FC Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[Kingman, 1982c] John FC Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

[Kong *et al.*, 2002] Augustine Kong, Daniel F Gudbjartsson, Jesus Sainz, Gudrun M Jonsdottir, Sigurjon A Gudjonsson, Bjorgvin Richardsson, Sigrun Sigurdardottir, John Barnard, Bjorn Hallbeck, Gisli Masson, et al. A high-resolution recombination map of the human genome. *Nature genetics*, 31(3):241–247, 2002.

[Kong *et al.*, 2012] Augustine Kong, Michael L Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature*, 488(7412):471–475, 2012.

[Lander *et al.*, 2001] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[Lao *et al.*, 2013] Oscar Lao, Eveline Altena, Christian Becker, Silke Brauer, Thirsa Kraaijenbrink, Mannis van Oven, Peter Nürnberg, Peter de Knijff, and Manfred Kayser. Clinal distribution of human genomic diversity across the netherlands despite archaeological evidence for genetic discontinuities in dutch population history. *Investigative genetics*, 4(1):9, 2013.

[Li and Durbin, 2011] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

[Li and Leal, 2008] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[Liang *et al.*, 2007] Liming Liang, Sebastian Zöllner, and Gonçalo R Abecasis. Genome: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–1567, 2007.

[Madsen and Browning, 2009] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384, 2009.

[Mathieson and McVean, 2012] Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3):243–246, 2012.

[McEvoy *et al.*, 2011] Brian P McEvoy, Joseph E Powell, Michael E Goddard, and Peter M Visscher. Human population dispersal âĂIJout of africaâĂİ estimated from linkage disequilibrium and allele frequencies of snps. *Genome research*, 21(6):821–829, 2011.

[McVean and Cardin, 2005] Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

[Menelaou and Marchini, 2013] Androniki Menelaou and Jonathan Marchini. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*, 29(1):84–91, 2013.

[Mitchell *et al.*, 2004] Maria K Mitchell, Peter K Gregersen, Stephen Johnson, and Ramon Parsons. The new york cancer project: rationale, organization, design, and baseline characteristics. *Journal of Urban Health*, 81(2):301–310, 2004.

[Moore and others, 1965] Gordon E Moore et al. Cramming more components onto integrated circuits, 1965.

[Moran, 1958] Patrick Alfred Pierce Moran. Random processes in genetics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 54, pages 60–71. Cambridge Univ Press, 1958.

[Moran, 1962] Patrick Alfred Pierce Moran. The statistical processes of evolutionary theory. *The statistical processes of evolutionary theory.*, 1962.

[NHGRI, 2013] National Human Genome Research Institute NHGRI. Dna sequencing costs, 2013.

[Palamara and Pe'er, 2013] Pier Francesco Palamara and Itsik Pe'er. Inference of historical migration rates via haplotype sharing. *Bioinformatics*, 29(13):i180–i188, 2013.

[Palamara *et al.*, 2012] Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik PeâĂŹer. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 2012.

[Parida *et al.*, 2011] Laxmi Parida, Pier Palamara, and Asif Javed. A minimal descriptor of an ancestral recombinations graph. *BMC bioinformatics*, 12(Suppl 1):S6, 2011.

[Pemberton *et al.*, 2010] Trevor J Pemberton, Chaolong Wang, Jun Z Li, and Noah A Rosenberg. Inference of unexpected genetic relatedness among individuals in hapmap phase iii. *The American Journal of Human Genetics*, 87(4):457–464, 2010.

[Perry *et al.*, 2008] George H Perry, Amir Ben-Dor, Anya Tsalenko, Nick Sampas, Laia Rodriguez-Revenga, Charles W Tran, Alicia Scheffer, Israel Steinfeld, Peter Tsang, N Alice Yamada, et al. The fine-scale and complex architecture of human copy-number variation. *The American Journal of Human Genetics*, 82(3):685–695, 2008.

[Pickrell *et al.*, 2009] Joseph K Pickrell, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z Li, Devin Absher, Balaji S Srinivasan, Gregory S Barsh, Richard M Myers, Marcus W Feldman, et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, 19(5):826–837, 2009.

[Pool and Nielsen, 2009] John E Pool and Rasmus Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719, 2009.

[Pool *et al.*, 2010] John E Pool, Ines Hellmann, Jeffrey D Jensen, and Rasmus Nielsen. Population genetic inference from genomic sequence variation. *Genome research*, 20(3):291–300, 2010.

[Price *et al.*, 2006] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[Price *et al.*, 2010] Alkes L Price, Gregory V Kryukov, Paul IW de Bakker, Shaun M Purcell, Jeff Staples, Lee-Jen Wei, and Shamil R Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838, 2010.

[Price *et al.*, 2011] Alkes L Price, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS genetics*, 7(2):e1001317, 2011.

[Purcell *et al.*, 2007] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[Ralph and Coop, 2013] Peter Ralph and Graham Coop. The geography of recent genetic ancestry across europe. *PLoS biology*, 11(5):e1001555, 2013.

[Rasmussen and Siepel, 2013] Matthew D Rasmussen and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *arXiv preprint arXiv:1306.5110*, 2013.

[Risch *et al.*, 2003] Neil Risch, Hua Tang, Howard Katzenstein, and Josef Ekstein. Geographic distribution of disease mutations in the ashkenazi jewish population supports genetic drift over selection. *The American Journal of Human Genetics*, 72(4):812–822, 2003.

[Roach *et al.*, 2010] Jared C Roach, Gustavo Glusman, Arian FA Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010.

[Sabeti *et al.*, 2002] Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.

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

[Sheehan *et al.*, 2013] Sara Sheehan, Kelley Harris, and Yun S Song. Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, 2013.

[Slatkin, 2004] Montgomery Slatkin. A population-genetic test of founder effects and implications for ashkenazi jewish diseases. *The American Journal of Human Genetics*, 75(2):282–293, 2004.

[Spielman *et al.*, 1993] Richard S Spielman, Ralph E McGinnis, and Warren J Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506, 1993.

[Steinrücken *et al.*, 2012] Matthias Steinrücken, Joshua S Paul, and Yun S Song. A sequentially markov conditional sampling distribution for structured populations with migration and recombination. *Theoretical Population Biology*, 2012.

[Sun *et al.*, 2012] James X Sun, Agnar Helgason, Gisli Masson, Sigríður Sunna Ebenesersdóttir, Heng Li, Swapan Mallick, Sante Gnerre, Nick Patterson, Augustine Kong, David Reich, et al. A direct characterization of human mutation based on microsatellites. *Nature genetics*, 44(10):1161–1165, 2012.

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

[Tenenbaum *et al.*, 2000] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[The Genome of the Netherlands Consortium, 2014] The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the dutch population. *Nature Genetics, in press*, 2014.

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

[van Dongen, 2000] Stijn Marinus van Dongen. Graph clustering by flow simulation. 2000.

[Velez *et al.*, 2012] C Velez, PF Palamara, J Guevara-Aguirre, L Hao, T Karafet, M Guevara-Aguirre, A Pearlman, C Oddoux, M Hammer, E Burns, et al. The impact of converso jews on the genomes of modern latin americans. *Human genetics*, 131(2):251–263, 2012.

[Venter *et al.*, 2001] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[Voight *et al.*, 2006] Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLoS biology*, 4(3):e72, 2006.

[Wakeley, 2009] John Wakeley. *Coalescent theory: an introduction*, volume 1. Roberts & Company Publishers, 2009.

[Ward Jr, 1963] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[Watson and Crick, 1953] JD Watson and FHC Crick. A structure for deoxyribose nucleic acid. *Nature*, 171, 1953.

[Wiuf and Hein, 1999] Carsten Wiuf and Jotun Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, 1999.

[Wright, 1931] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

[Wright, 1943] Sewall Wright. Isolation by distance. *Genetics*, 28(2):114, 1943.

[Zaitlen *et al.*, 2013] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L Price. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS genetics*, 9(5):e1003520, 2013.

[Zuk *et al.*, 2012] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.