Genes in space

Selection, association and variation in spatially structured populations



Iain Mathieson
Lincoln College
University of Oxford

Thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Trinity Term 2013

Heartfelt thanks to Gil McVean, who could not have been a better supervisor.

To Cecilia Lindgren and Jonathan Flint for support and advice.

To the Wellcome Trust for generous funding.

Abstract

Genes in Space

Selection, association and variation in spatially structured populations

Iain Mathieson, Lincoln College. *D.Phil.* Trinity 2013.

Spatial structure in a population creates distinctive patterns in genetic data. There are two reasons to model this process. First, since the genetic structure of a population is induced by its historical spatial structure, it can be used to make inference about history and demography. Second, these models provide corrections to other analyses that are confounded by spatial structure. Since is it is now common to collect genome-wide data on many thousands of samples, a major challenge is to develop fast, scalable, approximate algorithms that can analyse these datasets. A practical approach is to focus on subsets of the data that are most informative, for example rare variants.

First we look at the problem of estimating selection coefficients in spatially structured populations. We demonstrate this approach using classical datasets of moth colour morph frequencies, and then use it in a model incorporating both ancient and modern DNA to estimate the selective advantage of one of the best known examples of local adaptation in humans, lactase persistence in Europeans.

Next, we turn to the problem of association studies in spatially structured populations. We demonstrate that rare variants are more confounded by non-genetic risk than common variants. Excess confounding is a consequence of the fact that rare variants are highly informative about recent ancestry and therefore, in a spatially explicit model, about location.

Finally, we use this insight into rare variants to develop methods for inference about population history using rare variant and haplotype sharing as simple summary statistics. These approaches are extremely fast and can be applied to genome-wide data on thousands of samples, yet they provide an accurate description of the history of a population, both identifying recent ancestry and estimating migration rates between subpopulations.

Contents

1 Introduction **1**

1.1 Models of spatial structure 2

1.1.1 Island models 3

1.1.2 Stepping stone models 6

1.1.3 Continuous models 8

1.1.4 Non-equilibrium models 8

1.1.5 Inference in models of spatial structure 10

1.2 Empirical measures of population structure 11

1.2.1 F_{ST} and allied measures 11

1.2.2 PCA 13

1.2.3 Sharing statistics 16

1.2.4 Other measures 17

1.3 Clustering methods 18

1.3.1 Generic clustering algorithms 18

1.3.2 Model based clustering algorithms 19

1.3.3 Explicit spatial clustering algorithms 20

1.4 Human population structure 21

1.5 Conclusion 23

2 Maximum likelihood estimation of selection coefficients **25**

2.1 Estimating selection coefficients 25

2.2 Single population maximum likelihood estimators 27

2.2.1 Model and notation 27

2.2.2 Maximum likelihood estimators 28

2.3 Structured population maximum likelihood estimators 31

2.3.1 Model and notation 31

2.3.2 Maximum likelihood estimators 33

2.4 Estimation with incomplete observations 35

2.4.1 Single population . . . 36
2.4.2 Structured population . . . 40
2.5 Selection on colour morphs in moths . . . 44
2.5.1 *Panaxia dominula* . . . 44
2.5.2 *Biston betularia* . . . 46
2.6 Extending to haplotype data . . . 50
2.7 Conclusion: Selection and structure . . . 55

3 Association studies of rare variants **57**
3.1 Association studies . . . 57
3.2 Simulating genotypes in the lattice model . . . 62
3.3 Association studies in the lattice model . . . 67
3.4 Corrections for structure in rare variants . . . 70
3.5 Conclusion: rare variants in association studies . . . 73

4 Rare variant sharing, history and structure **77**
4.1 The 1000 Genomes Project . . . 78
4.2 Allele sharing and structure . . . 78
4.3 f_2 sharing in the 1000 Genomes . . . 85
4.4 Estimating haplotype ages from f_2 sharing . . . 92
4.4.1 Method . . . 92
4.4.2 Results: Simulated data . . . 97
4.4.3 Results: 1000 Genomes data . . . 97
4.4.4 Discussion . . . 102
4.5 Estimating migration rates from f_2 sharing . . . 103
4.5.1 Method . . . 103
4.5.2 Results . . . 107
4.5.3 Discussion . . . 109
4.6 Conclusion: Rare variant sharing . . . 111

5 Conclusions **113**
5.1 Models: beyond the stepping stone . . . 113
5.2 Data: more, and more ancient . . . 115
5.3 Humans: when and where . . . 116

A Local selection in subdivided populations **119**

B Further examples of the effect of spatial risk on association studies **123**

C Distributions of 1000 Genomes haplotype ages **129**

D A note on tools **133**

References **134**

List of Figures

1.1 Wright's island model . . . 3
1.2 Kimura & Weiss's stepping stone model . . . 6
1.3 Wright & Malécot's isolation by distance model . . . 8
1.4 The pain in the torus . . . 9
1.5 The evolution of PCA . . . 15
1.6 The Li and Stephens model . . . 17
1.7 Principal components of human population structure . . . 21

2.1 Single population MLE . . . 30
2.2 The Wright-Fisher lattice model . . . 31
2.3 The Wright-Fisher lattice model . . . 34
2.4 Likelihoods in the lattice model . . . 35
2.5 Simulated data results in one population . . . 39
2.6 Simulated data results on a lattice . . . 43
2.7 *Panaxia dominula* morphs . . . 45
2.8 *Panaxia dominula* analysis . . . 46
2.9 *Biston betularia* morphs . . . 47
2.10 *Biston betularia* analysis . . . 49
2.11 Analysing LCT using ancient DNA . . . 53

3.1 Correcting for population structure in GWAS . . . 59
3.2 Published GWAS . . . 60
3.3 Simulating the coalescent in a spatially structured population . . . 62
3.4 Principle components in the lattice model . . . 64
3.5 Spatial distribution of rare and common variants . . . 64
3.6 Allele sharing of rare and common variants . . . 66
3.7 F_{ST} as a function of M . . . 67
3.8 Inflation in P-value by allele frequency . . . 69
3.9 Correlation between genotype and non-genetic risk . . . 70

3.10 Corrections for population structure 72
3.11 Rare variant burden tests 73

4.1 Effect of structure on allele sharing 81
4.2 Comparing PCA and sharing measures 82
4.3 Clustering histories based on allele sharing. 83
4.4 Doubleton sharing in 1000 Genomes 86
4.5 IBD sharing in 1000 Genomes 87
4.6 Genotype covariance in 1000 Genomes 88
4.7 IBD and f_2 sharing 90
4.8 IBD and f_2 sharing empirically 91
4.9 Finding haplotypes from f_2 sharing 93
4.10 Estimating f_2 haplotype age 94
4.11 Simulated data results 98
4.12 IBD and f_2 haplotype counts 100
4.13 Selected haplotype age distributions 101
4.14 Migration model 1 104
4.15 Migration model 2 105
4.16 Migration model 3 106
4.17 Estimated migration rates from simulated data 108
4.18 Estimated migration rates from 1000 Genomes data 109
4.19 Migration model 4 110

A.1 Equilibrium frequencies in the two deme model 121
A.2 Expected trajectories for overdominance and subdivision 122

B.1 Effect of smoothness of the risk distribution 124
B.2 Effect of the size of the risk area 125
B.3 Effect of multiple patches of risk 126
B.4 Effect of migration rate 127
B.5 Inflation when there are rare causal variants 128

C.1 Haplotype age distributions 129

Chapter 1

Introduction

What makes a structured population? Genetic structure implies that ancestry is non-random, so that some individuals are more likely to be related than others. There are many effects that cause this, for example complex mating systems, family structure, selection and incompatibility loci. In this thesis, we focus on the effect of spatial structure in the population which, together with migration, is typically responsible for much of the structure in naturally occurring populations. In fact, such structure is ubiquitous in nature and a fundamental property of population genetic systems.

Why do we care about population structure? There are two main reasons. First, genetic structure is interesting in itself because it can be used to make inference about population history in terms of events such as population splits, migrations and size changes. This is generally a difficult inference problem because different demographic effects can produce similar patterns of genetic structure, and in some cases may be impossible to distinguish. Nonetheless, this is an important area of research, and has recently produced important insights into the demographic history of both humans and other species. Second, population structure is a confounding factor for many of the other analyses which we want to make. These include selection scans and association studies where undiscovered genetic structure in the data obscures the effects we are trying to discover. In these cases we are not interested in understanding the genetic structure in itself, but as a covariate in a model for something else, and we may not care about understanding the process which generates it.

It has always been clear to geneticists that spatial structure is an important component of genetic variation and there are several large bodies of work approaching it from different directions, from model-based analyses to purely empirical descriptions. In this chapter we describe some of these approaches. We begin by describing the most popular models of population structure, dating back to the origins of population genetics in the early twentieth century. Next we describe some empirical measures which can be used to visualise and quantify the structure present in a populations. A common approach to summarising

structure is to cluster individuals into populations, or groups which are in some way similar, and we discuss common methods for doing this. Finally, as an illustration, we summarise what is currently known about population structure in humans, and how this relates to history and geography.

1.1 Models of spatial structure

Historically, much research has focused on analysing theoretical models of spatially structured populations. While these models are simplified, understanding their behaviour provides useful intuition about the behaviour of more complex models. In this section we describe the three most popular classes of models but before that, we briefly outline the historical context of these models. Many of the interesting results about structured populations are close analogues of those which were first derived in unstructured models. Indeed many classic results derived in unstructured models hold in structured models, or are empirically true in natural populations, even those with more complex dynamics.

The first real mathematical models of genetic variation were described in the 1920's and 30's. This is largely the era of Fisher, Haldane and Wright. Using relatively simple models and mathematical techniques, many results were derived about the distribution of variation in a population, and response to selection. As we will see later in this section, many of these results were extended to simple models of structured populations as well.

The second major development in terms of modelling strategies was the use in the 1960's by Kimura and others of diffusion theory to model the change in allele frequencies in a population over time. This technique was used to prove results such as the probability of, and expected time to, the fixation of a mutant allele in a population. In the context of spatially structured populations, Maruyma, Slatkin and others investigated island and stepping stone models.

The third and most recent major development in technique was the use of Kingman's coalescent in the 1980's. Rather than modelling allele frequencies directly, coalescent theory models the genealogy relating samples. This is a much more natural way to model the relationships between samples, and this technique turned out to be extremely powerful. The effects of selection and recombination can be incorporated through more complex models (the ancestral selection and recombination graphs; called the ASG and ARG, respectively), though these tend to be computationally intractable for the purposes of inference. Much recent work in population genetics has focused on finding approximations to these complex models so that they can be practically used on the large datasets which can be generated with current technology. The coalescent allowed many of the earlier results about spatially structured populations to be rederived and extended.

1.1.1 Island models

The term "island model" refers to a model which considers a population to be split into some (finite or infinite) number of discrete islands (or "demes", we will use these terms interchangeably). There is migration between demes, and individuals can migrate from any deme to any other deme uniformly. That is, there is no sense of distance in that all islands are equally far apart from each other.

The infinite island model (Figure 1.1) was first described by Wright (1931, 1951). Here there are an infinite number of islands, each with a finite diploid population of size N, each of which exchanges a proportion m of its population each generation with a 'mainland' containing an infinite population. We write $M = 2Nm$ for the total number of migrants exchanged. An equivalent formulation which is often used is an infinite number of islands, with each island receiving and donating migrants at the same rate, uniformly with the other islands[†].

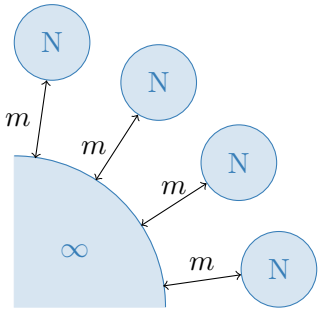


Figure 1.1: Wright's (infinite) island model. Disconnected demes of size N exchange migrants with an infinite 'mainland' at rate m per generation.

A useful statistic to measure the differentiation between subpopulations is the probability of *identity by descent* (Wright, 1921; Malécot, 1948, IBD) which is the probability that two alleles descend without mutation from a common ancestor. We can often characterise models of population structure by comparing the IBD probability for alleles from the same subpopulation with the IBD probabilities for alleles from different subpopulations. Another useful and related statistic is Wright's fixation index, F_{ST} (Wright, 1922, 1951). This has many possible definitions, which are related but may or may not correspond exactly to each other in different situations. One possible definition is the within-population correlation between individual genotypes. Another is as the fraction of variance explained by between-group variation[‡]. A third definition is in terms of *identity by state* (IBS) probabilities, or pairwise differences. Here we define F_{ST} in terms of IBD probabilities, since they are often the easiest quantity to analyse theoretically.

$$F_{ST} = \frac{P_W - P_T}{1 - P_T} \tag{1.1}$$

where P_W, and P_T are the probabilities of IBD for two samples from the same deme ("within") and from the entire population ("total"), respectively. Some authors have used

[†]To see why this formulation is equivalent, consider the point of view of any particular island. In the second formulation the sum of all the other islands is equivalent to the mainland in the first formulation.
[‡]This is like the use of F statistics in classical statistical methods such as ANOVA.

P_B, the probability of IBD for alleles drawn from different demes ("between") instead of P_T. The two definitions are the same in the limit of large number of demes. In the infinite island model, F_{ST} is given by

$$F_{ST} \approx \frac{1}{1+4Nm} \text{ (Wright, 1931).} \tag{1.2}$$

This expression drops off very quickly as Nm increases which tells us that when populations sizes are large, a relatively small number of migrants is sufficient to prevent to stop the island populations diverging greatly from the overall population. This is the source of the common assertion that one migrant per generation is sufficient to prevent two populations diverging.

A natural extension to the infinite island model is the finite island model (Maruyama, 1970). Now there are only a finite number d of demes but, as before, each deme exchanges, every generation, a proportion m of its population with the rest of the demes, with each migrant moving uniformly at random to one of the other $d-1$ demes. The equivalent expression to Equation 1.2 is

$$F_{ST} \approx \frac{1}{1+4Nm\left(\frac{d}{d-1}\right)} \text{ (Takahata and Nei, 1984)} \tag{1.3}$$

which converges rapidly to Equation 1.2 as $d \to \infty$.

We can analyse the island model in a coalescent framework by computing the distribution of coalescence times for the lineages in a sample. The above definition of F_{ST} in terms of IBD probabilities can be expressed as the ratio of the moment generating functions of coalescence times (Hudson, 1990; Wilkinson-Herbots, 1998). Slatkin (1991) showed that in the limit of zero mutation (when IBD is equivalent to identity by state; IBS), this reduces to a ratio of expected coalescence times,

$$F_{ST} = 1 - \frac{\mathbf{E}\left[T_W\right]}{\mathbf{E}\left[T_T\right]} \tag{1.4}$$

where T_W and T_T are the expected coalescence times of two lineages sampled from the same deme, and from the whole populations, respectively. When two lineages are in the same deme, the expected time to coalescence is $2Nd$, and does not depend on m. For two lineages in different demes, the expected time to coalescence is the sum of $2Nd$ and the expected time until they find themselves in the same deme, which is approximately $\frac{d-1}{2m}$ if m is small. Substituting these into Equation 1.4 recovers Equation 1.3. This separation of behaviour for within- and between-deme lineages was developed by Wakeley (1998, 1999) into a framework for understanding the behaviour of the genealogy of a sample of size larger than two. Starting with lineages spread across the demes, there is a first "scattering" stage when lineages either coalesce within a deme, or switch to another deme and then, when

there is only at most a single lineage in each deme, a "collecting" phase where the scattered lineages migrate around until they find themselves in the same deme and coalesce. Typically the "scattering" phase is short compared to the "collecting" phase.

That the expected coalescence time for two lineages in the same deme does not depend on the migration rate is somewhat surprising. It is a consequence of the fact that as the migration rate increases, the probability of migrating apart before coalescing increases, but the expected time to come back to the same deme decreases and these two effects cancel out. In fact, this is an example of a more general invariance principle extending beyond the island model. Maruyama (1971a) demonstrated that the number of heterozygotes in a population is independent of spatial substructure. Strobeck (1987) and Slatkin (1987) showed that in any conservative isotropic migration model, the expected number of nucleotide differences (and, equivalently, the expected coalescence time) of two lineages sampled from the same deme is independent of the migration rate. Nagylaki (1982) generalised these results to migration models which are neither conservative nor isotropic and Nagylaki (1998) proved a more general result, as follows. Suppose T_{ii} is the expected coalescence time for two lineages sampled from deme i, the population size of deme i is N_i, and the weighted mean coalescence time T_0 is defined by

$$T_0 = \frac{\sum_i \pi_i^2 N_i T_{ii}}{\sum_i \pi_i^2 N_i} \tag{1.5}$$

where π_i is the equilibrium distribution of a lineage over all demes[†]. Then $\mathbf{E}[T_0] = 2N_e$ where $N_e = \left(\sum_i \pi_i^2 N_i^{-1}\right)^{-1}$ is independent of the migration rate, in the sense that it depends on the relative rates (through π_i) but not on the absolute values. Of course, the second and higher moments of the within-deme coalescence times, and all the moments of the between-deme coalescence times, still depend on the migration rate. Generating functions for the distribution of pairwise coalescence times in the island models are given in Herbots (1997).

The coalescent used in these analyses is not the same process as the simple Kingman coalescent, but the *structured coalescent*. Generically this is a coalescent process where the lineages have labels, which can change as we go back in time, and only pairs of lineages with the same label can coalesce. In the case of structured populations, we identify the labels with demes, and the rates of change of labels with migration rates. The migration rates are thus migration rates of lineages backwards in time, rather than of individuals forwards in time, though in many cases they are equivalent. We define this process more carefully when we use it in Chapter 3. That this process exists as the limit of the ancestral process of some reasonable model of population dynamics is not obvious. Herbots (1997) showed that this is indeed the ancestral limit of a model in which each generation in each deme evolves as a

[†]i.e. the largest eigenvector of the backwards migration matrix

Wright-Fisher population of size N_i, and each pair of demes i, j exchange a fixed number q_{ij} of migrants each generation. This is a *conservative* migration model (Nagylaki, 1982), in the sense that every generation, the same number of individuals leave each deme as arrive. This means that migration does not affect the deme sizes, a strong assumption, though probably reasonable for small migration rates. Notohara (1990) had previously demonstrated that the same limiting ancestral process existed in a more complicated model which was not necessarily conservative.

1.1.2 Stepping stone models

While island models have the advantage of being tractable, they are unrealistic. In particular, for most populations, the assumption that an individual from any part of the range migrates to any other part of the range uniformly is likely to be false. Most populations will exhibit *isolation by distance* (Wright, 1938, 1943), where individuals are more likely to be closely related to individuals who are spatially close them. This is the crucial phenomenon which models of spatial structure seek to explain and we will spend the rest of this section describing such models.

The most natural extension of the island model to include isolation by distance is the so-called "stepping stone" or "lattice" model (Kimura, 1953; Malécot, 1959; Kimura and Weiss, 1964, Figure 1.2). Here, demes are laid out in a regular pattern, and migration can only occur between neighbouring demes. From any single deme, an individual can migrate to any of the four neighbouring demes. We have depicted a two-dimensional model here, since that is the most common situation for natural populations and the one we will focus on in the rest of this thesis. One-dimensional models are simple to analyse and may be applicable to some populations. Three- and higher-dimensional models may also be occasionally useful, though they are more difficult to analyse. Like island models, lattice models can be either finite or infinite, though finite models are more tractable. To construct a finite model, either migration stops at the edges of a finite range, or individuals "wrap" round the edges, turning the range into a torus (in two dimensions). Standard results about the behaviour of random walks imply that infinite models are impractical. In particular, in an infinite two dimensional model, the expected coalescence time of any two lineages is infinite, and in three- and higher-dimensions, there

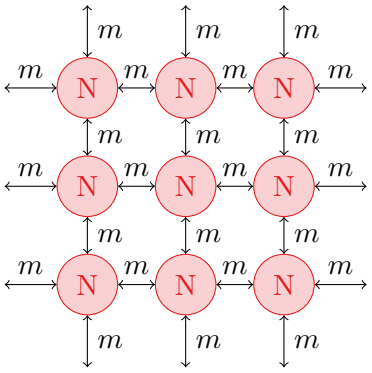


Figure 1.2: Kimura & Weiss's stepping stone model. Demes, each of size N exchange migrants with all neighbouring demes at rate m per generation.

is a non-zero probability that two lineages will never converge (Norris, 1998, Chapter 1.6; The simple symmetric random walk is null recurrent in two dimensions and transient in three or more.).

All the results derived earlier for the island model have natural analogues in the stepping stone model. In the stepping stone model, the degree of differentiation between subpopulations depends on the number of populations, whereas even in the finite island model it is largely independent. Suppose that there are d demes arranged in a $K \times K$ grid. Cox and Durrett (2002) show that, averaging over all demes,

$$F_{ST} \approx \frac{1}{1 + \frac{4Nm\pi}{\log(d)}}^{\dagger} \tag{1.6}$$

which implies that F_{ST} will be small if $4Nm > \frac{\log(d)}{\pi}$. In contrast, Equations 1.2 and 1.3 for the finite and infinite island models tell us that, for large d, F_{ST} is small when $4Nm > 1$. Further, for large Nm satisfying these conditions, $F_{ST} \approx \frac{\log(d)}{4Nm\pi}$, which is increasing in d while the equivalent expression for the finite island model, $\frac{1}{4Nm}\frac{d}{d-1}$ tends rapidly to $\frac{1}{4Nm}$ as $d \to \infty$. Note this is true only in the two dimensional model. The equivalent expression in one dimension is $\frac{1}{1+\frac{4Nm}{d}}$, which depends even more strongly on d, confirming the result of Weiss and Kimura (1965) that populations with linear structure are more differentiated than populations with structure in higher dimensions[‡].

Since, unlike the island model, some demes are closer than others, the statistics discussed above are now position-dependent. For example, the expected coalescence time for lineages in different demes depends on how distant the demes are. Weiss and Kimura (1965) derive the correlation between demes' allele frequencies as a function of distance in one, two and three dimensions. Maruyama (1970) and Maruyama and Kimura (1971) derive the IBD probabilities in various stepping-stone-like models, including the two-dimensional torus. Wilkinson-Herbots (1998) rederives these results using the structured coalescent.

Several authors have investigated models which are intermediate between the island and stepping stone models. In fact the original description of Kimura and Weiss (1964) was of just such a model, with both a short range migration rate m_1 to neighbouring demes and a long range migration rate m_∞ to all other demes. Maruyama (1971b) investigated a similar model. Unsurprisingly, the dynamics in these cases tend to be intermediate between the two models.

[†]Equation 2.11 of Cox and Durrett (2002), with $\sigma^2 = \frac{1}{2}$. Note in this equation, π actually means the constant π, i.e. 3.14159...

[‡]Intuitive since there is effectively only one path to reach any distant deme. Or to put it another way, *"each individual occupied the whole of the narrow path, so to speak, which constituted his Universe, and no one could move to the right or left to make way for passers by, it followed that no Linelander could ever pass another."* (Abbott, 1884).

1.1.3 Continuous models

The stepping stone model of the previous section seems unrealistic, albeit less so than the island model. Most populations we are interested in are not separated into discrete demes, but are distributed continuously in space. It is tempting therefore to investigate models of continuous spatial structure. Wright (1943) investigated such a model (Figure 1.3), as did Malécot at around the same time. In this model, individuals live in continuous space, and each individual draws its parents from a randomly distributed distance away or, in Malécot's formulation, each individual has independently a Poisson-distributed number of children, which disperse a random distance from the parents. Though this model seems intuitively more realistic that the island model, it has the problem that the assumptions about independence and random dispersion are inconsistent with a uniformly distributed population. Felsenstein (1975) demonstrated that as a result of this, individuals will eventually form into discrete clumps rather than being evenly spread across space (Figure 1.4).

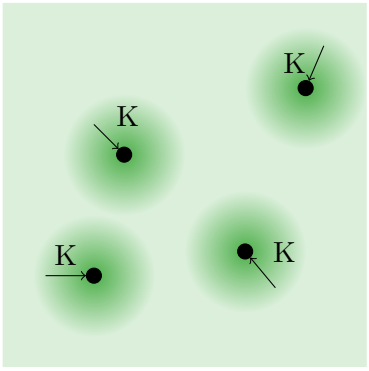


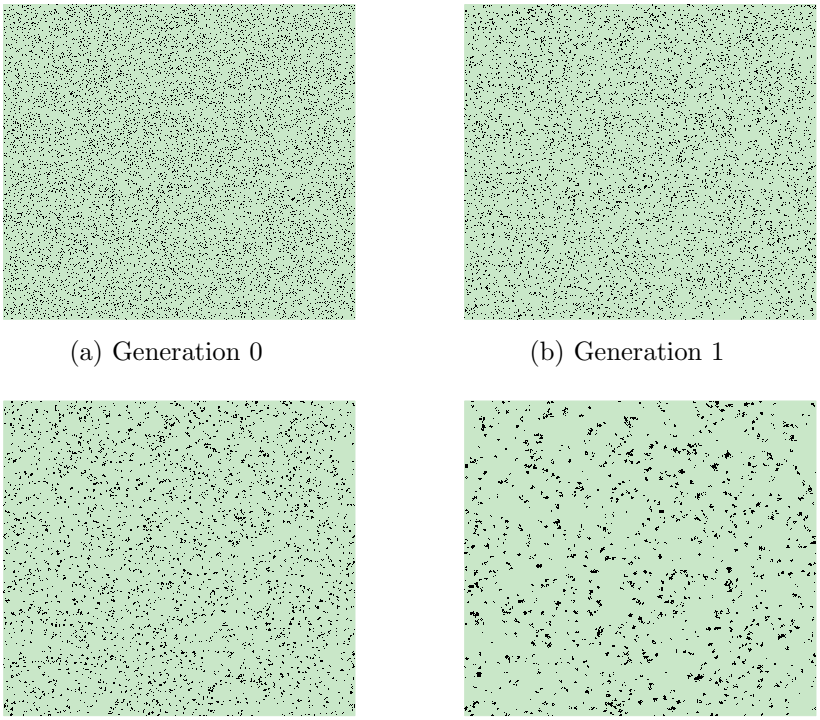
Figure 1.3: Wright & Malécot's isolation by distance model. Individuals live in continuous space and draw their parent(s) from a random distance with variance K.

There are a number of reasons why this is an undesirable property. For one, this effect is biologically unrealistic. Typically we might expect that natural populations would have a constraint on population density which would prevent this (availability of food, for example). More technically, the migration process is not reversible, which means that it is difficult to obtain the backwards process analogous to the structured coalescent for the stepping stone model.

Essentially, the problem is that this model does not have any constraints on local population density and therefore the state of uniform density is an unstable equilibrium. Contrast this with the stepping stone model, where each deme is constrained to have a fixed, constant, population size. This problem can be corrected by incorporating local density constraints into the model and such models have been introduced by Barton *et al.* (2002, 2010). However, these models can be intractable and in this thesis we concentrate on stepping stone and island-like models, which are easier to analyse and to simulate from.

1.1.4 Non-equilibrium models

All the models we have described so far in this section have been *equilibrium* models. That is, we have assumed that the population has reached a steady state in which the flow of



(a) Generation 0

(b) Generation 1

(c) Generation 4

(d) Generation 16

Figure 1.4: "The pain in the torus" (Felsenstein, 1975). The points represent the spatial locations of individuals on the (unwrapped) torus. We start with 10,000 individuals and every generation, each individual i has $N_i \sim \text{Po}(1)$ offspring which disperse a distance $D \sim \text{N}\left(0, 10^{-6}\right)$ away in a random direction. After only a few generations, the distribution has gone from uniform to noticeably "clumped".

migrants in and and out of an area is constant (at least in expectation), and the migration rates, population sizes and other parameters do not change over time. These assumptions are invalid for many real populations, and several authors have investigated models which do not conform to them. This has largely been driven by the desire to model ecological datasets, and describe phenomena such as extinction and colonisation events.

Extinction-recolonisation models have been investigated by several authors (Slatkin, 1977; Whitlock and McCauley, 1990; Le Corre and Kremer, 1998; Wakeley and Aliacar, 2001; Pannell, 2003). These are extensions to the island and stepping-stone models incorporating a process in which the entire population in one deme dies out and is replaced by individuals sampled from one or more of the other demes. The dynamics of the variation in the newly recolonised deme are determined by two factors. First the number and source of the colonists; if the number of colonists is small, and they are all from a single population, then the newly recolonised deme has little variation and F_{ST} is high. Conversely, if there are many colonists and they are drawn from all demes, then F_{ST} can be very low. The second factor is the migration rate. If there is little diversity in the recolonised population, then migration from the other demes acts to increase it, whereas if there is a lot of diversity then migration acts to reduce it, eventually tending to the equilibrium value. Interestingly, it is possible for migration to cause the diversity in the recolonised population to "overshoot" the equilibrium value before converging to it. The coalescent interpretation of this result is that coalescence rates are time varying. For lineages which have recently moved to the recolonised deme coalescence is very fast but as we go back in time the rate decreases, leading to bimodal distributions of both coalescence times and pairwise differences between sequences (Marjoram and Donnelly, 1994; Austerlitz *et al.*, 1997). Wakeley and Aliacar (2001) places these results in the context of the "scattering/collecting" model of structured populations, but demonstrate that patterns of diversity in these models can be very different from those obtained from simple migration models. In particular it is possible to observe more variants at intermediate frequencies, consistent with low coalescent rates at intermediate times.

1.1.5 Inference in models of spatial structure

The main inference question for the models described in this section is that of estimating migration rates between subpopulations. Historically there have been a number of approaches. As we described earlier, it is often possible to calculate the expected value of F_{ST} as a simple function of migration rates. Given a dataset, it is simple to compute F_{ST} and invert this relationship to provide an estimate of migration rates. Early analyses of spatially structured populations tended to follow this approach. More recently Slatkin

(1985) described a method for estimating migration rates from rare allele sharing (an idea to which we return in Chapter 4). Both Tufto *et al.* (1996) and Rannala and Hartigan (1996) used the covariance between allele frequencies in different demes. More recently coalescent approaches, typically using MCMC to sample from the structured coalescent, have been popular (Beerli and Felsenstein, 1999, 2001; Bahlo and Griffiths, 2000; Kuhner, 2006).

1.2 Empirical measures of population structure

In the previous section, we looked at theoretical models which had been proposed as models for the spatial structure of natural populations. While it is possible to make inference in these models, the question of whether we have specified an appropriate model is always an issue. The dynamics of real population are complex, almost inevitably more so than the models described in this section. Given this, there is value in a complementary "top-down" approach to the analysis of genetic data which can investigate the structure in the data while making few assumptions about the underlying process. In this section, we describe methods which take this approach. Rather than starting with a model and asking what its behaviour is, these methods start with data and ask what useful statistics we can compute, typically guided by standard statistical techniques, rather than population genetic intuition. As we might expect, this non-model based approach provides many results, but not always clarity of interpretation.

Though we discuss these methods in the context of spatial structure, they could be (and are) just as easily applied to any other form of population structure. Of the three methods we discuss, F_{ST} provides population level information, whereas PCA and sharing statistics provide individual level information which can be averaged over the individuals in a population. If populations are not *a priori* known, then we can use these measures to determine how to organise our individual samples into populations. We describe this clustering problem in the next section.

1.2.1 F_{ST} and allied measures

F_{ST} is a popular measure of differentiation between populations. It is simple to calculate, intuitive, and easy to analyse in specific models. On the other hand, it is often a source of confusion since, as we mentioned in the previous section, there are many different definitions of F_{ST}. Most authors do not state how they calculate it and different estimators can give very different answers. Confusingly, some authors treat it (implicitly or explicitly) as a parameter, and others as a statistic. It can therefore be quite difficult to compare results quantitatively across studies. Furthermore, the expected value of the statistic depends on

diversity, and on the choice of markers, so even if the same estimator is used, it cannot be compared across different populations.

F_{ST} was originally defined by Wright (1951) in the language of path coefficients, similar to the probability of identity by descent. In this sense it is a property of a particular mating system, rather than a statistic. As Wright observed, it has a natural interpretation in terms of ratios of variances[†] and Cockerham (1969) defined it as a statistic in terms of population allele frequencies,

$$F_{ST} = \frac{\mathrm{Var}_i(f_i)}{f(1-f)} \tag{1.7}$$

where f_i is the allele frequency in deme i, $\mathrm{Var}_i(f_i)$ represents the variance over the index i, and f is the mean allele frequency in the whole population, assuming a single bialleleic locus. Nei (1973) defined a related statistic G_{ST} for multiple alleles at a single locus. Suppose that allele j has average frequency f^j and frequency f_i^j in deme i. Then G_{ST} is a weighted average of F_{ST} for each allele,

$$G_{ST} = \frac{\sum_j \mathrm{Var}_i(f_i^j)}{\sum_j f^j(1-f^j)}. \tag{1.8}$$

In practice, G_{ST} and F_{ST} are used interchangeably. We are almost entirely concerned with bialellic variants, for which they are identical. Since $2f(1-f)$ is the probability of heterozygosity in a (randomly mating) population, these can be expressed in terms of heterozygosity probabilities,

$$F_{ST} = 1 - \frac{H_W}{H_T} \tag{1.9}$$

where H_w is the probability of heterozygosity if we took two gametes from the same population, and H_T is the probability of heterozygosity when sampled from the whole population. This leads to a simple estimator of F_{ST} from sequence data

$$F_{ST} = 1 - \frac{\Pi_W}{\Pi_T} \text{ (Hudson } \textit{et al.}\text{, 1992)} \tag{1.10}$$

where Π_W and Π_T are the number of pairwise sequence differences within populations and from the whole population[‡]. Weir and Cockerham (1984) provide probably the most widely used estimator, derived with a moment matching approach. Some authors compute F_{ST} by substituting observed frequencies into Equation 1.7 and then averaging over loci, either unweighted, or weighted by $f_i(1-f_i)$. The former weights rare variants relatively more so produces a lower estimate. There are also a number of extensions, R_{ST} for microsatellite

[†]Like a classical F-statistic.

[‡]Hudson *et al.* (1992) actually define their estimator as $1 - \frac{\Pi_W}{\Pi_B}$ where Π_B is the number of differences for sequences from *different* populations, but we have changed it to be consistent with our other definitions. As we mentioned in Section 1.1.1, the two are equivalent in the limit of infinite d.

data, Φ_{ST} for haplotype data and Q_{ST} for continuous data (reviewed by Holsinger and Weir (2009)).

Generally, the criticism of F_{ST} can be partitioned into two parts. The first is as described above, that there are many different, and inconsistent definitions and estimators. The second, more fundamental, criticism is that the quantity which is measured by F_{ST} is not interpretable, because it depends on diversity (Charlesworth, 1998; Jost, 2008). Specifically, when diversity is high F_{ST} is low. A related issue, important in the context of human SNP data, is that F_{ST} varies according to the allele frequencies used to estimate it. Suppose we have two equally sized demes, labeled 1 and 2, and some allele has frequency f_1 and f_2 in the two demes. Then for a total population frequency $\frac{1}{2}(f_1 + f_2) = \varepsilon$, Equation 1.7 implies that $F_{ST} < \frac{\varepsilon}{1-\varepsilon}$†.

Despite these problems, F_{ST} is widely used and, so long as its deficiencies are understood, it provides a useful method of making quantitative comparisons between populations.

1.2.2 PCA

Principal component analysis (Pearson, 1901; Hotelling, 1933) is a technique for reducing the dimensionality of large datasets. It is an orthogonal coordinate transformation which projects the data onto new axes, in such a way that the first axis explains as much of the variation in the data as possible, the second explains as much of the remaining variation as possible, and so on. Formally, let X be a $m \times n$ matrix with n observations, each of dimensionality m, such that each variable has zero mean, $\sum_{j=1}^{n} X_{ij} = 0 \; \forall i$. Then the *principal components* (PCs) Y are given by the transformation $Y = W^T X$ where W is an orthogonal $m \times m$ matrix chosen as follows. Let w_i be the i^{th} column of W, then w_1 satisfies $w_1 = \text{argmax}_{||w_1||=1} \left\{ ||w_1^T X||^2 \right\}$ and for $1 < i \leq m$, w_i satisfies $w_i = \text{argmax}_{||w_i||=1} \left\{ \left|\left| w_i^T \left[I - \sum_{j=1}^{i-1} w_j w_j^T \right] X \right|\right|^2 \right\}$. Equivalently, the principal components are the ordered eigenvectors of the sample covariance matrix. Since, typically, the first few principal components explain most of the variation in the data, we can often make an effective visualisation of the data by plotting the first few PCs against each other. This makes it very easy to visualise structure in the data, and detect obvious clusters and outliers. The main disadvantages of PCA are first, that is is sensitive to the relative scalings of each dimension and second, that since the PCs are arbitrary linear combinations of the original variables, they are typically not identifiable with any components of the orignal data, or parameters of the underlying models. In the context of genetic data, the first issue is usually

†Substituting f_1 and f_2 into Equation 1.7 gives $F_{ST} = \frac{(f_1 - f_2)^2}{(f_1+f_2)(2-f_1-f_2)}$ and if $f_1 + f_2 = 2\varepsilon$ then $(f_1 - f_2)^2 < 4\varepsilon^2$ so $F_{ST} \leq \frac{4\varepsilon^2}{2\varepsilon(2-2\varepsilon)} = \frac{\varepsilon}{1-\varepsilon}$. An example of this effect is shown in Figure 3.7

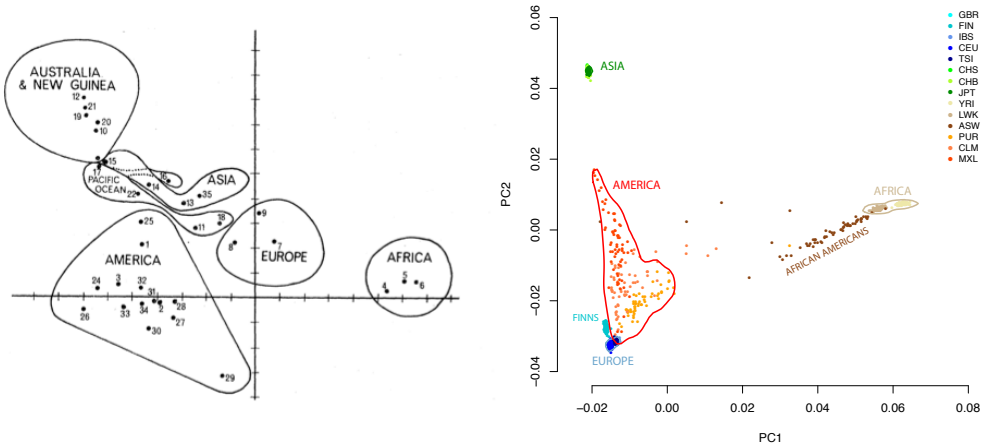
dealt with by weighting each marker by $\frac{1}{f(1-f)}$ where f is the marker frequency, so that each marker has the same variance.

The application of PCA to genetic data was pioneered by Cavalli-Sforza in the mid-1960s (Cavalli-Sforza and Barrai, 1964; Cavalli-Sforza, 1966). Using relatively small number of markers, from a wide range of populations, he investigated the structure and relationships between different human groups. Qualitatively, the results are remarkably similar to the results from PCA of modern SNP datasets (Figure 1.5). In the next few decades, PCA was commonly used to visualise genetic data (Harpending and Jenkins, 1973; Menozzi *et al.*, 1978; Melnick and Kidd, 1985; MacHugh *et al.*, 1998), though datasets were typically small enough that other visualisations could be used and early papers often quaintly include the entire raw dataset in printed tables. PCA[†] really became invaluable after the development of SNP genotyping chips around the year 2000 which allowed first hundreds of thousands, then millions of SNPs to be typed quickly and cheaply, in contrast to the tedious processes of allozyme typing or RFLP[‡] analysis. Such datasets, with the number of markers of order 10^5 or 10^6, absolutely require a dimensionality reduction technique to be summarised and visualised, and PCA is usually the most convenient and enlightening. Much of this increased interest was driven by the growing industry of genome wide association studies (GWAS), which used PCA to produce both visual summaries of genotype data, and quantitatively to correct for confounding due to population structure (Price *et al.*, 2006), an application to which we return in Chapter 3.

As we mentioned above, the interpretation of principal components is often non-obvious, and the principal components of genetic data are no different. Cavalli-Sforza and colleagues typically interpreted variation in principal components to indicate underlying dynamic processes, so that a cline in principal component space would indicate an admixture event, a selective gradient or, in particular, a migration event (i.e. a range expansion) (Menozzi *et al.*, 1978; Piazza and Menozzi, 1981). It is true that these events lead to the patterns described (see Patterson *et al.* (2006) for a modern analysis). However, as was eventually pointed out first by Sokal *et al.* (1999) and later Novembre and Stephens (2008) clines and other more complex regular patterns also appear naturally in the first few principal components of variation when spatially structured populations are at equilibrium. This is a problem for analysis. For example, the first two principal components of European genotypes almost perfectly recreate geographic North-South and East-West axes (Lao *et al.*, 2008; Novembre *et al.*, 2008; Wang *et al.*, 2010), but it is not clear whether this is as result of range expansion in both these directions, constant population structure with migration,

[†]And other more-or-less equivalent methods of dimensionality reduction such as multidimensional scaling (MDS).

[‡]Restriction fragment length polymorphism.



(a) Figure 82 from Cavalli-Sforza (1966). (b) Calculated from 1000 Genomes data.

Figure 1.5: Comparing **a**: PCA on allele frequencies of 15 allozyme loci in 35 populations. **b**: PCA on 2.1 million SNP genotypes for 1,092 individuals from the Phase 1 release of the 1000 Genomes Project. Despite the difference in the size of the datasets, the results are qualitatively similar. There are Asian, European and African clusters, and Americans are a diffuse (i.e. variably admixed) group.

---

or (most likely) a combination. In fact, there are several examples from different populations in which PCA (at least qualitatively) replicates geographic variation (Tian *et al.*, 2008; Chen *et al.*, 2009; Price *et al.*, 2009). McVean (2009) provided a unifying framework to understand what PCA actually represents, by showing that the principal components are simply a function of the expected coalescence times between lineages. Thus, any models which lead to the same expected coalescence times are indistinguishable using PCA. The distance measure in PCA space is equivalent to the covariance of the genotype vectors of the samples.

One major problem with using PCA to analyse genetic data is that it is very sensitive to both ascertainment of markers and sampling scheme. Choosing different individuals to base the analysis on can lead to very different conclusions. For example, DeGiorgio and Rosenberg (2013) demonstrate that for a population undergoing a range expansion the first principal component can be aligned either parallel or perpendicular to the direction of expansion, depending on sampling scheme. We must be therefore be extremely cautious about trying to make quantitative inference from PCA. It is perhaps most useful when it is used genome-wide to establish a null distribution for the relatedness between individuals, and markers are then searched for extreme deviations from this distribution, as in an association study, or a selection scan. Nonetheless, it remains a convenient and powerful tool for effectively summarising genetic data.

1.2.3 Sharing statistics

These are some of the simplest measures of difference between populations. To compute allele sharing probabilities, we simply count the probability that two individuals from the same population share an allele, and compare to the probability that two individuals from different populations share it. If the population is structured, then there is less sharing between populations than within populations. Clearly allele sharing probabilities are closely related to F_{ST} and PCA, since all can be expressed in terms of expected coalescence times. Allele sharing in pedigrees is used in linkage analysis, and association studies in unrelated samples are effectively tests for excess allele sharing within, relative to between, cases and controls. Historically, allele sharing has been used to infer population structure and model parameters (Slatkin, 1980, 1985; Bowcock *et al.*, 1994) and is still commonly reported as a data summary (Figure 3 of the 1000 Genomes Project Consortium (2012), for example). In principal, since there is a correlation between allele age and frequency, sharing at different frequencies gives us information about history as well as structure. Gravel *et al.* (2011) use this information (summarised in the joint allele frequency spectrum) to make inference about the demographic history of different populations.

Similar in spirit to allele sharing measures are IBD (or haplotype) sharing measures. These search for chunks of genome which are shared between individuals, which are indicative of recent ancestry at that locus. Given dense marker data, or sequence data, these measures can be extremely informative and reveal extremely fine scale structure, though they can be difficult and time consuming to implement. One class of approaches (Price *et al.* (2009); Lawson *et al.* (2012) for example) follows Li and Stephens (2003) and models each individual sequence in a sample as a mosaic of all the other sequences (Figure 1.6). This is equivalent to asking for each sample, at each site, which other sample is its genealogical nearest neighbour. This problem can be efficiently solved using hidden Markov model (HMM) techniques. Averaging, or counting chunks, over the genome then provides a distance measure between individuals. A related approach (Browning and Browning (2011) for example) is to attempt to detect specific shared chunks, without necessarily forcing the whole genome to be assigned. The extent of sharing between individuals can again be used as a distance metric. An advantage of this approach is that the distribution of chunk sizes is directly informative about demographic history (Ralph and Coop, 2013).

These statistics are all closely related to each other. In fact, in the case of unlinked markers, the expected sharing matrix from the mosaic approach (the *coancestry* matrix), the expected rescaled covariance matrix used to generate the principal components and the matrix of expected coalescence times are all equal (McVean, 2009; Lawson *et al.*, 2012). Additionally, as we mentioned earlier, if individuals are grouped into populations, then F_{ST}

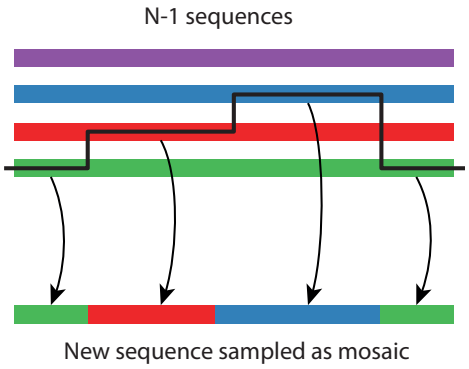


Figure 1.6: Suppose we have a sample of N sequences. Consider each in turn and model it as made up from a mosaic of the other $N-1$ sequences. This is equivalent to approximating the conditional sampling distribution (CSD) of the N^{th} lineage with a CSD where the other $N-1$ lineages never coalesce (a "trunk" genealogy).

can be expressed as a function of the expected coalescence times between groups, and so could be calculated in terms of these matrices (Slatkin, 1991). One of the main differences between IBD-type methods and PCA/covariance based methods is that the latter do not model linkage along the genome. This extra information is responsible for the superior performance of IBD methods in applications like clustering and imputation. We return to these statistics in much more detail in Chapter 4, and investigate carefully the links between allele sharing, IBD sharing, coalescence times and other measures of structure.

1.2.4 Other measures

One important feature of PCA is that it is explicitly linear. In contrast, Yang *et al.* (2012) developed an interesting approach which fits parameterised nonlinear functions to the spatial distribution of allele frequencies. This is likely to be a better fit to the data than the first few principal components, particularly when there is between-population variation in migration rate or population density, though the price of this is that the parameters will inevitably be difficult to interpret. This sort of approach is widely used in the ecological and landscape genetics literature, where the feature of interest is often correlation between landscape features and genetic variation.

All the measures we have described can be related to relatively simple functions of the genealogy of the sample. This is not surprising, since the genealogy is what generates the structure in the data. Given this, a natural approach would be to start with an appropriate metric on the space of trees, and derive the function of the sequences which best recovers this metric. In fact, there is a natural metric on the space of binary trees (Billera *et al.*, 2001) which can be used to define a (single) principal component for a set of trees (Nye, 2011). The appropriate metric on sequences would have the property that the expected distance between sequences from two trees would be the same as the distance between the two trees in tree space.

1.3 Clustering methods

In the previous sections, we have assumed that all the individuals in our sample were *a priori* labelled with respect to populations, and we looked at ways to investigate the relationship between populations. Often though, individuals are not labelled and we want to know whether it makes sense to arrange them into populations and if so, which individuals should go in which populations. We refer to this general class of problems as *clustering problems*. For now, we sidestep the question of what we mean by a population in a genealogical sense[†] and look at the three main types of method which are typically used, assuming that we can find a sensible working definition of a population, at least within each specific method. We describe three main classes of clustering algorithm. First, we look at generic clustering algorithms. These are clustering algorithms which do not assume any particular underlying model for the data. Second are model based clustering algorithms, which fit a specific model of population differentiation. The underlying assumption here is of exchangeability of individuals within populations. These typically perform better than generic algorithms. Finally, we look at explicitly spatial clustering algorithms, popular in the ecological genetics literature.

1.3.1 Generic clustering algorithms

The generic clustering problem can be stated as follows. Given a set of objects X, and a distance metric d (or, more or less equivalently, a similarity metric s) on this space, partition X into K subsets $X_1 \dots X_K$ in order to optimise some function of $\{X_1 \dots X_K\}$ and d. The number of clusters K can be regarded as a fixed value, or as a parameter to be optimised, depending on the application. We are trying to formalise a fundamentally subjective problem and there are few generally applicable results. As a consequence, the choice of metric and optimising function are somewhat arbitrary, though there may be canonical choices for specific applications. There is a very large literature on this class of problems in various contexts (see Xu and Wunsch (2005) for a recent review) and a large number of different methods. The choice of distance metric is largely independent of the choice of clustering algoritm, so we have two choices to make. In a genetic context, the distance metric is usually one of the metrics described in the previous section, either the covariance of the sequences (equivalent to clustering in PCA space), or a sharing measure.

Given a distance metric, the next choice is which clustering algorithm to use. K-means clustering places K points ("means") in the sample space, and assigns each observation to one of the means, minimising the total distance from the observations to their assigned

[†]We return to this question in Chapter 4.

means. Though easy to fit[†] and intuitively reasonable, if suffers from a number of well known issues, particularly sensitivity to outliers and the tendency to produce equal sized clusters, even when inappropriate. Thus, in practice, K-means clustering is rarely used.

Another class of generic methods are hierarchical clustering methods. These partition the observations into a series of hierarchical sets, which can be interpreted as a tree. By cutting off the tree at a particular height, we can partition the data into K clusters. Hierarchical clustering is attractive for genetic data, because it seems to reflect the underlying tree structure which generates the data. The algorithm starts with every sample in its own cluster and at each iteration merges clusters based on the distance between them. There are many ways of defining the distance between clusters, but popular choices are the maximum distance between individuals in the cluster (as implemented in *PLINK* (Purcell *et al.*, 2007) for example), or the average distance between clusters (called UPGMA[‡]).

In both of these cases, if K is unknown, it is usually determined using the Caliński criterion (Caliński and Harabasz, 1974) which compares the between and within group variances. In fact, if the variances are computed in genotype space, this is equivalent to choosing K to maximise F_{ST}, with an additional term to penalise large K.

1.3.2 Model based clustering algorithms

We now turn to clustering algorithms which explicitly model the structure of allele frequencies in the subpopulations. This is a common statistical problem. In genetics the archetypal model based algorithm is *STRUCTURE* (Pritchard *et al.*, 2000; Falush *et al.*, 2003). This fits a model in which the expected allele frequencies in each cluster are parameters, and the individual observations are sampled from these clusters. The cluster frequencies and assignments are fitted using Markov chain Monte Carlo (MCMC). *STRUCTURE* also allows us to model some individuals as admixed (i.e. drawing ancestry from more than one cluster). There is an *ad hoc* procedure to estimate K, but typically we need to explicitly try different values of K in order to get a sense for how robust the clusters are. The *ADMIXTURE* package developed by Alexander *et al.* (2009) takes a maximum likelihood approach to fitting the same model.

A similar method, *fineSTRUCTURE*, is described by Lawson *et al.* (2012). This takes a similarity matrix as input, and fits a model in which the expected similarities for each cluster are parameters and the observed individual similarities are samples. Any similarity matrix can be used, but the method performs best using the matrix which is constructed by running a Li and Stephens type model for each sequence (Figure 1.6), and then counting

[†]At least approximately, using a greedy or EM algorithm. The full problem is NP-hard (Mahajan *et al.*, 2012).

[‡]Unweighted pair group method with arithmetic means.

the number of chunks copied from each individual. Because this is a simpler model, it is possible to fit K along with the cluster expectations and assignments using reversible jump MCMC. Although they model only the total coancestry for each individual, Lawson *et al.* (2012) show that in the case of unlinked markers their coancestry matrix actually contains all the information used by *STRUCTURE* and thus that the two methods are equivalent. However, by modelling linkage along the genome while constructing the similarity matrix, *fineSTRUCTURE* has better performance.

Lawson and Falush (2012) review the performance of these clustering algorithms. In general, they find that model based approaches significantly outperform the generic methods, regardless of which distance metric is used. One important point is that for linked markers, failing to model linkage significantly decreases performance. Lawson and Falush (2012) focus on dense human SNP data, for which these methods tend to be optimised, and it is possible that different algorithms will perform better on different datasets.

1.3.3 Explicit spatial clustering algorithms

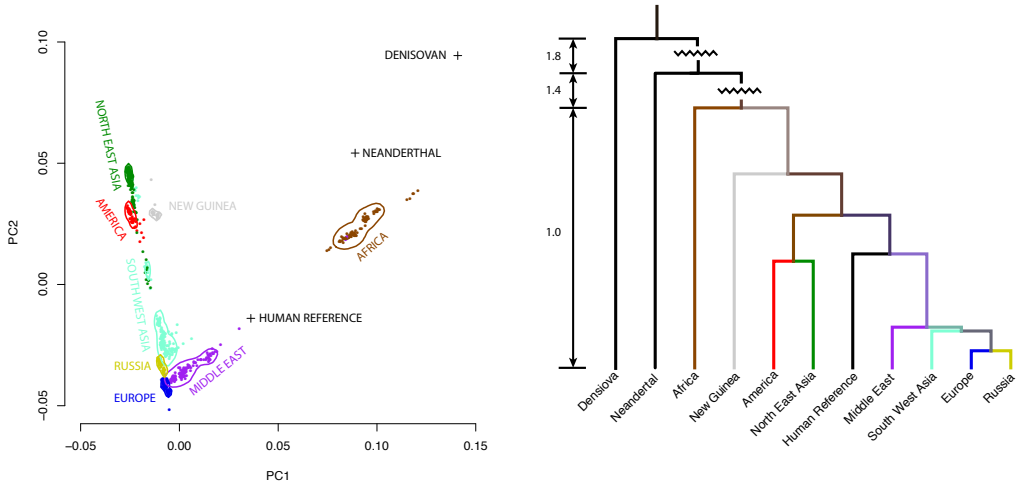
The clustering methods we have described so far use only genetic information. Where information about the spatial origin of the samples is available, it can be used to inform clustering, assuming that the clusters are likely to have at least some spatial basis. We can think of this as putting a prior on cluster assignment based on spatial position, rather than the uniform prior implied by all the methods described so far. This approach has been largely developed and used by ecological geneticists, and has not been much applied in human genetics[†].

One approach, implemented in the software *GENELAND* (Guillot *et al.*, 2005) is to generate random spatial tilings and fit a model which merges the tiles into clusters, and samples individuals from clusters, in the same way as the model-based clustering methods described above. A similar approach (François *et al.*, 2006; Chen *et al.*, 2007) bases tiles on the sampling position of the individuals, and uses an explicit model for how likely neighbouring individuals are to belong to the same cluster[‡]. This tends to form less regular clusters and may be more appropriate when complex spatial structure is expected.

These methods are probably too computationally expensive to run on genome-wide SNP datasets, but the idea of placing a prior on cluster membership based on geographic origin could easily be extended to any of the other model-based clustering methods described

[†]Perhaps partly because of a belief that humans migrate more than other species so spatial clusters are less stable, or because privacy concerns prevent the collection of this data, or because the methods developed are too computationally intensive for large human datasets.

[‡]The Potts model, which in this context is a distribution over the space of cluster memberships, with the Markovian property that the distribution of membership of any individual, conditional on its neighbours, is independent of all the other individuals.



(a) (b)

Figure 1.7: Visualisations of human population structure using a dataset collated by Patterson *et al.* (2012). This contains a total of 828 samples from 54 human populations (largely from the Human Genome Diversity Project) plus the human reference sequence and two extinct human species. Each sample is typed at a set of 620,765 SNPs ascertained across the populations. We have arbitrarily grouped the modern human populations into eight regional groups. **a**: First two principal components, computed from 93,862 LD-filtered SNPs. PC1 accounts for 3.9% of the variation in the data and PC2 for 2.8%. **a**: UPGMA tree joining population centres of gravity using the first ten principal components. For display purposes we have cut the branches linking the two extinct human species, which are longer than the rest of the tree (relative branch lengths on left of tree).

---

above and this would be an interesting approach to the modelling of spatial structure in humans.

1.4 Human population structure

All the methods of looking at population structure which we have discussed are generic, in the sense that they can be applied to data from any species. However they have almost all, at some point, been applied to humans. We probably have more data about humans than about any other natural species and, although it is still patchy, we can make some robust inference. In this section, we review the current state of knowledge about human population structure, and discuss the implications of this structure for other analyses of genetic data. Much of the rest of this thesis concentrates on investigating these effects and, although we will not always be working on human data, it will be useful to have a sense of the magnitude of the issue in humans. Though as we shall see, there is a large spatial component to human structure, there is also a large contribution from demographic history, in particular bottlenecks and admixture events, which we will have to take into account.

On a worldwide scale, the first two principal components of genetic variation typically distinguish African from non-African populations, and Europeans from Asians (Figures 1.5b and 1.7a). These reflect the oldest splits in human history and are consistent with estimates that West Africans and Eurasians diverged around 140 thousand years ago (kya) and that Europeans and East Asians diverged around 23kya (Gutenkunst *et al.*, 2009)[†]. Non-African populations have lower diversity than African populations, most likely the result of a population bottleneck related to both a limited number of individuals migrating out of Africa (the "out of Africa" bottleneck, at most 100kya) and the last glacial maximum, ending around 20kya (Li and Durbin, 2011). On these scales, the 1000 Genomes Project Consortium (2010) reported pairwise F_{ST} of 0.07 between European and West African populations, 0.08 between East Asians and West Africans, and 0.05 between East Asians and Europeans. These numbers are unweighted averages over all sites, and are therefore lower (empirically, by a factor of about 2) then the corresponding weighted average. Subsequent principal components tend to pull out specific populations, reflecting more recent splits. The tree in Figure 1.7b is a very rough approximation of the relationship between major human groups. The first split in modern humans after the African split is between Papuans and non-Papuans, reflecting the fact that they (and Aboriginal Australians) are representative of the first human migration out of Africa, largely replaced in the rest of Asia by subsequent migrations (Rasmussen *et al.*, 2011). More sophisticated model-based inference uses the structure in this modern human data to make inference about population splits, migrations, admixtures, size changes, and other demographic events on global scales (Gutenkunst *et al.*, 2009; Li and Durbin, 2011; Pickrell and Pritchard, 2012; Sheehan *et al.*, 2013, and many more). One of the most exciting recent developments in human genetics has been the sequencing of members of two extinct human groups, the Neanderthals and the Denisovans, and the discovery that many modern human populations have ancestry derived from one or both of these groups (Green *et al.*, 2010; Reich *et al.*, 2010; Meyer *et al.*, 2012; Sankararaman *et al.*, 2012; Wall *et al.*, 2013). Actually dating all these demographic events requires a molecular clock, either a mutational clock traditionally calibrated to human-chimp divergence using the fossil record, but increasingly commonly by direct measurements of the mutation rate, or a recombinational clock which uses the information contained in the size of shared haplotypes. In Chapter 4, we describe a method to combine both these sources of information to infer coalescence time distributions across the genome.

While most human population structure is probably due to drift in neutral variants, at least some appears to be due to natural selection, with spatially varying environment

[†]These, and all dates in this context, have very large confidence intervals, not least because they scale with the assumed average generation time. Archaeological evidence can provide lower bounds, but they may not be very tight.

likely to play a large role (Fumagalli *et al.*, 2011; Hancock *et al.*, 2011). Lack of power to detect regions under selection makes it difficult to quantify how much of the total variation is due to selection rather than drift, though the relatively small effective population size of humans (on the order of 10^4) suggests that it will not be large. We investigate further the problem of detecting selection in the presence of spatial structure in Chapter 2.

On a smaller scale, within a continent or country, as we described in Section 1.2.2 patterns of variation still reflect geographic structure and migration patterns. Just as on the global scale, these patterns can be used to make inference about structure and history. For example, Reich *et al.* (2009) demonstrated that population structure in India is consistent with the existence of two geographically distinct ancestral populations and with many modern populations descending from founder events. Reich *et al.* (2012) strongly support the hypothesis that Native Americans descend from three waves of migration from Asia. In Europe, Ralph and Coop (2013) see patterns of relatedness largely consistent with known historical events, whereas Pickrell *et al.* (2012) detect structure in Khoisan populations in southern Africa which presumably represents some, previously unknown, population history or isolation. Even at this level, structure can be due to differential selection. Turchin *et al.* (2012) show that variants which affect height are more differentiated across Europe than neutral loci, most likely as a result of selection. Undoubtedly in the next few years these, and many more similar results, will be stitched together into a rich narrative of human genetic history.

On these scales F_{ST} is extremely low, for example 0.004 between European countries (Novembre *et al.*, 2008) and 0.0026 between different regions of Iceland (Price *et al.*, 2009). Nonetheless, not only is this level of structure sufficient to make inference about historical events but we can even distinguish individuals within a country according to their geographic origin. Importantly, this level of structure is also sufficient to create inflation of test statistics in population-based association studies, and a major driving force in the study of human population structure in recent years has been the need to control for its confounding effect in genome-wide association studies (GWAS), an issue to which we return in Chapter 3, in particular the difference between common variants, which tend to be shared across populations, and rare variants, which do not.

1.5 Conclusion

In this chapter we described two classes of approaches for investigating population structure. The first is a purely theoretical model-based approach. We construct migration models and analyse their theoretical properties, generally in terms of summary statistics like coalescent times, or F_{ST}. This provides useful quantitative insights into the effect of structure, for

example the observation that if $4Nm > 1$ then F_{ST} is low, which suggests that very low levels of migration are sufficient to prevent populations from diverging substantially. However, while it is possible to perform inference in these models, it is difficult to be sure that we are capturing all the important dynamics of the model in our analysis. For the purposes of demographic inference the solution tends to involve the use of models which are less generic although these typically require, either explicitly or implicitly, a strong prior on the relationships between individuals or populations. An alternative approach is to extend the generic models of Section 1.1 to more complex dynamics, such as selection, or variable migration.

The second approach to understanding population structure is a purely empirical one. Tools like PCA, sharing measures, and clustering algorithms provide a way to investigate structure qualitatively, without thinking too hard about the processes which generate it. While this is useful for applications like association studies where structure is a nuisance to be controlled, it is ultimately rather unsatisfactory. Moving the interpretation of these measures to quantitative parameter estimates is an important area of research.

The rest of this thesis describes three attempts to partially bridge the gap between model-based and empirical views of population structure. In Chapter 2, we extend the stepping stone model to include spatially varying selection, and develop a model for inference in this setting. Then, in Chapter 3, we investigate the effect of spatial structure on confounding in association studies, in particular the effect of rare variants. Here, understanding the processes which drive structure helps us to realise why standard analyses break down at a particular point, and provides an example where empirical summaries of the data may be misleading. Finally, in Chapter 4, we show how data summaries from dense genotyping and sequence data, particularly allele and haplotype sharing metrics, can be used to estimate demographic and population genetic parameters with very few assumptions about the prior distributions of those parameters.

Chapter 2

Maximum likelihood estimation of selection coefficients

In this chapter, we consider the problem of estimating selection coefficients from time series data of allele frequencies in both the single and structured population cases. It is curious that the development of such estimators has received surprisingly little attention. Historically, most of the work in this area has been concerned with the calculation of quantities like fixation probabilities, rather than the estimation of selection coefficients, and many of the results derived in this chapter are surprisingly simple.

Initially, we assume that the population allele frequency is known exactly at every generation. This allows us to derive almost exact MLEs. However this situation is rarely encountered and perhaps is why these results have not been derived elsewhere. Typically, we have only a sample of individuals from a population, and we do not have data at every generation. This problem is naturally cast as a hidden Markov model (HMM), with the true allele frequency as a hidden state. We demonstrate both how to solve this model for a single population and an efficient extension to the structured case. We demonstrate both of these methods on classical datasets of moth colour morph frequencies. Finally we discuss how the methods we have developed might be integrated into analysis of genomic data from a single timepoint, which we illustrate with the well-studied example of recent selection on the lactase persistence phenotype in humans.

The work in this chapter largely follows Mathieson and McVean (2013). Most of the text in Sections 2.1-2.5 is from this reference, as well as Figures 2.2, 2.5, 2.6, 2.8 and 2.10.

2.1 Estimating selection coefficients

Detecting selection and estimating selection coefficients are important problems in many areas of genetics. In humans, genome-wide scans for selection identify regions of the genome which have been important in human evolution and provide clues about the location of

important functional variants (Bustamante *et al.*, 2005; Nielsen *et al.*, 2005; Voight *et al.*, 2006; Sabeti *et al.*, 2007). In pathogen research, understanding selection can help to understand and control the evolution and spatial spread of drug resistance in both pathogens and vectors. In cancer, intra-tumour selection is an important driver of tumour growth and development (Bignell *et al.*, 2010). With only a sample from one timepoint, for example with the human selection scans described above, it is difficult to obtain quantitative estimates of selection coefficients. However, with time series data on allele frequencies, for example from experimental evolution experiments, ecological observations, or ancient DNA, it is much easier (Bollback *et al.*, 2008; Illingworth and Mustonen, 2011; Malaspinas *et al.*, 2012). We return to the question of how this data might be jointly analysed with single timepoint data in Section 2.6.

Spatially structured populations are particularly interesting in terms of selection because a species' range may span many different environments, each subject to different selective pressures. Humans are a particularly good example of this and it is well established that adaptation to spatially varying selective pressures such as climate has been an important driver of recent human evolution (Hancock *et al.*, 2011). Such selection underlies many of the most obvious genetic differences between different human populations, for example differential susceptibility to infectious and autoimmune disease, and anthropometric differences like height and skin colour[†] (Jablonski and Chaplin, 2010; Turchin *et al.*, 2012).

If we assume we know the allele frequency exactly at every generation, then estimating the selection coefficient is relatively simple (Wright and Dobzhansky, 1946; DuMouchel and Anderson, 1968; Schaffer *et al.*, 1977). However typically this is not the case, and the analysis of time series data of allele frequencies uses a hidden Markov model (HMM) framework (Williamson and Slatkin, 1999; Bollback *et al.*, 2008). The allele frequency trajectory is modelled as a Markovian process, either a discrete process like the Wright-Fisher or Moran models, or as a diffusion (i.e. the limiting case of the discrete models). The observations are modelled as binomial observations from this population. Williamson and Slatkin (1999) use this approach to compute a likelihood surface for the effective population size N_e, assuming no selection. A similar approach to estimating N_e is used by Anderson *et al.* (2000) and Wang (2001). In order to estimate the selection coefficient s, Bollback *et al.* (2008) use numerical techniques to compute a likelihood surface and estimate $2N_e s$. Malaspinas *et al.* (2012) use an approximate transition density to compute the likelihood. They do this for

[†] "*As [man] ranged farther from his original home, and became exposed to greater extremes of climate, to greater changes of food, and had to contend with new enemies, organic and inorganic, useful variations in his constitution would be selected and rendered permanent, and would [...] be accompanied by corresponding external physical changes. Thus arose those striking characteristics and special modifications which still distinguish the chief races of mankind.*" (Wallace, 1864).

a grid of parameter values to estimate s and other parameters. Our method differs from these approaches in that we use an expectation-maximisation (EM) algorithm to maximise the likelihood, rather than numerical search.

As we described in Chapter 1, there are many ways in which to model spatially structured populations. The continuous spatial spread of a selected mutation is usually described using the travelling wave theory of Fisher (1937). This powerful tool can be extended to more complex situations such as the spread of multiple competing alleles (Ralph and Coop, 2010) or the existence of spatially varying selection pressures (See box 1 of Novembre and Di Rienzo (2009) for a brief review of such models). However these models can be difficult to fit to data. Our analysis here uses the lattice model of population subdivision, largely because in this model it is possible to derive relatively simple expressions for the MLE, and the discrete nature of the model means that it is possible to calculate an efficient solution to the HMM problem.

2.2 Single population maximum likelihood estimators

In this section we derive maximum likelihood estimators for the selection coefficient of a single allele in a single unstructured populations.

2.2.1 Model and notation

Consider a haploid Wright-Fisher population of constant size $2N_e$. We are interested in the frequency of a single allele with two types, A and a. Suppose the a allele has frequency f_t at generation t for $t = 0, \ldots, T$. The a allele has selection coefficient s so the relative finesses of the A and a genotypes are 1 and $1 + s$. Then at each generation t, the number of type a individuals is drawn from a binomial distribution with size $2N_e$ and probability $\frac{(1+s)f_{t-1}}{1+sf_{t-1}}$. We observe a sample of n_t individuals at generation t of which a_t are of the selected type, so that $\frac{a_t}{n_t}$ is an empirical estimate of f_t. We can represent missing observations by setting $n_t = 0$. For sufficiently large $2N_e$, and f_t not too small, we can approximate the distribution of the difference $f_{t+1} - f_t$ by a normal distribution with mean $sf_t(1 - f_t)$ and variance $\frac{f_t(1-f_t)}{2N_e}$.

In order to consider the effects of non-additive selection, we consider a diploid population of size N_e ($2N_e$ chromosomes). Assume that the three genotypes AA, Aa and aa have relative fitnesses 1, $1 + 2hs$ and $1 + 2s$ respectively where h is the heterozygous effect. The factor of 2 ensures that in the case of additive selection ($h = \frac{1}{2}$) the dynamics are the same as in the haploid case described above. $h = 0$ corresponds to a fully recessive allele and $h = 1$ to a fully dominant allele. In this case, the number of type a chromosomes

in the next generation is no longer binomial [†], but for large N_e, it is well approximated by a binomial distribution with size $2N_e$ and probability $\frac{(1+2s)f_t^2+(1+2hs)f_t(1-f_t)}{(1+2s)f_t^2+2(1+2hs)f_t(1-f_t)+(1-f_t)^2}$. As in the haploid case, for large N_e we can approximate the distribution of the difference $f_{t+1}-f_t$ by a normal distribution with mean $2sf_t(1-f_t)(f_t+h(1-2f_t))$ and variance $\frac{f_t(1-f_t)}{2N_e}$. See Ewens (1979) for a fuller discussion of this model and Nagylaki (1990) for a full explanation of the diploid case.

2.2.2 Maximum likelihood estimators

In the haploid model, or the diploid model with $h=\frac{1}{2}$, if the allele frequency of the selected allele in generation t is f_t, then the allele count in generation $t+1$ has a binomial distribution

$$\mathbb{P}\{f_{t+1}=f|f_t\}=\binom{2N_e}{2N_ef}\left(\frac{f_t+sf_t}{1+sf_t}\right)^{2N_ef}\left(\frac{1-f_t}{1+sf_t}\right)^{2N_e(1-f)}. \tag{2.1}$$

From this we can write down the log-likelihood $\ell(s)$ of s given the full trajectory (that is, sampling every individual at every generation), dropping terms that do not depend on s,

$$\ell(s)=2N_e\sum_{t=1}^{T}\{f_t\log(1+s)-\log(1+sf_{t-1})\}. \tag{2.2}$$

Since we have assumed N_e to be constant, the log-likelihood depends on N_e only though a constant multiple, and the MLE does not depend on N_e. If N_e were varying (but known), then in all the following analysis we could simply weight the terms in Equation 2.2 by N_e at each generation. Differentiating Equation 2.2 with respect to s, and setting equal to zero, we obtain the following equation satisfied by the MLE for s, which we denote by $\hat{s}$,

$$\sum_{t=1}^{T}\left\{\frac{f_{t-1}(1+\hat{s})}{1+f_{t-1}\hat{s}}\right\}-\sum_{t=1}^{T}f_t=0. \tag{2.3}$$

Writing $q(\hat{s})$ for the expression on the left hand side of Equation 2.3 and assuming that $0<f_t<1$, we have $q(-1)=-\sum_{t=1}^{T}f_t<0$, $\lim_{\hat{s}\to\infty}q(\hat{s})=T-\sum_{t=1}^{T}f_t>0$ and $q'(\hat{s})=\sum_{t=1}^{T}\frac{f_{t-1}(1-f_{t-1})}{(1+f_{t-1}\hat{s})^2}>0\ \forall\ \hat{s}>-1$. Since q is continuous for $\hat{s}>-1$, this implies that there is exactly one solution to Equation 2.3 in the range $-1<\hat{s}<\infty$, and therefore the MLE is unique. There is no simple analytic solution to Equation 2.3 but assuming

[†]It is no longer binomial because of the sampling variance of the genotypes. It is binomial conditional on the genotype counts. The probability in the main text assumes Hardy-Weinberg equilibrium. In fact, conditional on the a allele count $F_t=2N_ef_t$, the distribution of the genotype counts is given by $\mathbb{P}(N_{Aa}=x|F_t)=\frac{N_e!F_t!(2N_e-F_t)!2^x}{x!(N_{AA})!(N_{aa})!}$ where $N_{\bullet\bullet}$ is the genotype count for the $\bullet\bullet$ genotype and the number of homozygotes $N_{AA}=\frac{1}{2}(2N_e-F_t-N_{Aa})$ and $N_{aa}=\frac{1}{2}(F_t-N_{Aa})$ are fully determined by F_t and N_{Aa} (Haldane, 1954). Then conditional on these genotype counts, the distribution is binomial with size $2N_e$ and parameter $\frac{2(1+2s)N_{aa}+(1+2hs)N_{Aa}}{2((1+2s)N_{aa}+(1+2hs)NAa+N_{AA})}$.

$|s| < 1$, we can obtain an approximate solution by expanding the expression in powers of $\hat{s}$. Expanding to first order yields the solution

$$\hat{s} = \frac{f_T - f_0}{\sum_{t=0}^{T-1} f_t \left(1 - f_t\right)} + O(\hat{s}^2). \tag{2.4}$$

This estimator can also be obtained by considering an approximate processes where frequency increments are normally distributed (Watterson, 1982) where it is the exact solution to an approximate process rather than an approximate solution to an exact process. In the diploid case, if the frequency of the a allele is f_t then the expected frequency of heterozygotes is $2f_t\left(1 - f_t\right)$ and Equation 2.4 is therefore simply the total change in allele frequency, divided by the half the sum of the expected heterozygosity over all generations. It is interesting therefore to investigate the behaviour of this quantity. This was first described by Haldane (1956) for the case of $s \neq 0$ and $h = \frac{1}{2}$ but the more general derivation we give here is due to Maruyama and Kimura (1971) . Consider the continuous time diffusion limit. Then the equivalent quantity is $\int_{t=0}^{T} f_t(1 - f_t)dt$. Denote this quantity by H and define H_x to be $H|F_T = x$ and $H_x^{(n)}(p) = \mathbb{E}\left[(H_x)^n \,|f_0 = p\right]$ to be the expectation of the n^{th} moment of H conditional on $f_0 = p$ and $f_T = x$. showed that if $s = 0$ then $H_1^{(1)}(p) = \frac{4N_e}{3}\left(1 - p^2\right)$. Though they do not give the explicit result in the case $s \neq 0$, they do show that

$$H_1^{(1)}(p)u(p) = \left(1 - u(p)\right)\int_0^p \psi_K(f)u(f)df + u(p)\int_p^1 \psi_K(f)\left(1 - u(f)\right)df \tag{2.5}$$

where

$$G(p) = e^{-4Nsp} \tag{2.6}$$

$$u(p) = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}} \tag{2.7}$$

$$\psi_K(p) = \frac{2}{s}\left(e^{4Nsp} - 1\right) \tag{2.8}$$

which can be easily solved to give

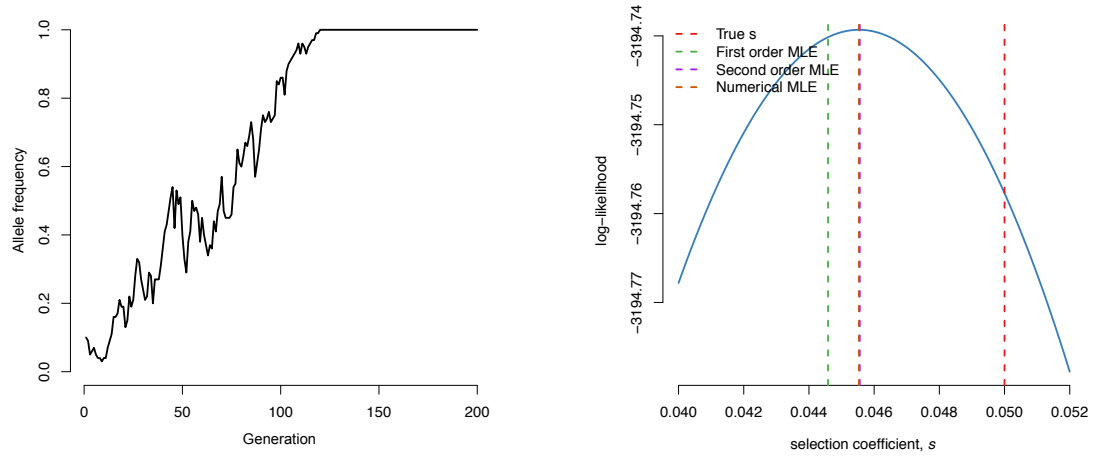
$$H_1^{(1)}(p) = \frac{\left((1 - p)\left(1 - e^{-4Ns(p+1)}\right) + (1 + p)\left(e^{-4Ns} - e^{-4Nsp}\right)\right)}{s\left(1 - e^{-4Ns}\right)\left(1 - e^{-4Nsp}\right)} \tag{2.9}$$

and therefore that

$$\lim_{N_e \to \infty} H_1^{(1)}(p) \to \frac{1 - p}{s}. \tag{2.10}$$

Maruyama and Kimura (1975) gives the more general result for $x \neq 1$. They give the result for $s = 0$ that $H_x^{(1)}(p) = \frac{4N_e}{3}\left(2x - x^2 - p^2\right)$ and their method can also be used to derive the general result for $s \neq 0$. This expression is extremely lengthy, and we do not give it here, but the limiting value for large N_e is given by

$$\lim_{N_e \to \infty} H_x^{(1)}(p) = \frac{x - p}{s}. \tag{2.11}$$



(a) (b)

Figure 2.1: (a) A simulated Wright-Fisher allele frequency trajectory with $2N_e = 100$, $s = 0.05$, $f_0 = 0.1$, $T = 200$ and (b) the results of applying our estimators to it showing the log-likelihood and MLE. The second order MLE is almost exactly on top of the numerical maximum (error $\approx 10^{-5}$).

---

which is consistent with our approximate MLE. These methods can also be used to derive the variance of H, though again, these expression are extremely lengthy and not particularly informative about the behaviour of the estimator.

An obvious question is whether we can derive a more accurate estimator from Equation 2.3. In fact, by expanding to second order in $\hat{s}$ and solving, we obtain

$$\hat{s} = \frac{k_1 - \sqrt{k_1^2 - 4k_2\left(f_T - f_0\right)}}{2k_2} + O(\hat{s}^3) \text{ where } k_i = \sum_{t=0}^{T-1} \left(f_t\right)^i \left(1 - f_t\right). \tag{2.12}$$

This expression seems not to have a simple interpretation, compared to Equation 3. It is slightly more accurate (Figure 2.11).

As mentioned above, an alternative way to obtain the estimator in Equation 2.4 is to consider an approximate process where frequency increments are normally distributed (Watterson, 1982). This makes it easier to consider the case of a general dominance parameter h. Setting $h_t = f_t + h\left(1 - 2f_t\right)$, the log-likelihood of the observations is given by

$$\ell(s) = 2N_e \sum_{t=1}^{T} \frac{\left(f_t - f_{t-1} - 2sf_{t-1}\left(1 - f_{t-1}\right) h_{t-1}\right)^2}{f_{t-1}\left(1 - f_{t-1}\right)}. \tag{2.13}$$

Differentiating with respect to s and setting equal to zero gives the MLE

$$\hat{s} = \frac{\sum_{t=0}^{T-1} h_t\left(f_{t+1} - f_t\right)}{2\sum_{t=0}^{T-1} h_t^2 f_t\left(1 - f_t\right)} \tag{2.14}$$

where $h_t = f_t + h\,(1 - 2f_t)$. Note that this reduces to Equation 2.4 when $h = \frac{1}{2}$ since then $h_t = \frac{1}{2}\ \forall t$. In practice the normal approximation is poor when f_t is close to 0 or 1, but the exact binomial result tells us that the error in the estimator is at most $O(S^2)$.

2.3 Structured population maximum likelihood estimators

In this section we consider estimating selection coefficients in a population distributed over a discrete lattice. We derive analogous results to the previous section for MLEs for the selection coefficient in this model.

2.3.1 Model and notation

Consider a lattice model consisting of K^2 single populations each of size $2N_e$, arranged in a $K \times K$ grid (Figure 2.2). Each deme has either two, three or four neighbouring demes, depending on where it is located on the grid. At each generation, from each population, M individuals migrate to each neighbouring deme. We also define the proportional migration rate $m = \frac{M}{2N_e}$. Note that this is a closed space rather than a torus so individuals at the edges do not "wrap" round to the other side of the grid. We index the demes by $i, j \in \{1, \dots K\}$ and write κ_{ij} for the set of indices of demes which neighbour deme i, j. The frequency of the selected allele in population i, j at generation t is f_t^{ij}. The migration rate is constant over all demes and time. The selection coefficients are constant over time, but not necessarily across demes. We write s^{ij} for the selection coefficient in deme i, j. We also write n_t^{ij} for the size of the sample taken from deme i, j at generation t and a_t^{ij} for the number of that sample of the selected type. We allow the dominance parameter h to be $\neq \frac{1}{2}$, but we assume it is constant across demes.

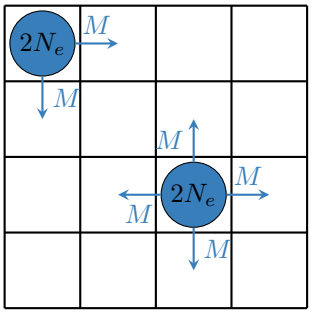


Figure 2.2: The Wright-Fisher lattice model. Shown for $K = 4$. Each deme has a constant population size of $2N_e$ and in each generation, exactly M individuals migrate to each of the neighbouring demes.

Writing down the exact likelihood in this model is impractical, since we would have to sum over all possible migrations. However as in the single population case, we can write down approximate models based on normal distributions of frequency increments. We first consider the case where $h = \frac{1}{2}$. Then assume that f_t^{ij} is normally distributed with mean

μ_t^{ij} and variance $\left(\sigma_t^{ij}\right)^2$ and

$$\mu_t^{ij} = \left(1 - m|\kappa_{ij}|\right) f_{t-1}^{ij} + s^{ij} f_{t-1}^{ij}\left(1 - f_{t-1}^{ij}\right) + m \sum_{i',j' \in \kappa_{ij}} f_{t-1}^{i'j'} \tag{2.15}$$

$$\text{and } \left(\sigma_t^{ij}\right)^2 = \frac{f_{t-1}^{ij}\left(1 - f_{t-1}^{ij}\right)}{2N_e}. \tag{2.16}$$

Here we are making four approximations. First, we are modelling allele frequency changes as normally rather than binomially distributed. This approximation is valid in the diffusion limit of large N_e and intermediate f_t. Second, we are ignoring the contribution of selection and migration to the variance of the change in frequency. The contribution from selection disappears in the diffusion limit. The contribution from migration is of order m times the difference in frequency between neighbouring demes. This disappears in the limit of an infinite number of demes, as long as allele frequencies vary continuously in space. Third, we are assuming that selection and migration are independent, so there is no $s^{ij}m$ term in μ_t^{ij}. Finally, we are assuming that the frequency changes in one generation are independent across demes, where in fact changes in neighbouring demes are negatively correlated due to the conservative migration. We could relax some or all of these assumptions and obtain a process closer to the "true" binomial process. However we prefer this form because it leads to a particularly simple derivation of the MLE, as we will see in the next section.

Since we allow the selection coefficients to vary arbitrarily, this model can exhibit complex behaviour (Figure 2.3). In particular, if some of the s^{ij} are positive and some are negative, then the allele frequencies can reach an internal (i.e. not at frequency 0 or 1) equilibrium state[†]. This can only happen in the single population model if the population is diploid and the allele is overdominant ($h > 1$). In this case, we say the allele is under balancing selection, and the equilibrium value of f is given by $\frac{h}{2h-1}$. The allele frequency trajectory is just one of several quantities which look very similar in model of population subdivision compared to models of overdominance (Schierup and Charlesworth, 2000; Nordborg, 1997). We can calculate the equilibrium frequency values $\bar{f}^{ij}$ by setting Equation 2.15 equal to zero,

$$s^{ij}\bar{f}^{ij}\left(1 - \bar{f}^{ij}\right) = m \sum_{i',j' \in \kappa_{ij}} \left\{\bar{f}^{ij} - \bar{f}^{i'j'}\right\} \ \forall i,j \tag{2.17}$$

and if there is a solution to these equations then it is a possible equilibrium value. Since the sum of the RHS of the above equation over all demes is 0, this implies that

$$\sum_{i,j} s^{ij}\bar{f}^{ij}\left(1 - \bar{f}^{ij}\right) = 0 \tag{2.18}$$

[†]If the migration rate is small, see Appendix A for an analysis of this effect in a two-deme model.

which provides some intuition about the relationship between possible equilibrium values. Note that if the s^{ij} are all of the same sign then the only possible solution is for all of the $\bar{f}^{ij}$ to be 0 or 1 so there is no internal equilibrium. If all the $S_ij = 0$ then there is an (unstable) equilibrium for all f_{ij}.

2.3.2 Maximum likelihood estimators

Given the model defined in the previous section, the distribution of allele frequency increments is Gaussian, and we can write the log-likelihood for the trajectory as

$$\ell(\mathbf{s}, m) = \sum_{t=1}^{T} \sum_{i,j=1}^{k} \frac{\left(f_t^{ij} - \mu_t^{ij}\right)^2}{\left(\sigma_t^{ij}\right)^2} \tag{2.19}$$

where μ_t^{ij} and σ_t^{ij} are as defined in Equation 2.15. Equation 2.19 is quadratic in s^{ij} and m, and therefore has a unique solution. We can solve for the MLE of s^{ij} with m known:

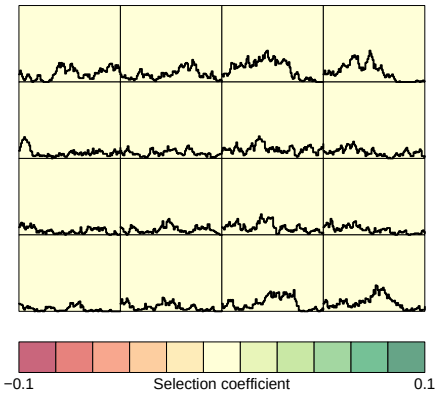
$$\hat{s}^{ij} = \frac{f_T^{ij} - f_0^{ij}}{\sum_{t=0}^{T-1} f_t^{ij}\left(1 - f_t^{ij}\right)} + m\left(\frac{\sum_{t=0}^{T-1}\left(|\kappa_{ij}| f_t^{ij} - \sum_{i',j' \in \kappa_{ij}} f_t^{i'j'}\right)}{\sum_{t=0}^{T-1} f_t^{ij}\left(1 - f_t^{ij}\right)}\right). \tag{2.20}$$

Notice that the first term in Equation 2.20 is the same as that in Equation 2.4, which is the estimator for the selection coefficient if we ignored the other demes. The second term is a correction for the migration from other demes. If the allele frequency in neighbouring demes is higher than that in the deme ij then our estimate of s^{ij} is reduced since some of the increase in allele frequency is likely due to migration rather than selection. Similarly, we can obtain the MLE for m with s^{ij} known:

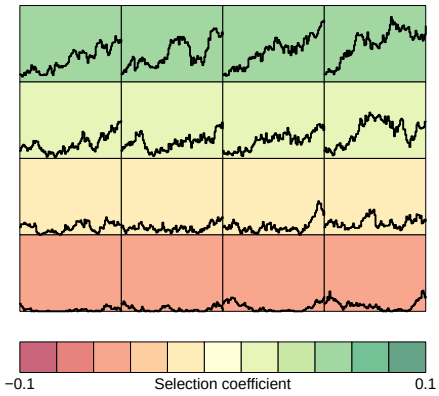
$$\hat{m} = \frac{\sum_{t=1}^{T}\sum_{i,j=1}^{K}\left\{\frac{\left(s^{ij} f_{t-1}^{ij}\left(1-f_{t-1}^{ij}\right)+f_{t-1}^{ij}-f_t^{ij}\right)\left(|\kappa_{ij}| f_{t-1}^{ij} - \sum_{i',j' \in \kappa_{ij}} f_{t-1}^{i'j'}\right)}{f_{t-1}^{ij}\left(1-f_{t-1}^{ij}\right)}\right\}}{\sum_{t=1}^{T}\sum_{i,j=1}^{K}\left\{\frac{\left(|\kappa_{ij}| f_{t-1}^{ij} - \sum_{i',j' \in \kappa_{ij}} f_{t-1}^{i'j'}\right)^2}{f_{t-1}^{ij}\left(1-f_{t-1}^{ij}\right)}\right\}}. \tag{2.21}$$

This expression does not have a simple interpretation like Equation 2.20 but the numerator is close to the covariance of the observed movements in allele frequency with the expected change due to migration (weighted by $f(1-f)$). That is, if the observed changes in allele frequency were uncorrelated with the relative frequencies in neighbouring demes, we would estimate the migration rate to be zero. Conversely, if the allele frequency always increases when neighbouring demes have higher frequency, we estimate the migration rate to be large.

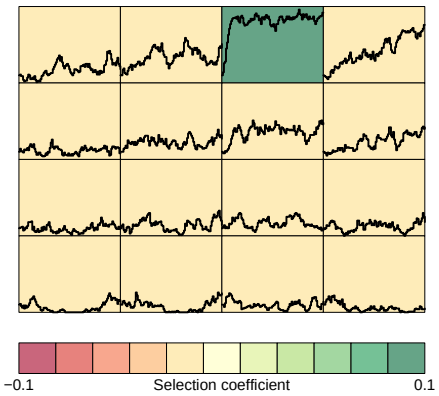
If both s^{ij} and m are unknown, we could either solve for the maximum of Equation 2.19 directly or iterate computation of the two individual estimators. Note that the likelihood is flatter when there are more observations at very low or very high frequencies. Figure



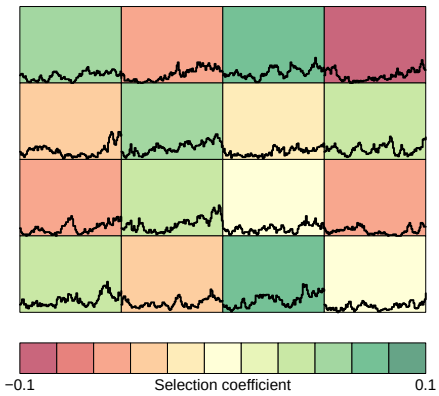
(a) The same selection coefficient in every deme. No selection and neutral drift. The allele will eventually drift to fixation or extinction.



(b) A north-south gradient of selection coefficients. In this case, allele frequencies reach an internal equilibrium.



(c) $s < 0$ everywhere apart from one deme where it is positive. Here a selective advantage in one deme keeps the allele from dying out, despite being selected against elsewhere.



(d) We allow for arbitrary distributions of selection coefficients. Though it is not obvious from this plot, there is still an equilibrium value for the frequency.

Figure 2.3: Examples of allele frequency trajectories in the Wright-Fisher lattice model. Here $K = 4$ and the black line in each square shows the allele frequency in that deme for 100 generations. Here $2N_e = 100$, $m = 0.04$, $f_0 = 0.1$ and the s^{ij} are between ± 0.1.

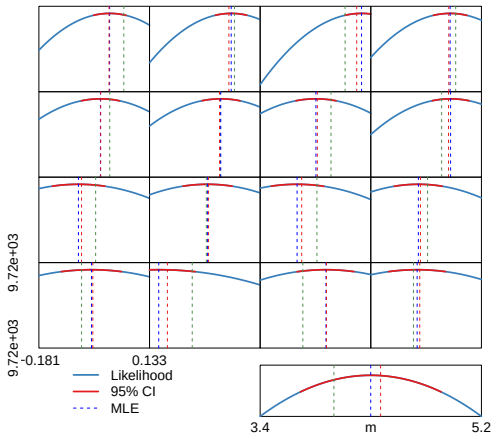


Figure 2.4: Illustrating the MLE in the lattice model. The main plot shows the log-likelihood (Equation 2.19) for a simulated trajectory from a population with the North-South selection gradient illustrated in Figure 2.3b. Each square shows the likelihood as a function of s^{ij} for that deme, with all the other demes held constant at the MLE. The red line shows the approximate 95% confidence interval ($\Delta\ell = 1.92$) for each s^{ij}. The green dashed line shows the true value of s^{ij}, the blue line shows the numerically maximised MLE and the red line shows the MLE computed using equations 2.20 and 2.21. The small panel below shows the likelihood for m, when all the s^{ij} are held constant at their MLEs.

2.4 shows the likelihood surface for a simulation from the selection coefficients shown in Figure 2.3b. Because we took $f_0^{ij} = 0.1$, the southernmost demes with negative selection coefficients have only observations close to 0 and the likelihood is flatter in those demes.

If $h \neq \frac{1}{2}$ then the MLEs for s^{ij} and m are

$$\hat{s}^{ij} = \frac{\sum_{t=0}^{T-1} h_t^{ij}\left(f_{t+1}^{ij} - f_t^{ij}\right)}{2\sum_{t=0}^{T-1}\left(h_t^{ij}\right)^2 f_t^{ij}\left(1-f_t^{ij}\right)} + m\left(\frac{\sum_{t=0}^{T-1} h_t^{ij}\left(|\kappa_{ij}|f_t^{ij} - \sum_{i',j'\in\kappa_{ij}} f_t^{i'j'}\right)}{2\sum_{t=0}^{T-1}\left(h_t^{ij}\right)^2 f_t^{ij}\left(1-f_t^{ij}\right)}\right) \tag{2.22}$$

$$\hat{m} = \frac{\sum_{t=1}^{T}\sum_{i,j=1}^{K}\left\{\frac{\left(2h_{t-1}^{ij}s^{ij}f_{t-1}^{ij}\left(1-f_{t-1}^{ij}\right)+f_{t-1}^{ij}-f_t^{ij}\right)\left(|\kappa_{ij}|f_{t-1}^{ij}-\sum_{i',j'\in\kappa_{ij}} f_{t-1}^{i'j'}\right)}{f_{t-1}^{ij}\left(1-f_{t-1}^{ij}\right)}\right\}}{\sum_{t=1}^{T}\sum_{i,j=1}^{K}\left\{\frac{\left(|\kappa_{ij}|f_{t-1}^{ij}-\sum_{i',j'\in\kappa_{ij}} f_{t-1}^{i'j'}\right)^2}{f_{t-1}^{ij}\left(1-f_{t-1}^{ij}\right)}\right\}} \tag{2.23}$$

where $h_t^{ij} = f_t^{ij} + h\left(1 - 2f_t^{ij}\right)$. Finally, if we constrain s^{ij} to be a constant say $s^{ij} = \tilde{s}\ \forall i, j$ then

$$\hat{\tilde{s}} = \frac{\sum_{i,j=1}^{K}\left(f_T^{ij} - f_0^{ij}\right)}{\sum_{i,j=1}^{K}\sum_{t=0}^{T-1} f_t^{ij}\left(1-f_t^{ij}\right)} \tag{2.24}$$

which no longer depends on m. Notice that because the function $f(1-f)$ is concave, this estimate is smaller than the estimate that we would get by averaging the frequencies across all demes and applying Equation 2.4.

2.4 Estimation with incomplete observations

So far in this chapter we have assumed that we knew the exact allele frequency in every deme at every generation. As we pointed out earlier, this is almost never the case, and if

we want to make inference in a realistic setting, we need to develop methods which allow us to use data where we do not know the allele frequency exactly, and we do not have data for every generation. In practice, what we have available are samples of individuals from the populations and this estimates of the underlying allele frequency. In this section we develop a method to use this type of data, first in the single population and later in the lattice case. In both cases, we pose the problem as a hidden Markov model (HMM). The advantage of the HMM approach is that it allows us to make inference in settings where there is a long gap between observations or many missing observations, without having to approximate the transition density over gaps of many generations.

A hidden Markov model has the following structure. For discrete time intervals indexed by t, some unobserved (hidden) process f_t takes values in a countable set. f_t is a Markov process so, conditional on f_{t-1}, f_t is independent of $\{f_0 \dots f_{t-2}, f_{t+1} \dots\}$. We do not observe f_t, but we do make observations a_t, where the distribution of a_t is known, conditional on f_t. In order to fully specify the HMM, we need to specify the state space, the *emission probabilities* of observing some value of a_t given f_t and the transition probabilities of moving hidden states between t and $t+1$; $\mathbb{P}\{f_{t+1} = f | f\}$. Useful algorithms for investigating this model include the *Viterbi algorithm*, and the *forward* and *backward* algorithms (together the *forward-backward* algorithm). The Viterbi algorithm (Viterbi, 1967) finds the most likely hidden path $\{f_0 \dots f_T\}$, given observations $\{a_0 \dots a_T\}$. The forward algorithm finds the probability of being in any hidden state at time t, conditional on the observations up to that time; $\mathbb{P}\{f_t = f | a_0 \dots a_t\}$. Similarly, the backward algorithm finds the probability of being in any hidden state at time t, conditional on the observations after that time; $\mathbb{P}\{f_t = f | a_{t+1} \dots a_T\}$. Then the forward-backward algorithm combines these two probabilities to give the full conditional distribution of hidden states at each time $\mathbb{P}\{f_t = f | a_0 \dots a_T\}$. For a full description and further introduction to these algorithms, see Durbin *et al.* (1998).

2.4.1 Single population

In order to apply standard HMM theory, we discretise the allele frequency space, assuming that $f_t \in \mathcal{G} = \{g_0, \dots g_D\}$, and the interval between points $\delta g = g_{i+1} - g_i$ is constant for all i. We typically use a grid size of $D = 100$. We define the HMM as follows:

1. The hidden states are the frequencies f_t. The observations are the number of alleles of the selected type a_t. We assume that N_e is known and that we have an estimate of s, which we take as fixed for this iteration. The sample size n_t is known.

2. The emission probabilities are binomial: $a_t \sim \text{Bin}(n_t, f_t)$.

3. The transition probabilities are defined by integrating the approximate normal continuous transition density between the midpoints of the intervals of the discretised points:

$$\mathbb{P}\left\{f_{t+1} = g | f_t\right\} = \int_{g-\frac{\delta g}{2}}^{g+\frac{\delta g}{2}} \phi\left(\frac{x-\mu_t}{\sigma_t}\right) dx \tag{2.25}$$

$$\text{where } \mu_t = f_t + s f_t(1-f_t) \text{ and } \sigma_t^2 = \frac{f_t(1-f_t)}{2N_e}. \tag{2.26}$$

There are several ways we can proceed from here. Given a value of s, we can compute the likelihood of the observations, so we could just find the value of s which maximises this likelihood, either by searching, or by standard numerical maximisation techniques. However, this would become impractical in the structured case, and a more efficient way to find the MLE is with an EM algorithm where, at each iteration, we update the estimate of s using the value which maximises the expected log-likelihood under the posterior distribution on the hidden variables f_t, conditional on the previous estimate of s. Suppose at iteration r, we have an estimate s_r of s, then taking the expectation of Equation 2.2, expanding to first order in s and maximising yields the EM update rule for the next estimate of s, analogous to Equation 2.4,

$$s_{r+1} = \frac{\mathbb{E}\left[f_T\right] - \mathbb{E}\left[f_0\right]}{\Sigma_{t=0}^{T-1}\mathbb{E}\left[f_t\left(1-f_t\right)\right]} \tag{2.27}$$

where the expectations are taken with respect to the posterior distribution of paths, conditional on the observations, and the selection coefficient s_r. These posterior probabilities can be computed using the forward-backward algorithm. This expression is identical to Equation 2.4 but with expectations replacing the actual frequencies. Taking into account the discretisation of the frequencies, our algorithm is as follows:

1. *Initialisation:* Choose s_0 to be some reasonable starting value. We linearly interpolate the frequency estimates and apply Equation 2.4.

2. *Recursion:* Given an estimate s_r for s, apply the forward-backward algorithm to the HMM described above to compute the probabilities $p_t^g = \mathbb{P}\left(f_t = g | a_0, \ldots, a_T, s_r\right)$. Then set

$$s_{r+1} = \frac{\Sigma_{g\in\mathcal{G}}\left[p_T^g g\right] - \Sigma_{g\in\mathcal{G}}\left[p_0^g g\right]}{\Sigma_{t=0}^{T-1}\Sigma_{g\in\mathcal{G}}\left[p_t^g g\left(1-g\right)\right]}. \tag{2.28}$$

3. *Termination:* Stop when $|s_{r+1} - s_r| < \epsilon$ for some predetermined tolerance ϵ and set our estimate of s equal to s_{r+1}.

The algorithm also naturally computes (as part of the forward algorithm), the likelihood of the data at each iteration given the observations and the current parameter values. Using

the final likelihood, and the fact that the difference in likelihood between two models is asymptotically χ^2 distributed, we can compute confidence intervals and p-values against various null hypotheses for our estimates.

If the dominance parameter $h \neq 0.5$ then the update rule in Equation 2.27 becomes

$$s_{r+1} = \frac{\Sigma_{t=0}^{T-1}\mathbb{E}\left[h_t\left(f_{t+1} - f_t\right)\right]}{2\Sigma_{t=0}^{T-1}\mathbb{E}\left[h_t^2 f_t\left(1 - f_t\right)\right]} \tag{2.29}$$

where $h_t = f_t + h\left(1 - 2f_t\right)$. Since h_t depends on f_t, the numerator now contains a $\mathbb{E}\left[f_{t+1}f_t\right]$ term which makes this expression harder than Equation 2.28 to compute in the discretised model. Fortunately, using the forward and backward probabilities, it is possible to compute the conditional transition probabilities $q_t^{gg'} = \mathbb{P}\left(f_{t+1} = g' | f_t = g, a_0, \ldots, a_T, s_r\right)$ and using these, we compute $\mathbb{E}\left[f_{t+1}f_t\right]$ in the discretised step using $\Sigma_{g\in\mathcal{G}}\left[p_t^g g\left[\Sigma_{g'\in\mathcal{G}}\left[q_t^{gg'}g'\right]\right]\right]$. Equation 2.28 is then replaced by

$$s_{r+1} = \frac{h\Sigma_{g\in\mathcal{G}}\left[p_T^g g\right] - h\Sigma_{g\in\mathcal{G}}\left[p_0^g g\right] + (1-2h)\,\Sigma_{t=0}^{T-1}\left[\Sigma_{g\in\mathcal{G}}\left[p_t^g g\left[\Sigma_{g'\in\mathcal{G}}\left[q_t^{gg'}g'\right]\right] - g\right]\right]}{2\Sigma_{t=0}^{T-1}\Sigma_{g\in\mathcal{G}}\left[p_t^g h(g)^2 g\left(1-g\right)\right]} \tag{2.30}$$

where $h(g) = g + h\left(1 - 2g\right)$.

In order to investigate the performance of this estimator, we simulated observations from a Wright-Fisher population for a variety of parameter values and evaluated the performance of the estimator under different conditions (Figure 2.5).

First (Figure 2.5a) we investigated the effect of changing the sample size while the frequency of sampling remained constant. As expected, the error decreases with sample size, though for all effective population sizes, the expected error remains more or less constant above a given sample size. This constant error is larger when the effective population size is smaller. In Figure 2.5b, we show the effect of changing the frequency of sampling, ranging from sampling just the start and end points to sampling at every generation, while keeping the sample size fixed. We see that the error decreases as the sampling frequency increases, but this really only makes a difference when the effective population size is small. For $2N_e = 10^4$, for example, changing the sampling frequency from every twenty generations to every generation makes virtually no difference to the error. One caveat is that this result relies on the start and end points being at some intermediate (between 0 and 1) frequency. If all we observed was that $f_0 = 0$ and $f_T = 1$ for some T, then it would be impossible to make a sensible estimate of s. We can see this further by varying the initial frequency (Figure 2.5c). Conditional on eventual fixation, the error increases as the initial frequency increases, demonstrating that it is the observations at intermediate allele frequencies that give us precision in our estimates. Again, this is particularly true when the effective population size is small.

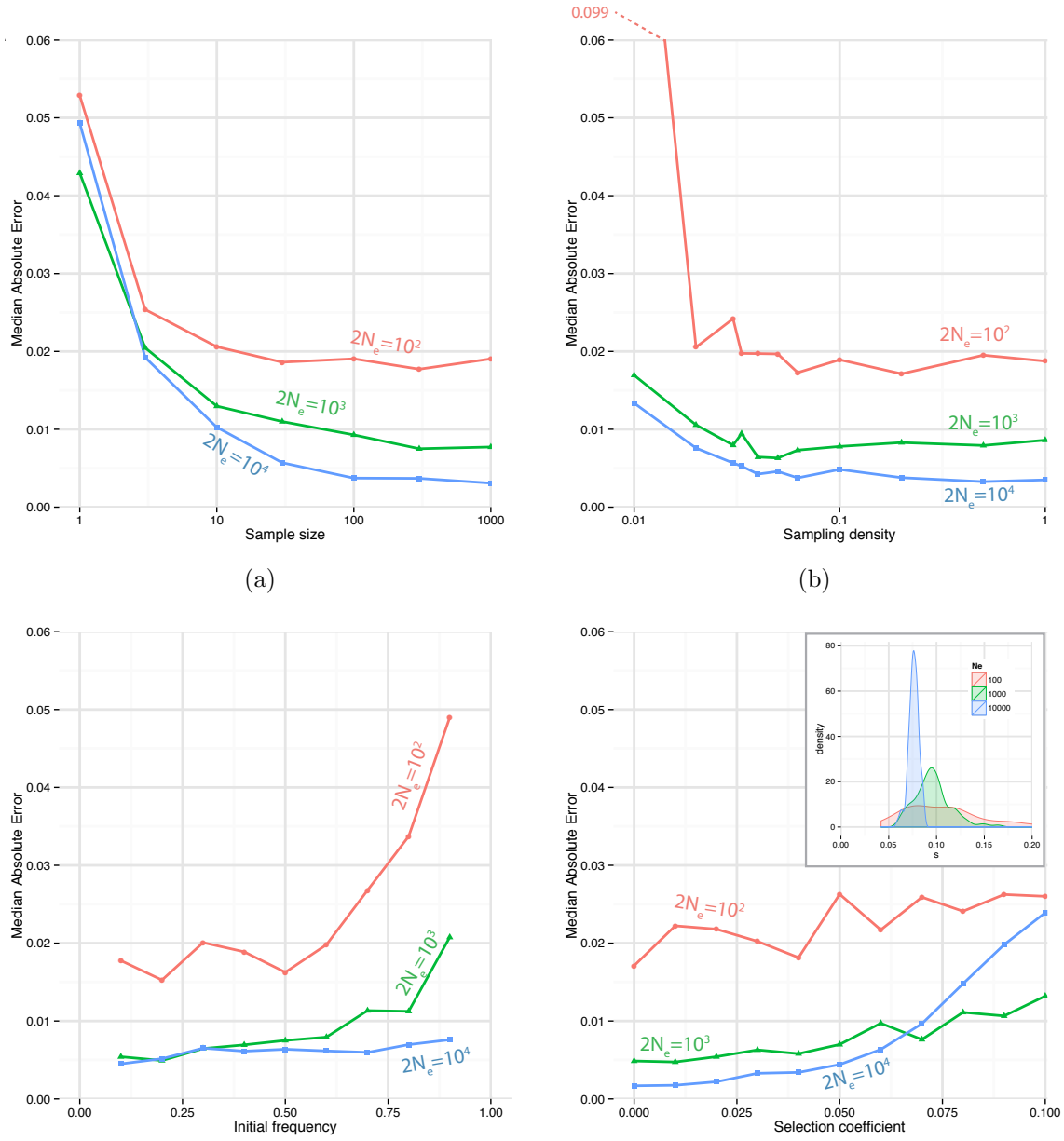


Figure 2.5: Performance of the single population estimator. **a-d**: Median absolute error of estimates of s, for a range of different parameter values. In each case, results are shown for effective population sizes $2N_e = 10^2, 10^3$ and 10^4. Simulations were performed in a standard Wright-Fisher model and each point is the median of 100 independent simulations. If not otherwise specified, parameters are constant as follows: initial frequency $f_0 = 0.5$, number of generations $T = 100$, selection coefficient $s = 0.05$, samples taken every 10 generations, size of each sample = 100. We stopped the algorithm when successive estimates of s differed by less that $\epsilon = 10^{-3}$. **a**: The size of the sample varies from 1 to 1000. **b**: The frequency of sampling varies from 0.01 (once every hundred generations, i.e. two observations), to 1 (every generation). **c**: The initial frequency f_0 varies from 0.1 to 0.9. **d**: The selection coefficient s varies from 0 to 0.1. **Inset**: density of $\hat{s}$ for $s = 0.1$.

We also investigated the effect that s has on the error (Figure 2.5d). As s increases, the expected error increases, for all population sizes, though the relative error is decreasing. For large N_e and large s, the estimator begins to perform poorly because although the variance of $\hat{s}$ decreases, the bias increases (Figure 2.5d inset). The bias comes from the fact that our estimator is only accurate to $O(s^2)$.

Overall, it seems that the main determinant of the accuracy of the estimator is the effective population size of the underlying population and that, provided we have a sufficiently large population and at least some observations at intermediate allele frequencies, we require neither large nor frequent samples.

Finally, we checked whether the discretisation and approximate transition density had a large influence on the result. In the single population model, if we set $D = 2N_e + 1$ and use the exact binomial transition probabilities (Equation 2.1) in the HMM rather than the approximate normal transition probabilities in Equation 2.26, then our model is exactly the one from which we simulated. We compared the estimates from this exact model to those from our approximate model. Using the same parameters as in Figure 2.5d, we found that the error increased with s, though modestly. For illustration, when $s = 0$ the expected error in s due to discretisation was $\approx 3 \times 10^{-5}$ and when $s = 0.1$ the expected error was $\approx 3 \times 10^{-3}$.

2.4.2 Structured population

Directly extending the EM algorithm from the single population case to the structured case is impractical for two reasons. First, the likelihood depends on both the s^{ij} and m, making the EM step difficult to calculate. Second, the state space of the HMM increases rapidly with the number of demes. If there are K^2 demes and D discretised states then the full HMM has D^{K^2} states, making it impractical to compute for anything but the smallest number of demes. Therefore we present an algorithm which makes two approximations to make the solution tractable. First, we update s^{ij} and m separately. The update rule for s^{ij} has the form of the EM update rule from the single population case and the update rule for m can be computed similarly. Second, when updating any frequency in any particular deme, we assume the allele frequencies in all the other demes are fixed to their most likely values from the previous iteration, which makes the HMM calculations independent across demes, reducing the complexity to DK^2.

As in the single population case, we discretise frequency space so that $f_t^{ij} \in \{g_0, \dots g_D\}$. We can then write down the emission probabilities as before, $a_t^{ij} \sim \text{Bin}\left(n_t^{ij}, f_t^{ij}\right)$. These are independent across demes. As mentioned above, in order to reduce the complexity, we look at each deme in turn, and assume that the allele frequency in all the other demes is

fixed at the frequency from the previous iteration of the algorithm. In other words, we update the frequency trajectory in each deme independently in turn, conditional on all the others, rather than updating them all simultaneously. We define the HMM for each deme as in the single population case, except that we set the mean in Equation 2.26 to

$$\mu_t^{ij} = (1 - m|\kappa_{ij}|)\, f_{t-1}^{ij} + s^{ij} f_{t-1}^{ij} \left(1 - f_{t-1}^{ij}\right) + m \sum_{i',j' \in \kappa_{ij}} \tilde{f}_{t-1}^{i'j'} \tag{2.31}$$

where $\tilde{f}_t^{ij}$ is a fixed parameter. This is identical to the true mean from Equation 2.15, except that $\tilde{f}_t^{ij}$ has replaced f_t^{ij}, which is what makes the demes independent. Analogous to Equation 2.27, the update rule for s^{ij} is given by

$$s_{r+1}^{ij} = \frac{\mathbb{E}\left[f_T^{ij}\right] - \mathbb{E}\left[f_0^{ij}\right]}{\Sigma_{t=0}^{T-1} \mathbb{E}\left[f_t^{ij}\left(1 - f_t^{ij}\right)\right]} + m_r \left(\frac{\sum_{t=0}^{T-1} |\kappa_{ij}| \mathbb{E}\left[f_t^{ij}\right] - \sum_{i',j' \in \kappa_{ij}} \tilde{f}_t^{i'j'}}{\sum_{t=0}^{T-1} \mathbb{E}\left[f_t^{ij}\left(1 - f_t^{ij}\right)\right]} \right) \tag{2.32}$$

where the expectations are taken with respect to the posterior distribution of allele frequencies conditional on the current estimate of s^{ij} and m, denoted s_r^{ij} and m_r. They can be computed over the discretised values of f_t using a similar expression to Equation 2.28. Similarly, if m is not known, the EM update rule is

$$m_{r+1} = \frac{\sum_{t=1}^{T} \sum_{i,j=1}^{K} \mathbb{E}\left[\frac{\left(s^{ij} f_{t-1}^{ij}\left(1 - f_{t-1}^{ij}\right) + f_{t-1}^{ij} - f_t^{ij}\right)\left(|\kappa_{ij}| f_{t-1}^{ij} - \sum_{i',j' \in \kappa_{ij}} \tilde{f}_{t-1}^{i'j'}\right)}{f_{t-1}^{ij}\left(1 - f_{t-1}^{ij}\right)} \right]}{\sum_{t=1}^{T} \sum_{i,j=1}^{K} \mathbb{E}\left[\frac{\left(|\kappa_{ij}| f_{t-1}^{ij} - \sum_{i',j' \in \kappa_{ij}} f_{t-1}^{i'j'}\right)^2}{f_{t-1}^{ij}\left(1 - f_{t-1}^{ij}\right)} \right]}. \tag{2.33}$$

Note that this estimator can be negative which, though it makes sense within the model with small Gaussian updates defined in Equation 2.15, does not have a sensible interpretation. We allow it to be negative at intermediate steps of the algorithm, but if the final estimate is negative, we set it to zero. The algorithm proceeds as follows:

1. *Initialisation:* Compute an initial guess for the f_t^{ij}, by taking the observed frequencies and linearly interpolating missing values. Call this $\tilde{f}_t^{ij}$, and make initial guesses for s^{ij} and m.

2. *Recursion:* Given estimates s_r^{ij}, m_r, and $\tilde{f}_t^{ij}$ for s^{ij} and m, solve the HMM for each deme as in the single population case. Compute the posterior probabilities $p_{ij,t}^g = \mathbb{P}\left(f_t^{ij} = g | a_0^{ij} \ldots a_t^{ij}, s^{ij}\right)$ as before, and the most likely path v_t^{ij}, using the Viterbi algorithm. Compute new estimates s_{r+1}^{ij} and m_{r+1} using the EM update rules above, and set $\tilde{f}_{t+1}^{ij} = v_t^{ij}$.

3. *Termination:* Stop when the change in log-likelihood between successive iterations is less than some specified amount ϵ.

Note that the calculation for each of the K^2 demes is independent, so it would be easy to parallelise this computation and compute the recursion step for each deme on a separate core.

Though we refer to the "likelihood" in this model, it is worth noting that this is not in fact a real likelihood, since we are effectively treating $\tilde{f}_t^{ij}$ as a parameter rather than (missing) data. Writing $\tilde{f}_{t,r}^{ij}$ for the r^{th} iteration of $\tilde{f}_t^{ij}$, and the likelihood ℓ as a function of s_r^{ij}, m_r and $\tilde{f}_t^{ij}$, we guarantee that $\ell\left(s_{r+1}^{ij}, m_{r+1}, \tilde{f}_{t,r}^{ij}\right) \geq \ell\left(s_r^{ij}, m_r, \tilde{f}_{t,r}^{ij}\right)$, but not that $\ell\left(s_{r+1}^{ij}, m_{r+1}, \tilde{f}_{t,r+1}^{ij}\right) \geq \ell\left(s_{r+1}^{ij}, m_{r+1}, \tilde{f}_{t,r}^{ij}\right)$, Consequently there is no guarantee that one full iteration of the algorithm will actually increase the likelihood, so this algorithm does not produce an MLE. In practice however, it seems to perform acceptably, a fact we verified using simulation.

As with the single population case, we tested this estimator by simulating observations from the structured Wright-Fisher model. We found that, as in the one dimensional case, the estimates are more accurate the more of the trajectory we see. When we set $f_0 = 0.5$ so that each deme saw roughly the same change in trajectory, we found that our estimates for both s and m were unbiased (Figure 2.6a), but that when we set f_0 to 0.1, so that we saw less of the trajectory in demes with lower selection coefficients, we found that our estimates of the low selection coefficients were significantly worse (Figure 2.6b), consistent with what we would expect from the results in Figure 2.5c. We assumed that we could guess an initial value for m within 0.01 of the true value. If we set the initial value for m much further from the true value, then the estimator performed poorly.

We investigated the performance of the estimator for different values of m (Figures 2.6c-d). As m increased, the error in our estimates of s and m increased. We also investigated the error in our estimates of s when m was known and fixed (Figure 2.6d). In this case, there is a modest improvement in accuracy, particularly for small m (comparing Figures 2.6c and 2.6d).

In general, our estimates of s perform better than our estimates of m. In practice we have assumed that we have a reasonably good prior on the value of m, whereas we have made no such assumptions about s. In particular, if our initial value of m is far from the true value, then we noticed that the algorithm can sometimes fail to converge to the correct value. In this case, because estimates of s and m are correlated, the estimates of s are biased. In particular, if m is too small, then the estimates of s^{ij} shrink towards their mean. To see why this is true, consider the limiting behaviour of the allele frequencies in the lattice case. As we mentioned earlier (Equation 2.18), in the lattice model it is possible to reach

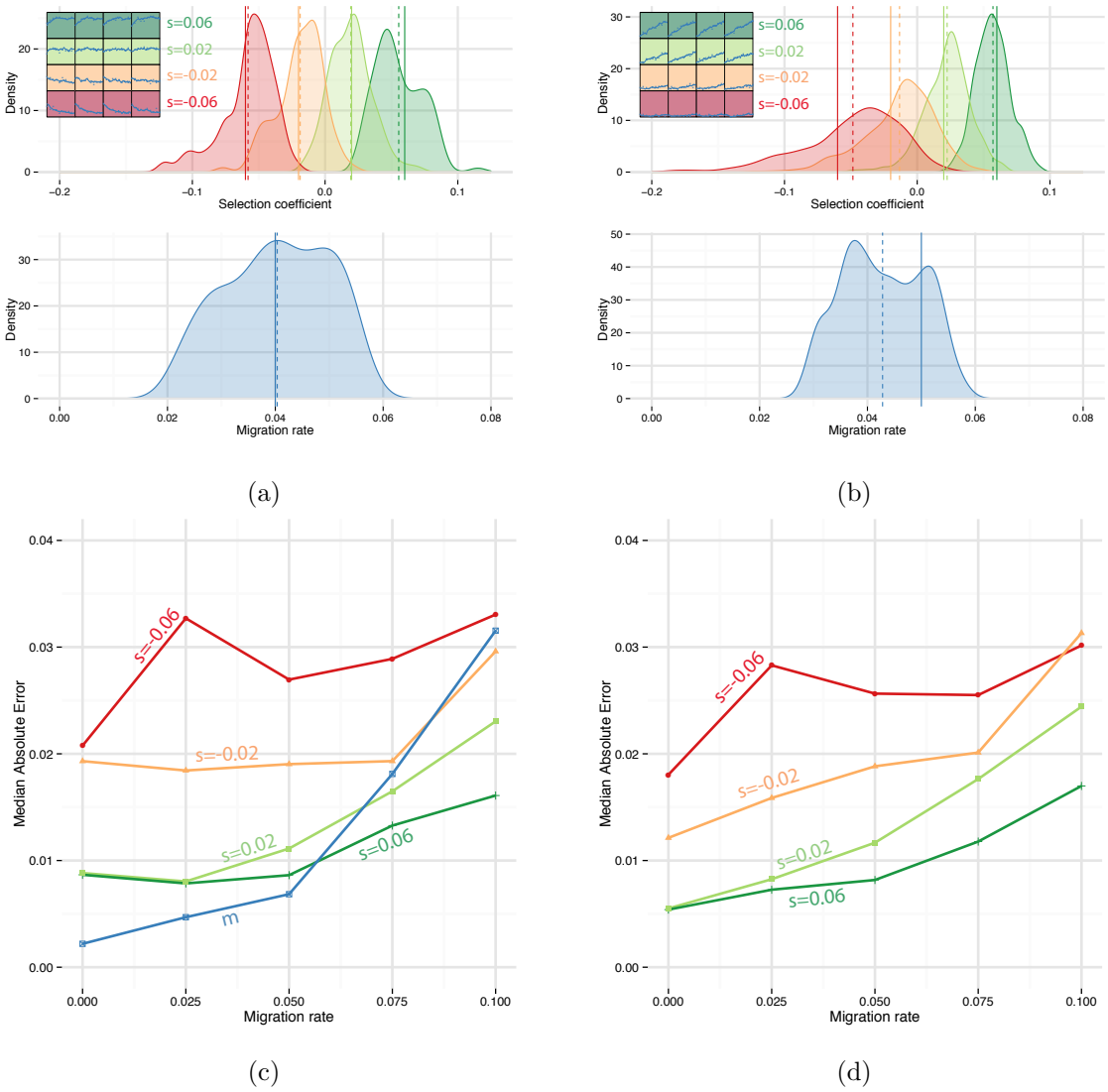


Figure 2.6: Performance of the structured population estimator. We simulated observations from the structured Wright-Fisher model described in the main text, with 16 demes. Here $2N_e = 1000$, s ranges across space from -0.06 to +0.06, m ranges from 0 to 0.1, samples of size 100 are taken every 10 generations. The algorithm is terminated when successive log-likelihoods differ by less than 0.1. The initial value for m is uniformly distributed on $[m - 0.01, m + 0.01]$. **a**: Density plots of the results of 100 simulations, with $m = 0.04$ and $f_0 = 0.5$. Solid vertical lines show the true values, and dashed lines the mean of the density **Upper panel**: Density of estimates of s. We combined the results across all demes with the same values of s, so the dark green density shows the results for 400 observations, i.e. 4 demes in each of 100 simulations. The inset shows the spatial distribution of selection coefficients, and an example path. **Lower panel**: Density of estimates of m. **b**: As **a** except $f_0 = 0.1$ and $m = 0.05$. **c-d**: Median absolute error of the estimates of s in each deme, with different lines for each true value of s, for different values of m. $f_0 = 0.1$. **c**: Error when m is unknown (but guessed to within 0.01), including error in m. **d**: Error in s when m is known and fixed.

an internal equilibrium where

$$\sum_{i,j} s^{ij} \bar{f}^{ij} \left(1 - \bar{f}^{ij}\right) = 0.$$

However, this does not fully determine all the s_{ij} and m. It determines the s^{ij} relative to each other but to know the absolute values, we need to know m and there is no information about m in these equilibrium values. Power to estimate m comes from observing fluctuations around the equilibrium value, but when $2N_e$ is large, these fluctuations are small, and if the sample size is small then the fluctuations due to sampling error are much greater than those due to changes in allele frequency. Therefore the best chance of estimating m using this method might be when we have very large samples from a very small population, a situation which is rarely encountered. Fortunately for most applications there are likely to be independent estimates of m, which we can use as starting points. These estimates might come from direct estimates of migration rates, for example using capture-recapture techniques, or historical data. On the other hand, if we had data for multiple alleles, some of which we knew not to be under selection, we could just fit our model to those loci, fixing $s = 0$.

2.5 Selection on colour morphs in moths

In this section, we demonstrate our method on data on colour morphs of two common British moth species, one (*Panaxia dominula*) in a small and apparently unstructured population, and one (*Biston betularia*) in a spatially structured setting. Large time-serial datasets of allele frequencies in natural populations are relatively rare and both these species are unusual in having had so many observations collected. Another advantage of these species is that they are known to have exactly one generation per calendar year, and all the living adult moths die every winter, so not only are they actually quite close to a Wright-Fisher population, but we do not have to work to convert observation time into generation time.

2.5.1 *Panaxia dominula*

For a dataset of allele frequencies in a single, closed population, we used observations of frequencies of the *medionigra* morph in a population of *Panaxia dominula* (scarlet tiger moth) at Cothill Fen near Oxford. Observations of this colony were first made in 1939 by E.B. Ford and R.A. Fisher and were collected annually, with some gaps, until at least 1999. The data are reported in Cook and Jones (1996) and Jones (2000) and have most recently been analysed by O'hara (2005). *P. dominula* is a colourful day-flying moth, which has exactly one generation per calendar year. The *medionigra* morph (Figure 2.7b)



(a) *typica* (b) *medionigra* (c) *bimacula*

Figure 2.7: The three morphs of *P. dominula*, corresponding to **a**: wild type, **b**: heterozygote and **c**: homozygote. Pictures taken from Fisher and Ford (1947).

---

is the result of a heterozygous polymorphic allele and is observed as a reduction of the size and number of white spots on the moths wings. A moth that is homozygous for the variant allele is much darker, but this *bimacula* morph (Figure 2.7c) is much rarer and almost never observed. When the study was started, the *medionigra* morph was present in the Cothill colony at a frequency of around 10%, but subsequently dropped sharply. The question of whether this rapid decline in frequency represented the effect of natural selection was the subject of a spirited debate between Fisher and Wright. Fisher held, assuming an effective population size of approximately 1,000 that the decline could only be due to selection whereas Wright argued that the effective population size might be much smaller and that the decline would then be consistent with drift (Fisher and Ford, 1947; Wright, 1948). We made the assumption that the effective population size is constant. As part of the collection of moth frequencies, the population size has been estimated using capture-recapture methods. Estimates have ranged from 216 to 60,000 Cook and Jones (1996). Even if the census population really took this range, it is not clear what relation this has to the effective population size. We therefore assumed that the effective population size, $2N_e$ was constant, but checked whether different values of N_e had an effect on our results.

The data and model fit are illustrated in in Figure 2.8, along with the likelihood surface for the true allele frequency trajectory, and the likelihood function for s. We stopped the algorithm when successive iterations of s differed by less that 10^{-3}. Taking $2N_e = 1,000$, we estimate a selection coefficient of -0.057, though with a fairly wide 95% confidence interval of (-0.113, -0.003). We thus reject the null hypothesis that $s = 0$ at the 5% level, but only just (twice the change in log-likelihood, $2\Delta\ell = 4.2$, approximate χ^2_1 p-value=0.04) and more or less agree with the original conclusion of Fisher and Ford (1947) that "the observed fluctuations in gene frequency are much greater than could be ascribed to random survival only". Our estimate is similar to the estimate of given by Cook and Jones (1996), and is consistent with results from other *P. dominula* colonies given in the same source. Wright

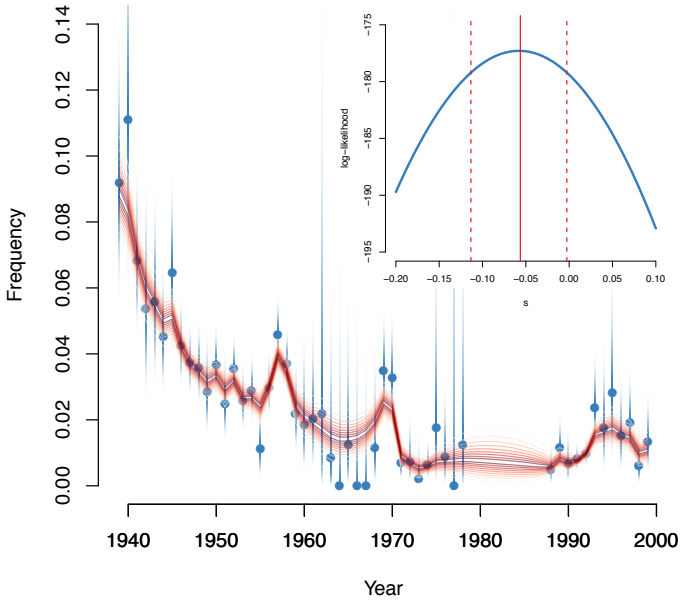


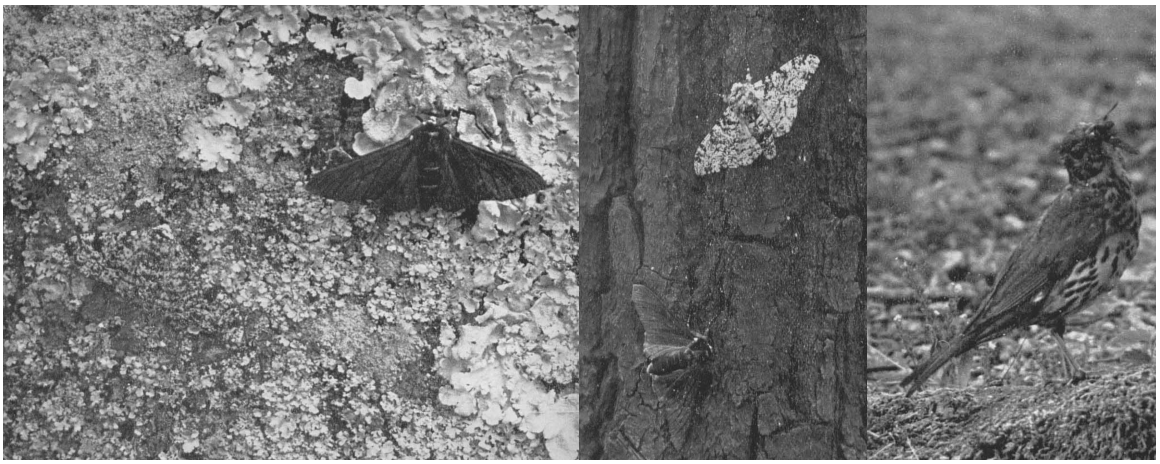
Figure 2.8: *Panaxia dominula* data. **Main plot**: *medionigra* frequency across generations. Blue dots show observed points with shaded support intervals. Red lines show posterior confidence intervals for the true frequency, from 10% (darkest) to 90% (lightest). **Inset**: log-likelihood as a function of s. For each value of s, we computed the likelihood of the observations using the forward algorithm for HMMs. Red solid and dashed lines show the MLE and the 95% confidence interval respectively. This figure shows the results for a model of additive selection though, as discussed in the main text, a recessive model may fit better.

(1948) argued that $2N_e$ might be of the order of 100 and O'hara (2005) estimated it to be of the order of a few hundred. If $2N_e = 100$, we would not reject $s = 0$ ($2\Delta\ell = 0.45$, approximate χ^2_1 p-value=0.50). A recessive model fits with a higher likelihood, (change in log-likelihood, $\Delta\ell = +2.5$ for $h = 0$ compared to $h = \frac{1}{2}$), but fits a large negative selection coefficient $\hat{s} \approx -1$, which is outside the range for which our approximations are valid, but may indicate that a model of recessive lethality (or near-lethality) is the best explanation for the data.

In the end then, our conclusion depends largely on the assumptions that we make about effective population size, which is essentially the conclusion we reach by reading Fisher and Wright's original papers. It is not surprising, given the form of our estimator, that the past 50 years of observations when f is very small do not add much to this estimate. However, the fact that the allele was still present after 60 years does give more support to the idea of selection on a recessive phenotype (or, indeed, a neutral phenotype which would have similar dynamics for small f). Simulating under a Wright-Fisher model of diploid selection using $f_0 = 0.1$, we find that under a fully recessive lethal model, approximately 58% of trajectories have not fixed at 0 by $T = 60$, compared to 41% under our best fit additive model.

2.5.2 *Biston betularia*

For our structured population dataset, we turned to another classical dataset on moth morph frequencies, this one of the species *Biston betularia* (peppered moth). A classic example of



(a) (b) (c)

Figure 2.9: The *typica* and *carbonaria* morphs of *B. betularia*. **a**: The two morphs resting on a lichen-covered tree trunk. **b**: Resting on a soot-covered trunk. **c**: A song thrush with a *carbonaria* in its beak. Pictures taken from Kettlewell (1956).

---

natural selection, the morph of interest here is the *carbonaria* morph, which appears very dark or black, in contrast to the *typica* morph, which has a complex speckled pattern on a light background. A full discussion of the extensive literature on this species is beyond the scope of this article (but see Cook (2003) for a review). Briefly, the *carbonaria* morph was identified in the North of England by 1848 and by 1958 was present at a frequency of approximately 100% in the North West, and at varying frequencies throughout the rest of the country (Kettlewell, 1958). Starting in the early 1970s, the frequency of the *carbonaria* morph began to decline, until today it is found very rarely. The change in frequency of the morph is almost certainly the result of a strong selective pressure, initially positive and changing direction some time in the mid 20^{th} century. The changes in frequency correlated strongly, both temporally and geographically with levels of industrial pollution and the classical explanation for this selective pressure was that the *carbonaria* form was better camouflaged against soot-covered tree trunks and thus subject to less predation by birds when industrial pollution was high (Kettlewell, 1955, 1956, ; Figure 2.9). In the second half of the 20^{th} century, and particularly following the Clean Air Act of 1956, levels of pollution dropped dramatically. Since the *carbonaria* allele is much more visible when there is no pollution, this can explain the subsequent drop in *carbonaria* frequencies. Though this explanation and Kettlewell's original experiments have been criticised, recent experiments appear to confirm this result (Cook *et al.*, 2012) and suggest a selection coefficient of around -0.1 for the *carbonaria* allele.

In addition to the *typica* and *carbonaria* morphs, there is also an intermediate form,

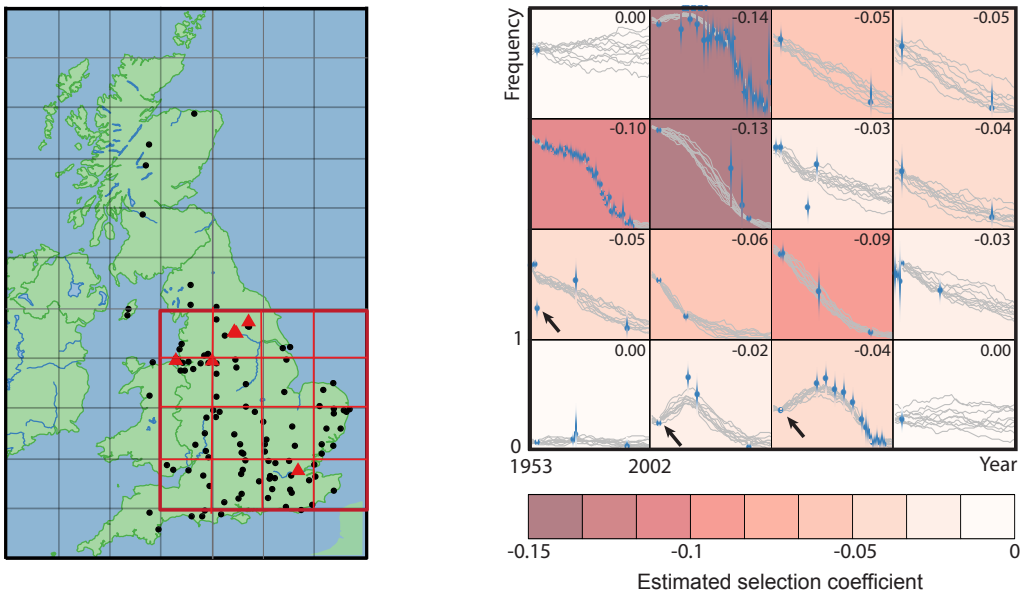
	$h = 0$	$h = \frac{1}{2}$	$h = 1$
Log-likelihood, s^{ij} varying	-660.0	-623.6	-613.6
Log-likelihood, s^{ij} constant	-695.6	-656.9	-669.2
Approximate χ^2_{15} P-value	2.7×10^{-9}	1.9×10^{-8}	9.6×10^{-17}

Table 2.1: Log-likelihoods for the *B. betularia* data, with approximate P-values for the null hypothesis of constant selection against the alternative of varying selection.

insularia, controlled by a different allele at the same locus (Lees and Creed, 1977). It is relatively rare ($< 10\%$ of observations), and we excluded it from our analysis.

We searched the literature for observations of the frequency of the *carbonaria* morph across England since 1958 (Bishop *et al.*, 1978; Clarke *et al.*, 1994; Cook *et al.*, 1999, 2002, 2005; Cook and Turner, 2008; Grant *et al.*, 1996, 1998; Kettlewell, 1958; Lees and Creed, 1975; Mani and Majerus, 1993; West, 1993, we were able to extract data from 8 of these 12 references). Many data points had been reported more than once, and we attempted to remove duplicate observations. For each observation we extracted the number of moths collected, the numbers of *typica* and *carbonaria* observed, and the location of the observation. Assuming a single dominant allele and Hardy-Weinberg equilibrium, we converted the carbonaria frequency to an allele frequency as $f = 1 - \sqrt{1 - f_c}$ where f_c is the *carbonaria* frequency. We then assigned the observations to large Ordnance Survey grid squares. One of the corner grid squares lies entirely in the North Sea, and had no observations. We filled this in by averaging over the two adjacent squares, reasoning that this would have caused the least disruption to the dynamics of the rest of the grid.

To analyse the *B. betularia* data using our lattice model estimator, we used $m = 0$ as an initial value, and stopped when successive log-likelihoods differed by less than 0.005. If we assume that $2N_e = 1000$, we estimate selection coefficients for the *carbonaria* allele varying spatially between 0 and -0.12. We also estimate that $\hat{m} = 0.00$. If we constrain s to be constant across the range, we estimate that $\hat{s} = -0.068$; however, we strongly reject the hypothesis that s is constant ($2\Delta\ell = 67$, approximate χ^2_{15} p-value $= 1.7 \times 10^{-8}$). Cook (2003) gives estimates for s from different sites ranging from -0.018 to -0.208. There are three datapoints, all consisting of observations from Kettlewell (1958) which have a large influence on the result that selection is not constant. If these are removed then the p-value is less significant ($2\Delta\ell = 36$, approximate χ^2_{15} p-value$=1.9 \times 10^{-3}$). The model of dominant selection fitted better than additive or recessive selection ($\Delta\ell$ = -36 and +10 for $h = 0$ and $h = 1$ compared to $h = \frac{1}{2}$). The fit of the dominant model is shown in Figure 2.10. In this case we reject the hypothesis of constant selection even more strongly ($2\Delta\ell = 111$, approximate χ^2_{15} p-value$=9.6 \times 10^{-17}$). The models are summarised in Table 2.1.



(a) (b)

Figure 2.10: *Biston betularia* analysis. **a**: Sample sites. The grey grid shows UK Ordnance Survey national grid reference squares. The red highlighted squares indicate the range we considered for our analysis. Black dots indicate sample sites. Red triangles indicate sites with observations in five or more years. We excluded sites outside the red area. **b**: Estimated selection coefficients for the *carbonaria* allele. Each grid square represents a single deme. Time in generations runs along the x-axis in each deme from 1953 to 2002. Allele frequency runs on the y-axis in each deme from 0 to 1. Blue dots are observations, collapsed over all sites in a deme for each year. Grey lines are sample paths from the final pseudo-posterior distribution. The background colour of each deme represents the MLEs of the selection coefficients, which are given in the top right corner of each deme. We assumed $2N_e = 1000$ (in each deme) and complete dominance (h=1). Black arrows indicate points of high influence (all from Kettlewell (1958)). If these points are removed then the $\hat{s}^{ij}$ in these demes lies between -0.12 and -0.10. The north west and south east squares each have observations in only one year, so the likelihood is almost completely flat.

Overall then, we found strong evidence that selection was not constant across the range, a conclusion which is robust to the assumptions we make about population size. However, it seems likely that several of our assumptions are violated in this population. In particular, given the rapid increase in *carbonaria* frequencies in the first half of the 20^{th} century followed by the rapid decrease observed since, it seems likely that the sign of s switched from positive to negative at some point, making our assumption of time-constant s since 1953 implausible. The highly significant p-values we obtained are likely due, at least in part, to poor model fit. In order to incorporate time-varying selection into this model, we could include an additional HMM step to fit s as a function of t, subject to some assumptions about the rate of change of s. Model comparisons indicated that selection was dominant, which is consistant with the fact that the allele is dominant for the *carbonaria* trait (though note that this does not imply that selection must act dominantly, since the allele may have pleiotropic effects). It is slightly puzzling that we estimated $m = 0$ (in fact, we estimated it to be slightly negative and set it to zero), but given the uncertainly in the estimate, this is certainly consistent with a very small, but nonzero, migration rate. We made a rather crude partition of the space into a grid and this may contribute to poor estimation of migration rate. A better approach would be to estimate the neighbourhood positions by fitting them as polygonal tilings following Guillot *et al.* (2005).

2.6 Extending to haplotype data

Essentially, inferences about selection depend on understanding the behaviour of allele frequencies over time. The method we have developed in this chapter uses this information directly, observing a single locus, and relies on having explicit temporal information. A somewhat orthogonal approach recognises that this temporal information is also encoded implicitly in the haplotype structure of the population. This information is difficult to extract but the advantage is that, in practice, it is almost always easier to obtain data from the current generation of a population than from previous generations. Specifically in humans, we have enormous quantities of sequence and genotype data available for modern humans but data from ancient humans is patchy and low quality. Additionally, it is much easier to obtain single-locus information than full sequence data from ancient DNA. Given this, it is natural to ask whether we can combine these two approaches to estimation and develop a method which uses both sparse, single locus allele frequency time series, and plentiful modern sequence data. This section describes, to our knowledge, the first attempt to develop such a method to make parametric inference about selection.

Since ancient DNA has only recently started to become available, most attempts to detect recent selection in humans have used only modern haplotype information. While it is

possible to make qualitative inference about selection (for example, ranking genomic regions) by computing various statistics, quantitative inference is difficult because the likelihood of the ancestral selection (and recombination) graph is impossible to calculate in practice. One method which is often used in this situation is approximate Bayesian computation (ABC). Introduced by Tavaré *et al.* (1997), also used by Weiss and von Haeseler (1998) (and similar in spirit to earlier stochastic methods for dealing with intractable likelihoods (Diggle and Gratton, 1984), ABC is a rejection method for drawing an approximate sample from a posterior distribution where we are unable to calculate the likelihood. Suppose we make an observation X from a distribution parameterised by θ, but we cannot calculate the likelihood $\ell(\theta; X)$. Let Θ be the random variable representing the distribution of θ. Starting from a prior on the model parameters, we sample sets of parameters $\{\theta_i\}$. Then, for each i, we simulate an observation x_i, and compute a vector of summary statistics t_i. We compute the same summary statistics for our observed data, t^* say. Suppose there are n parameters and m summary statistics. Now we consider the set $\Pi = \{\theta_i \ : \ d(t_i, t^*) < \epsilon\}$ where $d(\bullet, \bullet)$ is some distance metric on the space of summary statistics, and ϵ is chosen arbitrarily. We call this set the ϵ-ball and write $\pi_\epsilon(\theta)$ for the empirical distribution of the θ_i in Π. We call $\pi_\epsilon(\theta)$ the *truncated prior*. The original formulation of ABC treats this as a sample from the posterior of θ given the observations, which is equivalent to assuming that the likelihood is flat for the set of parameters leading to the summary statistics in the ϵ-ball. Improved sampling can usually be obtained by assuming a parametric form for the likelihood in this region, a technique known as *post-sampling adjustment*. We used the *ABC-GLM* method introduced by Leuenberger and Wegmann (2010) and implemented in Wegmann *et al.* (2010). They assume that, within the ϵ-ball, the likelihood of the summary statistics conditional on the parameters is multivariate normal with the form $T|\theta \propto MVN(A\theta + \mu, B)$ where A,B and μ are $m \times n$, $m \times m$ and $m \times 1$, respectively. A,B and μ are estimated from the retained simulations and then the posterior distribution of θ is estimated as $f_{\Theta|t^*}(\theta) \sim f_{T|\theta}(t^*)\,\pi_\epsilon(\theta)$.

Peter *et al.* (2012) use ABC to estimate selection coefficients and other parameters, taking various LD-based measures and sequence statistics as summary statistics. We take the same approach but with a prior generated using the method presented in this chapter with the time series data. Figure 2.5b implies that, so long as N_e is relatively large, we could generate a reasonably strong prior on s without many observations. An alternative, and equivalent, approach would be to perform the ABC reference simulations conditional on the observed allele frequency. Since as part of the HMM calculations we have calculated the forward and backward matrices, which is equivalent to knowing the full posterior distribution of paths, this would be simple to do. Of course, this approach would also naturally

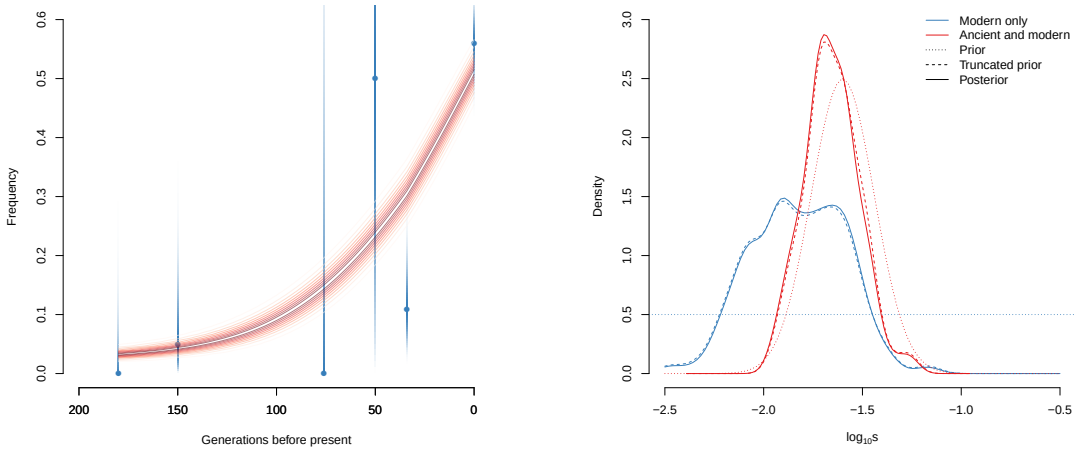
Source	Age (y)	Frequency	Location
Burger *et al.* (2007)	5400	0/16	Scandinavia
	2267	0/2	
	1500	1/2	
Malmström *et al.* (2010)	4500	1/20	Central Europe
Nagy *et al.* (2011)	1000	5/46	Hungary
Plantinga *et al.* (2012)*	5070	10/38	South-West Europe
	4450	2/14	

Table 2.2: Reported allele frequencies of the T allele of rs4988235 in ancient European samples. The age (in years before present) is the mean of the range in the source, and the frequency is the allele count, out of the total tested. *We did not use the sample from Plantinga *et al.* (2012) in our analysis, since it is from an Iberian population and therefore likely to be quite distinct from Northern Europe.

extend to the structured case, in which case we could perform coalescent simulations under a model similar to the one we use in Chapter 3, but including recombination and then go on to perform the ABC step as before.

We implemented this idea using the pipeline of Peter *et al.* (2012), which uses *mbs* (Teshima and Innan, 2009) for simulation and *ABCtoolbox* (Wegmann *et al.*, 2010) for inference, as described above. We investigated selection at the lactase persistence locus in northern Europeans, which is one of the best studied examples of a selective sweep in humans. The genetics of lactase persistence are reviewed by Ingram *et al.* (2009) and Swallow (2003) but we describe them briefly here. Lactase, encoded by the LCT gene on chromosome 2, is an enzyme which breaks down lactose, a sugar found in milk. In ancestral humans, and most modern humans, the gene is highly expressed during infancy, but expression declines with age and adult humans do not express LCT to high levels, and therefore cannot digest lactose. In some human populations, the expression of lactase persists into adulthood, allowing them to digest milk. In Northern Europeans, the lactase persistence phenotype is almost perfectly associated with the derived T allele of rs4988235, a C/T SNP in the gene MCM6, 14kb from LCT, which lies on an extremely long haplotype (Enattah *et al.*, 2002; Poulter *et al.*, 2003). In other populations, lactase persistence is associated with different mutations, strongly suggesting independent origins of the phenotype (Tishkoff *et al.*, 2006; Enattah *et al.*, 2008). Inheritance of the phenotype is usually described as dominant, but reports of intermediate enzymatic activity in heterozygotes suggest that perhaps $0.5 \leq h \leq 1$ (Ho *et al.*, 1982; Flatz, 1984).

To demonstrate how the methods we developed in this chapter could be useful for historical inference, we analysed both ancient and modern data on the evolution of lactase persistence allele, and then combined these analyses. First we searched the literature for



(a) (b)

Figure 2.11: Analysing the LCT gene using ancient DNA **a:** Our single population estimator applied to data from samples of ancient DNA from North-East Europe (Table 2.2). Blue points show allele frequencies and their confidence intervals. The red lines show intervals of equal posterior probability at 10% intervals for the underlying allele frequency. **b**: Comparing the ABC estimator of Peter *et al.* (2012) using a uniform prior on $\log_{10} s$ (blue line) with the same estimator using a normal prior estimated using the likelihood from the analysis in **a**.

---

reports of ancient DNA samples from Europe which had been genotyped for rs4988235 (Table 2.2). We analysed this data in the same way as the *Panaxia* data in Section 2.5.1 (Figure 2.11a). We assumed an effective population size $2N_e = 14,000$. Fitting the model for different values of h gave the MLE for h as 0.506, though the difference in likelihoods was not large (for $h = 1$, $\Delta\ell = -1.7$ compared to $h = 0.5$). Therefore for the rest of the analysis we assumed that $h = 0.5$. We estimated a selection coefficient of 0.022 (95% CI 0.015-0.099).

To analyse modern data we followed Peter *et al.* (2012), and estimated the selection coefficient (on a log scale) for rs4988235 in the 93 Finnish samples from the 1000 Genomes Project Phase 1 data release (1000 Genomes Project Consortium, 2012). We used a 120kb region around LCT (Chr2:136,535,946-136,657,220). Again we assumed that $h = 0.5$[†], and used variable recombination rates from the HapMap CEU recombination map (International HapMap Consortium, 2010). We chose the per-site per-generation mutation rate μ from a uniform prior on $\left(5 \times 10^{-9}, 2.5 \times 01^{-8}\right)$ and $\log_{10} s$ from a uniform prior on $(-2.5, -0.5)$. We estimated that $\log_{10} s = -1.82$ (-2.22,-1.46)[‡]. Equivalently, $s = 0.015$ (0.006,0.034). We

[†]We did attempt to estimate h, but there was virtually no information in the data and the posterior was unchanged from the prior.

[‡]Parameter estimates are posterior means, and numbers in brackets after them are 95% credible intervals, unless otherwise stated.

also estimated $\mu = 1.53 \times 10^{-8}$ (1.12×10^{-8}, 1.95×10^{-8}). The prior, truncated prior, and posterior distributions are shown in blue (Figure 2.11b).

Finally, to combine the ancient and modern datasets, we computed the log-likelihood surface (in $\log_{10} s$-space) for the ancient DNA data, as described in this chapter, and fitted a normal distribution in order to minimise the average pointwise squared deviation[†]. This had mean -1.64 and standard deviation 0.16. We repeated the Peter *et al.* (2012) ABC inference on the modern data using this prior (Figure 2.11b, red lines), rather than a constant prior. We estimated that $\log_{10} s = -1.66$ (-1.90, -1.43), i.e. $s = 0.021$ (0.012, 0.037).

In this example, the combined estimates are close to the estimate from the ancient DNA suggesting that, given this data, there is little extra information in the modern data. It appears the modern data are consistent with very low selection coefficients which are not supported by the ancient data. Since the most obvious signal of selection in this region is the large shared haplotype, we were concerned that perhaps we had looked at too small a region to capture this effect, and that this was why small selection coefficients were being accepted. We re-ran the analysis on a larger (300kb) region and obtained very similar results, suggesting that the size of the region used could not explain this effect. This region has a very low recombination rate (mean per-site recombination rate $\rho = 2.15 \times 10^{-9}$), and presumably this means that very long haplotypes are plausible, even with little selection. We could not estimate h well using either the ancient data or the modern data. The ancient data suggested it was close to 0.5, but we cannot be certain about this. In general if h is higher then the estimated selection coefficient is lower.

Published estimates of the selection coefficient on rs4988235 tend to be higher than we have estimated here. Bersaglieri *et al.* (2004) estimated s in Scandinavian populations to be 0.09-0.19, based on inter-population differentiation. Tishkoff *et al.* (2006) estimated $s = 0.069$ using a haplotype-decay statistic. Burger *et al.* (2007) estimated $s = 0.089$ using a deterministic trajectory based on a small ancient sample and Itan *et al.* (2009) estimated $s = 0.095$ in Northern European dairy farmers using a spatially explicit ABC approach, which incorporated data about the dates at which dairy farming began. While all these have large confidence intervals, our estimate is definitely at the low end. Peter *et al.* (2012) reported $s = 0.025$ which is slightly higher than our estimate using ostensibly the same pipeline and data but is similar to our estimate from the combined data. We cannot explain why our estimates are different, although we have used a different (though mostly overlapping) sample. We may also have made different assumptions about population sizes

[†]We could use the empirical density exactly because *ABCtoolbox* only accepts parametric distributions as priors. Given this, probably a more sensible strategy would be to minimise the Kullblack-Leibler divergence; $\int_{-\infty}^{+\infty} \log\left(\frac{p(x)}{f(x)}\right) f(x)dx$ where $p(x)$ is the prior and $f(x)$ is the observed density.

and generation times. and it may be that some of the difference is explained by this. Ultimately more ancient samples will allow us to easily answer this question and perform similar analysis on other regions thought to be under recent selection.

2.7 Conclusion: Selection and structure

Historically, it has been difficult to estimate selection coefficients in humans, even for alleles which are known to be strongly selected, though the increasing availability of historical allele frequency data in humans mean that it is a problem in which there is growing interest. In the near future it may finally become possible to test long-standing hypothesis about the relationship between adaptive differences between human populations and differential disease risk. However it is clear that doing so will require the development of models which incorporate many of the effects, both demographic (for example, migration) and genomic (for example, recombination rate variation), which can be confounded with the effect of selection. In this chapter we developed an estimator which takes into account a very simple equilibrium demographic model. However, human history is largely characterised by large migrations and intense bottlenecks, and it is unlikely that the model we developed here will be universally appropriate. Important future work therefore, would be to develop more similar methods for inference in more complicated demographic models. As we mentioned earlier, effects like time- and frequency- dependent selection could also be incorporated relatively easily.

An additional area for development is methods to integrate different sources of data. In Section 2.6, we demonstrated a simple way to combine time series data of allele frequencies with single-timepoint haplotype data. However in practice, there are many more sources of information which could be incorporated into the analysis. Examples of extra information about that could be incorporated in the case of lactase persistence might be epidemiological information about the health effects of the phenotype (Smith *et al.*, 2008), biological information about the functional effect of the mutation (Lewinsky *et al.*, 2005), and archaeological evidence about the consumption of milk and dairy products (Salque *et al.*, 2013; Dunne *et al.*, 2012). An example of this approach can be found in Itan *et al.* (2009). Which information we include will depend on the data and the questions that we are trying to answer, but the sort of modelling framework we described here ought to be able to incorporate virtually any data.

Another important area which we have ignored in this chapter is the question of different modes of selection. We looked here at the "classic selective sweep" scenario in which a single beneficial mutation arises in a population and sweeps, relatively rapidly, to fixation

(Maynard Smith and Haigh, 1974)[†]. However, such events seem to be rare in humans (Hernandez *et al.*, 2011) and there is evidence that other modes of selection may be important, for example soft sweeps (Hermisson and Pennings, 2005), balancing selection (Leffler *et al.*, 2013) and particularly, selection on polygenic traits (Turchin *et al.*, 2012). These effects are more difficult to include in the framework we have developed, but will likely form an important part of our understanding of recent human evolution.

Identifying and quantifying selection is interesting for many reasons, but one of them is that it can lead us to genomic regions which are phenotypically important. A region under strong selection is likely to be involved with a phenotype that that has a large effect on fitness. However, a more direct way to find genes which are involved in interesting phenotypes, and one which has certainly not been neglected in recent human genetics is an association study. Since an association study is just another way to look at the intersection of genotype, phenotype and environment, we might expect that spatial structure will be important to consider as well. In the next chapter we investigate this effect and in particular, describe the confounding effect of spatial structure in the presence of non-genetic risk, something which is of practical importance for most, if not all, association studies.

[†]Maynard Smith and Haigh (1974) were the first to correctly study the hitchhiking effect, though they did not use the term "selective sweep" which was introduced by Berry *et al.* (1991).

Chapter 3

Association studies of rare variants

In the previous chapter we investigated the effect of spatial structure on the estimation of selection coefficients. We now turn to the effect of this structure on association studies. It is well understood that spatial structure, and population structure in general, are important confounding factors in association studies, and methods for detecting and correcting for this structure are well developed. However, until recently most association studies have focussed on variants which are relatively common in the population and it is not well understood how population structure would affect studies of rare variants. We develop a structured coalescent framework to simulate genotypes from a lattice population[†] . Then, we simulate association studies under various models of non-genetic risk, and investigate how the non-genetic risk interacts with variants of different frequencies to produces false positive associations. We demonstrate that rare variants can be more confounded by non-genetic risk with a sharp spatial distribution, and that standard methods for correcting for population structure do not correct for this confounding. Another way of thinking about this effect is that, in a spatially structured population, rare variants are highly informative about spatial position and, more generally, about recent relatedness. In the next chapter, we go on to develop this idea and extract information about inter-population relationships from rare variant sharing.

The work in this chapter largely follows Mathieson and McVean (2012). Much of the text from Section 3.2 onwards is from this reference, as are Figures 4.6 to 4.11.

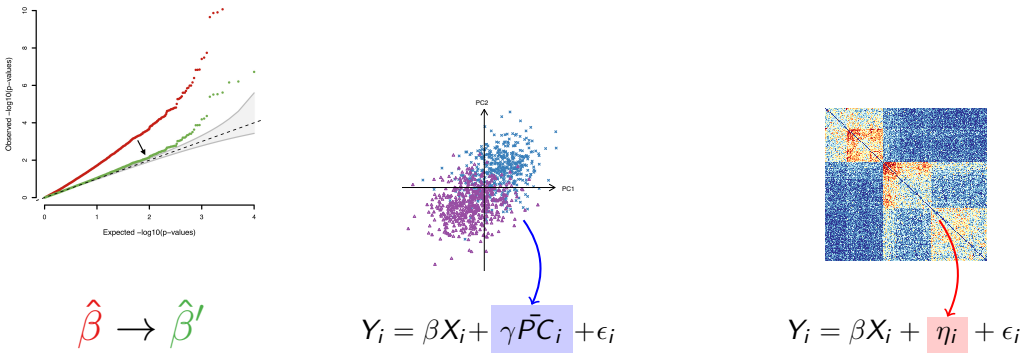
3.1 Association studies

Association studies aim to test whether a specific variant at a specific genomic locus has an effect on a trait of interest, within a population. The basic idea is to sample individuals

[†]The same population model we used in Chapter 2, but here, instead of the forward Wright-Fisher process, we will work with the backward coalescent process, which arises as a limit of the Wright-Fisher and other models.

from within that population and look for a correlation between the variant of interest and the trait. This idea has a relatively long history, but because identifying and typing variants was technically challenging and time-consuming, before the twenty-first century it was only possible to look at relatively small numbers of variants in any particular study. Early association studies suffered from two main issues. First, since it was only possible to test relatively small numbers of variants, it was necessary to have a good *a priori* idea of which loci might be involved. Second, and more damaging, was the issue of population structure. This works as follows; suppose the trait of interest was affected by some (either genetic or non-genetic) risk factor which varied across the population, and also that the population contains some underlying substructure. Then as a result of restricted migration or differential selection between subpopulations, loci with no effect on the trait can drift to different frequencies in different subpopulations and thus be correlated ("associated") with the trait, even though there is no real effect. For example, in an association study of *any* trait in a cohort that included both Europeans and another population, the LCT locus investigated in Section 2.6 would be significantly associated. Knowler *et al.* (1988) demonstrated a classic example of this sort of confounding where a marker appeared to be associated with type 2 diabetes in Native Americans, but really just represented admixture proportions. Because of these problems, disease mapping studies tended to use pedigree-based designs rather than association studies. Though these are not affected by population structure in the same way, they are underpowered compared to association studies and the sample sizes required to detect anything but the strongest effects are prohibitively large (Risch and Merikangas, 1996).

These problems were solved to a large extent by the development of microarray genotyping technology around the year 2000, which made it possible to type hundreds of thousands of SNPs across the genome. This meant that most of the genome could be tested for association simultaneously, rather than having to target specific candidate loci. The problem of population structure could be dealt with by choosing loci known (or strongly believed to be) unassociated with the trait, and using any association with those loci to correct for the population structure in the population. While this idea does not rely on having genome-wide data and in fact, was first proposed in the context of standard association studies by Pritchard and Rosenberg (1999), it becomes much more powerful with very large numbers of markers since we can then look for systematic patterns over the whole genome, symptomatic of underlying population structure. Several methods for doing this have been developed but in practice the most commonly used are genomic control (Devlin and Roeder, 1999; Bacanu and Devlin, 2000), principal component analysis (PCA) (Price *et al.*, 2006) and linear mixed models (Kang *et al.*, 2008). These differ conceptually to some extent in



(a) (b) (c)

Figure 3.1: Methods of correcting for population structure in GWAS. The basic test for an additive effect of a single variant is a simple linear model (or a logistic model for a case-control study). The trait value Y_i of individual i is modelled by $Y_i = \beta X_i + \epsilon_i$ where X_i is the genotype of the variant and ϵ_i is an IID (independently, identically distributed) error. The effect of that variant is significant if the estimated coefficient $\hat{\beta}$ is significantly different from zero, after an appropriate multiple testing correction for all the variants tested. **a**: Genomic control corrects for inflation in the test statistics $\hat{\beta}$ by scaling $\hat{\beta}$ for each SNP (shown here in a quantile-quantile plot) such that the median test statistic lies on the expected median of the null distribution. **b**: For PCA, we compute the principal components of the genotype matrix (first two shown here in a scatterplot - the two colours represent individuals from two subpopulations) and then include the first n principal components as terms in the regression. n is arbitrary, but 7 seems to be conventional. **c**: Linear mixed models are formed by adding an additional correlated error term η_i to the regression where the correlation matrix of the η_i is given by the kinship, or relatedness matrix of the individuals, estimated from all the markers tested and illustrated here as a heatmap.

---

that genomic control is a *post hoc* correction to the test statistics, whereas PCA and mixed models attempt to infer the structure, and then incorporate it into the model directly. Genomic control is computationally the simplest method to apply, but can result in a loss of power, and so is often used as a test for structure as much as a correction. PCA and mixed models work similarly, but PCA is intuitive and computationally simple and was historically preferred. The idea of using mixed models to model population structure comes from the animal breeding literature (Henderson, 1984, for example), where extreme structure (i.e. pedigrees) is the norm. Recently the use of mixed models for GWAS has become more common, due mostly to recent computational advances (Pirinen *et al.*, 2012; Lippert *et al.*, 2011) which mean that it is now feasible to fit them to genome-wide data. They have been reported to be more effective than either PCA or mixed models at controlling stratification (Sawcer *et al.*, 2011).

The first large-scale GWAS was the Wellcome Trust Case Control Consortium (2007), which conducted a case-control GWAS on seven common diseases using shared control. Driven mainly by disease studies, the explosion in the number of published GWAS since then (Figure 3.2) is a sign of how effective this experimental design can be at identifying

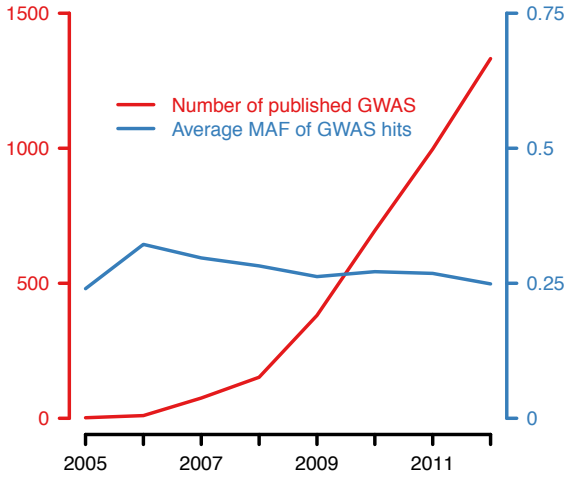


Figure 3.2: Since 2005, the number of published GWAS has risen dramatically (red line). Interestingly, the average minor allele frequency (MAF) of identified loci did not change greatly, despite denser genotyping, imputation and larger sample sizes. Data from the NHGRI GWAS catalog at www.genome.gov/gwastudies.

disease-related loci. The NHGRI GWAS catalogue (Hindorff, Junkins, Hall, Mehta, and Manolio, Hindorff *et al.*, www.genome.gov/gwastudies) which seeks to record the results of all published GWAS, listed at the time of writing (March 2013), 1,532 publications identifying a total of 8,698 SNPs associated with various traits and diseases.

Though there have been some examples of GWAS detecting single loci which explain most of the phenotypic variance of a trait (Klein, 2005), for many traits, most detected genetic effects have been relatively modest. Large scale meta-analyses in traits such as height (Lango Allen *et al.*, 2010), obesity (Speliotes *et al.*, 2010) and type 2 diabetes risk (Voight *et al.*, 2010) have detected many loci, each with an individually small effect. In fact, probably one of the most important contributions that GWAS have made to our understanding of the architecture of complex traits is the empirical demonstration that there do not exist many common variants with large phenotypic effects. Even for the best studied traits, all the loci identified through GWAS explain only a small proportion of the variance which is known to be due to genetic effects. This is known as the problem of the "missing heritability" (Manolio *et al.*, 2009). While there are many possible explanations for this, which are beyond the scope of this work[†] the suggestion that rare variants[‡] of large effect may contribute significantly to heritability is common, though intensely debated (Gibson, 2011). Perhaps the most convincing argument for why this might be true is that mutations with large deleterious phenotypic effects are likely to be selected against,

[†]Popular explanations include a long tail of common variants each with a small effect, large numbers of rare variants, epistasis, gene-environment interactions, bad heritability estimates and epigenetic effects.

[‡]Obviously, the definition of "rare" in this context is somewhat arbitrary and different authors use different definitions. A population frequency of $< 1\%$ seems to be generally considered rare. In the rest of this chapter, we'll have this in mind, but specify a range if it is important for the analysis.

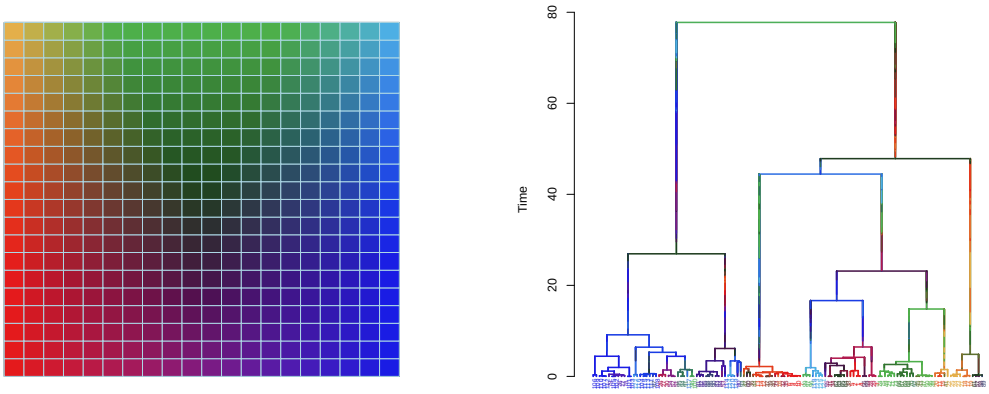
and therefore unlikely to rise to high frequency in the population. Therefore we might expect that variants of large effect are more likely to be found among rare than common variants. The fact that GWAS have typically found common variants (Figure 3.2) might be interpreted to mean that most causal variants are common, but in fact there are a number of reasons to expect that GWAS to date would not have detected rare variants, even if they were strongly associated.

The most important reason for the lack of rare variant associations is that the power to detect associations falls off dramatically as the frequency of the tested variant decreases. Therefore extremely large samples are required to have any chance of detecting rare variant associations. For example, in a case-control study of a common disease with 20% penetrance, to have 80% power to detect a variant at frequency 10% with an additive odds ratio of 1.1 requires a total sample size of around 12,000 whereas a variant at 1% frequency requires a sample size of 100,000 to achieve the same power[†]. Second is the problem of ascertainment. Genotyping chips typically do not contain rare variants and common imputation panels (such as the HapMap project) had low power to detect rare variants . Finally, a reason which is important but perhaps under appreciated is that rare variants tests are much more sensitive to simple technical errors such as incorrect genotypes or numerical errors, and therefore most GWAS analyses simply remove them to avoid noise in the results.

However, recently the collection of large samples and meta-samples, the collection of larger imputation panels (such as the 1000 Genomes project), and the hypothetical contribution to the missing heritability described above, have stimulated investigation into this problem and the development of novel statistical methods for testing rare variant associations (Ionita-Laza *et al.*, 2011; Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010; Neale *et al.*, 2011, and many more). Many of these methods seek to increase power by combining information across nearby variants and a common test design is a so-called "burden" or "collapsing" test which tests whether a particular gene (or pathway) tends to carry more rare variants in cases compared to controls.

Although, as we saw above, there has been much theoretical and practical work developing corrections for population structure in GWAS, there has been little investigation of the effect that such structure would have on rare variants. In the rest of this chapter, we investigate this problem in the context of spatial structure. We first develop a structured coalescent simulation framework, which we describe in the next section.

[†]Calculated using the *Genetic Power Calculator* at http://pngu.mgh.harvard.edu/∼purcell/gpc/ (Purcell *et al.*, 2003).



(a) A 20×20 grid.

(b) A genealogy of 400 samples

Figure 3.3: We simulate genealogies of a sample drawn from a square grid. The colours of the lineages indicate which grid square that lineage is in at any time. Individuals from nearby squares coalesce more quickly than those from distant squares. This captures the isolation by distance effect of spatial structure, that spatially close individuals tend to be more closely related. We can also see the distinction between the "scattering" phase where nearby individuals coalesce very fast, and the "collecting" phase, where a small number of remaining lineages take a long time to coalesce (Wakeley, 1998, 1999, though this is not quite the same model).

3.2 Simulating genotypes in the lattice model

We consider a population living on a two dimensional grid, just as in Chapter 2. Each individual lives in a single grid square, and each grid square can contain more than one individual. However now, rather than considering a forward Wright-Fisher model, we simulate genealogies using a structured coalescent model[†] (Figure 3.3). The main parameter of interest in the simulation is the migration rate M (scaled by N_e) which determines how likely an individual is to migrate relative to coalesce. For two individuals in the same grid square, the probability that one of them migrates to another square before they coalesce is $\frac{M}{1+M}$. This parameter has an important effect on the amount of structure. Clearly as $M \to \infty$, the effect of spatial structure decreases, and the samples become totally random.

The simulations proceed as follows: suppose we have a $K \times K$ grid and we wish to simulate a sample of $C = cK^2$ individuals where c is the number of individuals in each grid square. We simulate L loci on each of G genealogies for a total of LG loci. Each genealogy represents an independent genomic region, with no recombination inside each region. Index the grid squares by i, j and denote the number of lineages in grid square i, j at time t by $n^t_{i,j}$. Let $s_{i,j}$ represent the number of grid squares adjacent[‡] to i, j, so $s_{i,j} \in \{2, 3, 4\}$.

[†]Which arises in the limit of the structured Wright-Fisher model, just as in the single population case.
[‡]Next to, but not diagonal

Now we start at $t = 0$ and repeat the following steps until only one lineage remains.

1. At time t, the rate of coalescence within grid square i, j is $\lambda^t_{i,j} = \frac{n^t_{i,j}(n^t_{i,j}-1)}{2}$ and the total rate of coalescence is $\lambda^t_{\bullet,\bullet} = \sum_{i,j} \frac{n^t_{i,j}(n^t_{i,j}-1)}{2}$ where we use dots to represent summation over indices. The rate of migration for each grid square is $\mu^t_{i,j} = \frac{Mn^t_{i,j}s_{i,j}}{2}$ and the total rate of migration is $\mu^t_{\bullet,\bullet}$. The next event occurs at time $t + T$ where $T \sim Exp(\lambda^t_{\bullet,\bullet} + \mu^t_{\bullet,\bullet})$. It is a coalescence with probability $\frac{\lambda^t_{\bullet,\bullet}}{\lambda^t_{\bullet,\bullet}+\mu^t_{\bullet,\bullet}}$ and migration with probability $\frac{\mu^t_{\bullet,\bullet}}{\mu^t_{\bullet,\bullet}+\lambda^t_{\bullet,\bullet}}$.

2. If the next event is coalescence, it occurs in grid square i, j with probability $\frac{\lambda^t_{i,j}}{\lambda^t_{\bullet,\bullet}}$. In this grid square, choose two lineages uniformly and join them together. Return to 1 with t replaced with $t + T$.

3. If the next event is migration, it occurs in grid square i, j with probability $\frac{\mu^t_{i,j}}{\mu^t_{\bullet,\bullet}}$. In this grid square, choose one lineage uniformly, and move it to one of the (2,3 or 4) adjacent squares, chosen at random. Return to 1 with t replaced with $t + T$.

Once we have simulated a single instance of the genealogy, we generate genotypes at L random loci by sampling L nodes from the genealogy with replacement, selecting each node with probability proportional to the branch above that node, and setting each individual's genotype to 1 if they are descended from that node, or 0 otherwise, so that a genotype of 0 represents an ancestral allele and a genotype of 1 represents a derived allele. We end up with G loci which are unlinked, but within each locus we have L variants which are in perfect LD.

The first few principal components of the genotypes generated in this model are highly predictive of spatial position (Figure 3.4). the first principal component is an interior-exterior axis, the second and third are East-West and North-South, and the fourth is a saddle pattern. These are exactly the patterns which are both predicted theoretically and observed in the first few principal components of real data (Novembre and Stephens, 2008, and particularly Figure 1 of that reference).

We are interested in the ways in which rare and common variants differ. Variants of all frequencies have the same, or similar, principal components, but differ in other ways. In particular, rare variants tend to be more spatially clustered. Informally, since rare variants are typically younger, they have not had time to migrate far from their origin, and therefore individuals carrying those variants tend to be spatially restricted.

One way in which we can quantitatively demonstrate the difference between variants of different frequency is to look at allele sharing probabilities at different frequencies.The

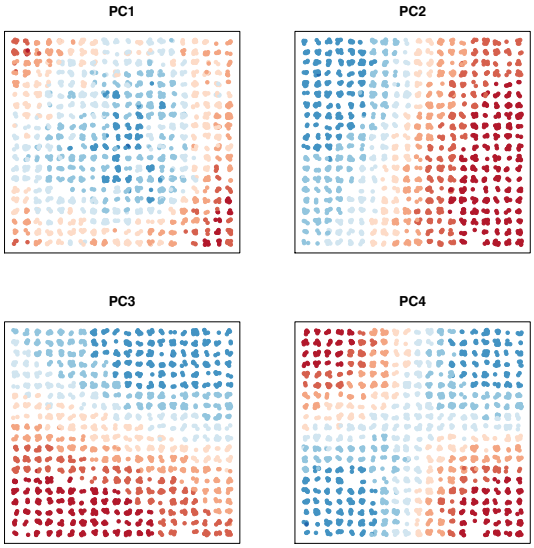
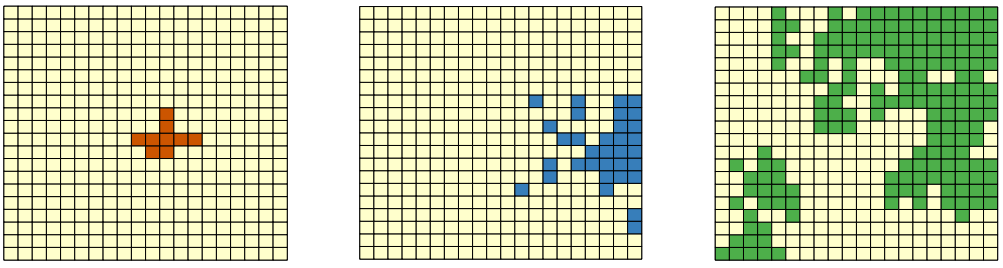


Figure 3.4: Plot of the first four principal components for a set of genotypes simulated from the lattice model. Each dot represents one sequence and there are five sequences in each of 400 grid squares. The colours represent the standardised value of that principal component for that individual, from red to blue (on an arbitrary scale).



(a) (b) (c)

Figure 3.5: Examples of the spatial distribution of **a**: rare, **b**: low frequency and **c**: common variants. In each case, coloured squares indicate demes where the variant is present. Note that the rare variant has a tightly clustered spatial distribution while the common variant has a much more irregular distribution.

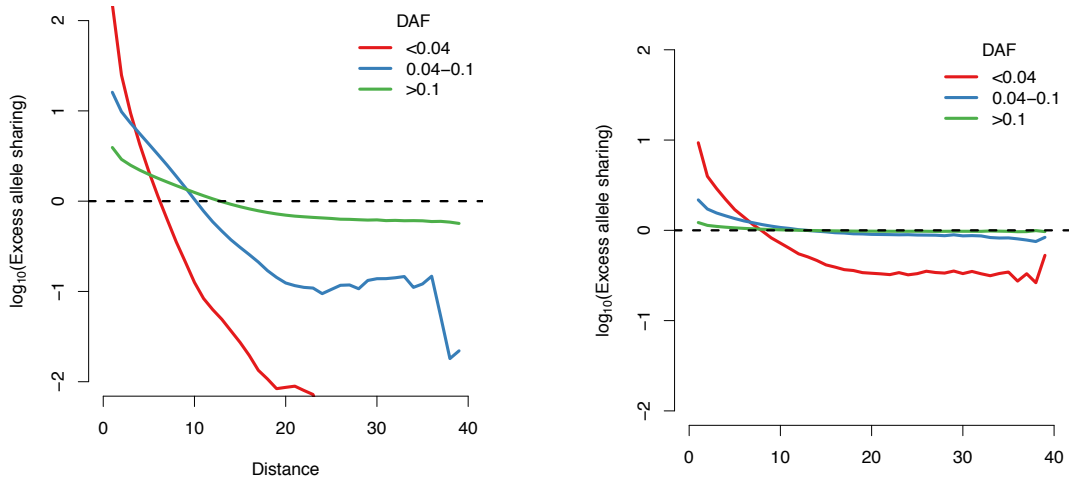
effect of spatial structure producing isolation by distance is that individuals will be more likely to share variants with other individuals who are spatially close to them. To quantify this, we calculate the probability that two individuals at a given distance share an allele compared to what would be expected from a randomly mating population, i.e. f.

Suppose we have sampled C chromosomes, then given the $LG \times C$ genotype matrix $X = \{X_k^{l,g}\}$ where $l \in \{1 \dots L\}$, $g \in \{1 \dots G\}$ and $k \in \{1 \dots C\}$, we first divide the variants into rare, low frequency, and common variants based on their allele frequencies. Suppose there are R rare variants in an $R \times C$ genotype matrix $\tilde{X} = \{\tilde{X}_k^r\}$ with allele frequencies $f_r = \frac{\tilde{X}_\bullet^r}{C}$ and the spatial distance between individuals i and j is given by $D_{i,j}$. Then we compute the excess allele sharing at distance d, Q_d as:

$$Q_d = \frac{1}{R} \sum_{r=1}^{R} \frac{\sum_{i=1}^{C} \mathbb{1}\{\tilde{X}_k^i = 1\} \sum_{j>i} [\mathbb{1}\{\tilde{X}_k^j = 1\} \mathbb{1}\{D_{i,j} = d\}]}{f_r \mathbb{1}\{\tilde{X}_k^i = 1\} \sum_{j>i} \mathbb{1}\{D_{i,j} = d\}} \tag{3.1}$$

where $\mathbb{1}\{A\}$ is the indicator function of the event A and recalling that $\tilde{X}_k^r \in \{0, 1\}$ with 0 and 1 representing the ancestral and derived allele respectively. So, for a given distance d, we count the total number of alleles which are shared between pairs of individuals at distance d, and divide by the total number of pairs of individuals which are separated by d. We then divide by the allele frequency which is the sharing probability in an infinite randomly mating population, to get the excess allele sharing Q_d. Figure 3.6, shows $\log_{10} Q_d$ as a function of d for different values of M. We see that all variants are more likely to be shared by individuals close together, but that the extent of excess sharing is much greater for rare variants compared to common variants. This excess increases as migration rate decreases, so when the migration rate is high, there is less excess sharing. However, even for an extremely high migration rate of $M = 10$, there is still noticeable excess sharing. Rare variants (frequency $< 4\%$) are almost ten times more likely to be shared with an individual in a neighbouring deme, compared to what would be expected in an unstructured population. This is consistent with both theoretical results about rare allele sharing (Slatkin, 1985) and empirical observations that show very low rates of rare allele sharing between different human populations (Bustamante *et al.*, 2011).

This observation raises the question of to what extend our simulations are realistic for natural populations. Clearly the lattice model is somewhat artificial, and in particular, natural populations are rarely at equilibrium. However, is still interesting to ask what values the parameter M might take on in a realistic population model. In order to do this, we calculated the statistic F_{ST} (using the Hudson *et al.* (1992) estimator) for different values of M, supposing that the entire grid was divided into halves, and computing F_{ST} between the two halves (Figure 3.7). There are two things to note about this plot. First, F_{ST} is lower for



(a) $M = 0.01$. (b) $M = 10$.

Figure 3.6: For $K = 20$, we simulated genotypes and computed the excess allele sharing (Equation 3.1) for variants at different distances apart. **a** shows the result of a very low migration rate leading to a highly structured population and **b** a high migration rate leading to a relatively unstructured population. Even in **b**, however, there is a noticeable increase in excess sharing, particularly for rare variants. Here "rare" is defined as $< 4\%$ frequency. Abbreviations; DAF - Derived allele frequency.

---

rare variants than for common variants. Second, F_{ST} decreases as M increases. Different authors estimate F_{ST} in different ways and this will affect the relative weighting of different frequencies. However it is usually closer to the value for common variants (since they are weighted more), and in any case, typically only common variants are available. We see therefore that for $M = 0.01$, which many of our simulations use, $F_{ST} > 0.1$, which is large for human populations, but would be around the level of inter-continental differentiation. On the other hand, when $M = 10$, F_{ST} is around 0.01. This is relatively low compared to the level within Europe of less than 0.02 (Nelis *et al.*, 2009), and it is quite plausible that we would encounter an association study on a population with this level of structure. Even at this level, we still observe significant excess allele sharing, particularly for rare variants (Figure 3.6b).

These results suggest that rare variants are highly informative about local population structure, and suggest that sharing of rare alleles could be a useful tool to investigate recent relatedness between individuals. We return to this idea in the next chapter. Another consequence of these observations is that rare variants, being themselves spatially restricted, will be more susceptible to confounding from nongenetic risk which is also spatially restricted, a result which we demonstrate in the next section.

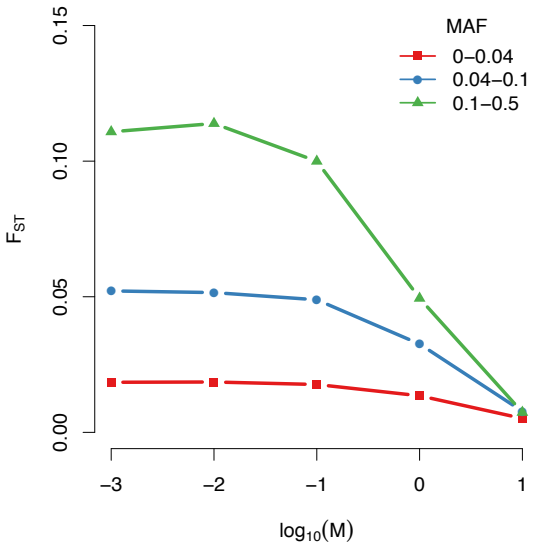


Figure 3.7: We divided the 20×20 grid into two and calculated F_{ST} between the two halves. We show separately F_{ST} for rare, low-frequency and common variants.

3.3 Association studies in the lattice model

In order to investigate the effect of this spatial structure on association studies, we simulate a quantitative trait for each individual that we sampled. We assume that the trait is normally distributed with variance 1. The mean of the trait for each individual depends on the the deme they are sampled from. We refer to the difference in trait value depending on location as the non-genetic risk. This could be in the context of a trait of interest (for example BMI), or as a binary disease risk in the context of an underlying liability model. We choose different geographic distributions of non-genetic risk to investigate their effects.

Formally, let $\phi : [1, C] \to [1, K] \times [1, K]$ be a function which maps each individual to the grid square which they originated in. Then, for individual k, the trait value $Y_k \sim N(R_{\phi(k)}, 1)$ where $R_{i,j}$ is the non-genetic risk in grid square i, j. That is, there is an additive environmental effect, but no genetic effect at all. In a real experiment, each individual would have a single value of Y_k which would be tested against every locus, but to reduce the uncertainty due to sampling error, in our simulations we resample the trait independently for each locus except where that would be inappropriate, for example when testing corrections (Section 3.4), in which case we average the results over many experiments instead.

We perform association tests for each locus by fitting a simple linear model to the genotype matrix X,

$$Y_k^{l,g} = \mu^{l,g} + \beta^{l,g} X_k^{l,g} + \varepsilon_k^{l,g} \text{ where } \varepsilon_k^{l,g} \sim N(0, \sigma_{l,g}^2) \text{ IID} \tag{3.2}$$

for some $\sigma_{l,g}^2$ and computing the P-values of the β-estimate. We then repeat this for $l = 1 \dots L$ and $g = 1 \dots G$. We look at the distribution of the P-values over all LG simulated

sites to see whether there is a significant departure from the expected null distribution. Figure 3.8 demonstrates the results for two contrasting distributions of non-genetic risk. One which has a smooth (Gaussian) distribution in space, and one which has a small, sharp distribution. For risk which is smoothly distributed, test statistics for low frequency variants are less inflated than test statistics for higher frequency variants. On the other had, when the risk has a sharp distribution, lower frequency variants are more inflated for smaller P-values. This tells us that, in the case of non-genetic risk with a sharp boundary, rare variants will have more false positives than common variants. Further, it is clear from Figure 3.8d that the greatest inflation in P-value occurs when the frequency of the variant is approximately equal to the frequency of the risk region. This is not surprising. Intuitively, rare variants are more likely to be highly spatially clustered (Figure 3.5a) and therefore when the non-genetic risk is highly clustered, some rare variants will have spatial distribution almost exactly equal to the spatial distribution of the risk, and they will be strongly associated with the trait. We demonstrate this in Figure 3.9 which shows the distribution of correlation coefficients between genotype and non-genetic risk in the two cases described here. Rare variants cannot be highly correlated with the smoothly varying risk, but they can be highly correlated with risk that occurs in small, sharp patches.

Appendix B demonstrates this behaviour in more complex situations. If we take the sharp distribution of risk from Figures 3.8b and 3.8d, and smooth it out, then the extra inflation in rare variants becomes less pronounced and the frequency at which variants are inflated is more spread out. (Figure B.1). Similarly, if we keep the same amount of risk (in the sense of squared difference in mean), but increase the size of the area over which it is distributed, we see less inflation in rare variants. In this case the inflation is concentrated in variants with frequencies which correspond to the size of the non-genetic risk area (Figure B.2). If we add more small patches of risk of the same size, the effect does not change (Figure B.3). Finally, if we increase the migration rate then the differential confounding effect is significantly decreased. For $M = 10$ the effect is still observable, though barely (Figure B.4).

Another effect worth noticing that when a trait really is controlled by rare variants, those causal rare variants will themselves tend to have a clustered spatial distribution and therefore will be confounded with other, non-causal, rare variants. Thus rare variants can show excess inflation even when there is no spatial non-genetic risk, but there is genetic risk from other, untyped, rare variants (Figure B.5).

Thus far we have demonstrated that, when non-genetic risk has a sharp spatial distribution, rare variants display excess confounding in association studies, relative to common variants. However, given that there are many effective controls for this sort of confounding

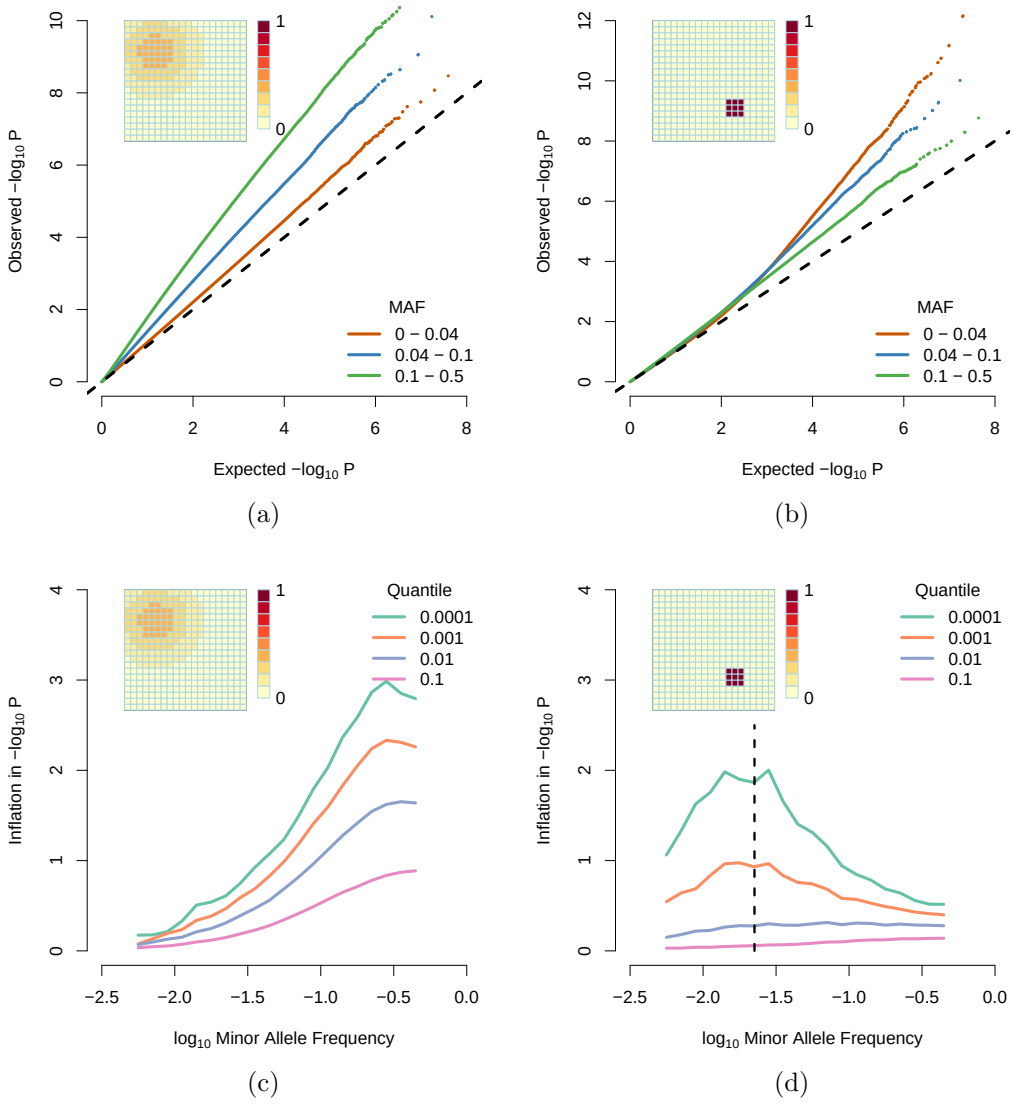


Figure 3.8: Inflation in P-value by allele frequency for an association test in a population living on a grid when there is no genotypic risk but some spatially varying source of non-genetic risk. The grid in the upper left corner of each plot shows the distribution of risk, with the scales next to them showing the effect of the risk on the mean measured in standard deviations of the phenotype. In a and c the risk has a smooth (Gaussian) distribution. In b and d the risk has a sharp boundary. a and b are qq plots of $-\log_{10}$ P-values and c and d show the inflation in $-\log_{10}$ P-value as a function of minor allele frequency for different quantiles. When the distribution of risk is smooth, lower frequency alleles are less inflated than higher frequency alleles. However, when the risk has sharp edges, small P-values are more inflated for low frequency alleles. From d we see that in the case of the small, sharp risk, the inflation is greatest for allele frequencies equal to the spatial frequency of the risk ($\log_{10} 0.225 \approx -1.65$ - dashed black line). These simulations were performed with migration rate $M = 0.01$ and each plot shows the results of 10^8 variants from 10^6 independent trees (i.e. $L = 10^2$ and $G = 10^6$). The maximum non-genetic risk in the smooth case is 0.7 and in the sharp case is 1. Abbreviations; MAF - Minor allele frequency.

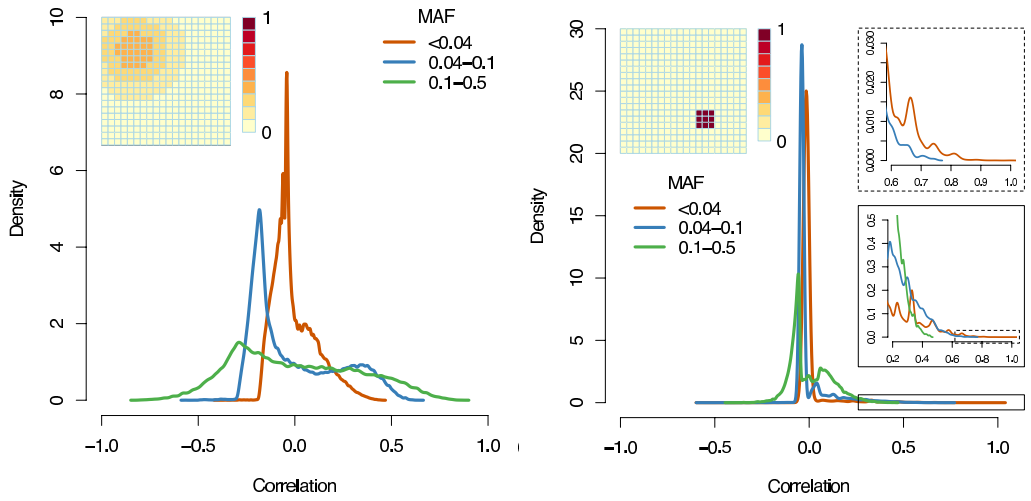


Figure 3.9: Distribution of the correlation coefficients between genotype and non-genetic risk for simulated genotypes and (a) Gaussian risk and (b) small, sharp, square risk. The inset panels in (b) show successive enlargements of the tail of the distribution and demonstrate that there is a long tail of rare variants which are highly correlated with the distribution of the risk. Parameters are the same as in figure 3.8 and in particular, $M = 0.01$. Abbreviations; MAF - Minor allele frequency.

---

in common variants, it's not clear that this would be a problem in practice. In the next section we demonstrate that these controls do not necessarily work in the cases we have described, and therefore that we may need to worry about this effect in real studies.

3.4 Corrections for structure in rare variants

As we mentioned in Section 3.1, there are several effective methods for correcting for structure in association studies. In this section, we investigate whether these corrections will work in the situation described above where sharp spatial distributions of nongenetic risk lead to differential inflation between rare and common variants, and in particular, make rare variants display excessive inflation.

In order to investigate this, we sample genotypes for multiple loci and genealogies as described above. However, now we sample only one realisation of the quantitative trait $Y_k \sim N(R_{\phi(k)}, 1)$ to use for each l, g. We compute single marker test statistics as described above and then perform the following corrections.

1. Genomic control. We take the P-values for each locus, $p^{l,g}$ and compute chi-squared statistics $X^{l,g} = F_{\chi^2}^{-1}(1 - p^{l,g})$ where F_{χ^2} is the cumulative distribution function on the chi-squared distribution with one degree of freedom. We then compute ad-

justed test statistics $\tilde{X}^{l,g} = \frac{X^{l,g}}{\lambda_{GC}}$ where the genomic inflation constant $\lambda_{GC} = \frac{\langle X^{l,g} \rangle}{F^{-1}_{\chi^2}(0.5)}$ and $\langle \bullet \rangle$ represents the median operator. Finally we compute adjusted P-values $\tilde{p}^{l,g} = 1 - F_{\chi^2}(\tilde{X}^{l,g})$.

2. Principal component analysis. We compute the principal components of the $LG \times C$ genotype matrix $X = \{X^{l,g}_k\}$, say $P^1 \dots P^2$ and then fit the linear model

$$Y_k = \mu^{l,g} + \beta^{l,g} X^{l,g}_k + \sum_{i=1}^{D} \gamma^{l,g}_i P^i + \varepsilon^{l,g}_k \text{ where } \varepsilon^{l,g}_k \sim N(0, \sigma^2_{l,g}) \text{ IID}$$

for some $\sigma^2_{l,g}$ and where D is the number of principal components we want to use. We then test the significance of the β-estimates as before. We have used 10 principal components. Using more reduces stratification, but at the cost of decreasing power to detect true associations. We also tried PCA, but calculating the principal components only from rare markers with MAF<4%.

3. Mixed models. The linear mixed model has the form

$$Y = \mu + \beta X + \varepsilon_G + \varepsilon_R \text{ where } \varepsilon_G \sim MVN(0, \sigma^2_G R) \text{ and } \varepsilon_R \sim MVN(0, \sigma^2_R I)$$

for some σ^2_G and σ^2_R where R is a fixed kinship matrix. Here we use the correlation matrix of the genotype vectors as an approximation to the kinship matrix. In this model, as well as the uncorrelated error term from the simple linear regression case, we have an additional error term with fixed correlation structure between individuals (the matrix R, in this case). The proportion of the residual variance explained by these *fixed effects* is $\frac{\sigma^2_R}{\sigma^2_R + \sigma^2_G}$. If we know this proportion then fitting the model is relatively easy, but estimating it can be computationally intensive.

We fitted the mixed model using the package *MMM* (Pirinen *et al.*, 2012).

The results of applying these methods in the two situations considered above are shown in Figure 3.10. In fact, none of the standard corrections work when the inflation is due to the small, sharp non-genetic risk distribution. n the case of genomic control this is because most variants are unaffected by this risk, so looking at the median does not work since it is the tail of rare variants which are highly inflated. For PCA and mixed models, we correct using average relatedness over the whole sample and again, for the highly clustered rare variants which deviate from this, the correction is not sufficient.

Given this, is there any way to correct for this structure? In principle, we could correct all inflation by using a sufficient number of principal components, but at a significant cost to power. Here we found that between 20 and 100 would be sufficient, but in practice

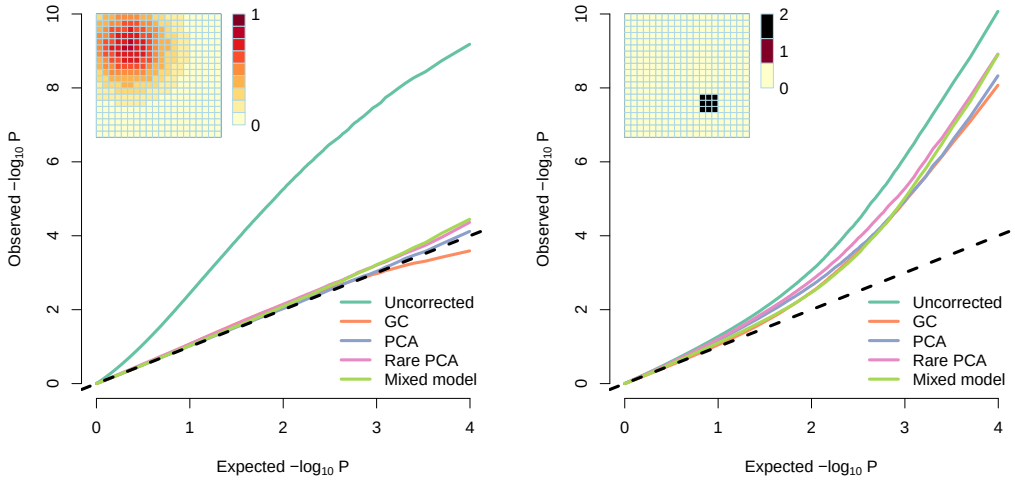


Figure 3.10: QQ plots showing the effect of standard corrections for population structure applied to the situations described above. In the case of smooth Gaussian risk, all methods are effective at correcting for structure, but in the case of small, sharp risk, none of them can correct. The simulations use the same parameters as in Figure 3.8 but the non-genetic risk is twice as intense. The plots show the pointwise average of 100 experiments, each one testing 10^4 loci ($L = 10$ and $G = 10^3$).

---

this would remove almost all our power to detect true associations. Listgarten *et al.* (2013) found that, in the setting described here, their *FastLMM* method would remove this inflation while retaining power to detect rare variants which were not spatially clustered, although it would lose power to detect variants which were spatially clustered as most rare variants are. In fact, in general it is hard to see that it would ever be possible to design a method which would not do this, at least in an association study design. Family based studies, though generally less powerful, are not affected by this kind of structure, so might be an effective way to follow up rare variant associations. One attractive strategy is to scan for rare variants in an association study, then go back and follow up rare variant associations in the families of individuals who carry that variant. This means that we can at least replicate associations in a way which is not affected by structure.

Another type of study design involves "burden" or "collapsing" tests, which seek to increase power by collapsing rare variants over a genomic region or pathway of interest. These are popular ways to conduct association studies on rare variants and it is natural to ask whether these will go some way towards ameliorating the effect of spatial structure on rare variants. There are many complex burden test designs, but we implemented one of the simplest (Morris and Zeggini, 2010) as an illustration. We simulate variants $X = \{X_k^{l,g}\}$ as described above, but retain only rare variants (MAF<4%). We imagine that each genealogy

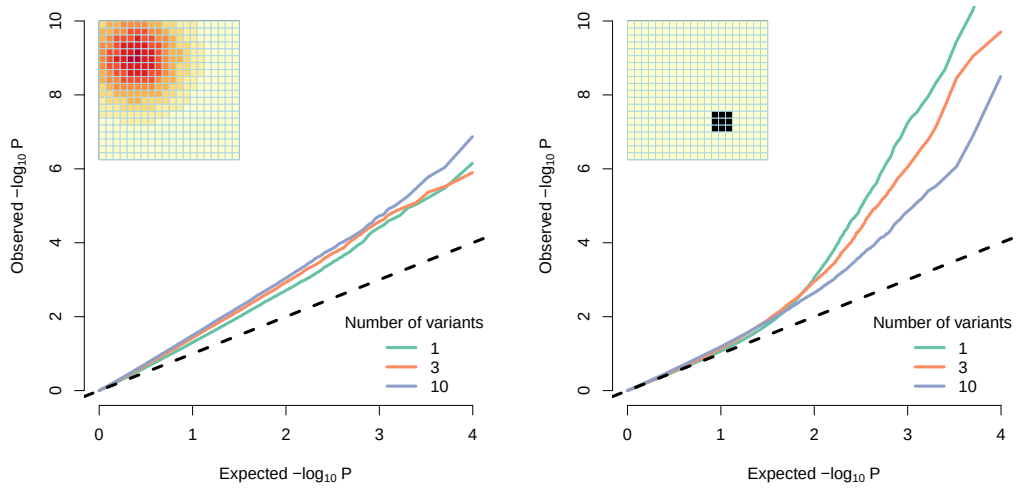


Figure 3.11: Inflation in P-values for the rare variant burden test described in Appendix 3.3 for the two different risk distributions described in the main text. The different lines on each graph represent different numbers of rare variants tested in each gene. Note that in the case of small, sharp risk (b), the burden test goes some way towards removing the effect, but does not do so entirely. Increasing the number of variants to more than 10 does not change the results. Parameters are the same as in Figure 3.10.

---

g represents an independent genomic region (a gene, say) and that each of the L variants on that genealogy represents a rare variant that segregates in the populations. We then compute the rare variants load B_k^g for each individual for each region by counting the number of derived alleles on each genealogy, so that $B_k^g = X_k^{\bullet,g}$. We simulate traits as described above and test association with the rare variant load by fitting the model

$$B_k^g = \mu^g + \beta^g B_k^g + \varepsilon_k^g \tag{3.3}$$

where $\varepsilon_k^g \sim N(0, \sigma_g^2)$. The burden test does slightly reduce inflation due to the sharp distribution of risk (Figure 3.11), but does not remove it entirely. This is because even one rare variant, highly correlated with the non-genetic risk, is enough to make the test significant. More sophisticated tests might do better, but on the other hand, probably at a cost to power.

3.5 Conclusion: rare variants in association studies

We have shown that, under certain conditions, rare and common variants exhibit differential patterns of stratification in association studies. However, these results are qualitative, and we must also ask whether these conditions are likely to be met in practice. This leads to

two questions. First, is the amount of structure we simulated realistic and second, are the risk distributions we used realistic?

As we discussed in Section 3.2, the populations we investigated in this chapter are probably more structured than most human populations. The level of differentiation across the space is (roughly, in terms of F_{ST}), equivalent to the differentiation between continents. However, note that even with a relatively low level of structure, there is still noticeable excess sharing of rare alleles between spatially close populations (Figure 3.6). Since the extent of excess allele sharing increases as frequency decreases, this implies that even in a relatively unstructured population, we just need to look at sufficiently rare variants in order to find excess sharing. Therefore what matters is not just how much structure there is in a populations, but the rarity of the variants in which we are interested. The frequency of variants which can be examined will depend on the population and, though simulations like this can provide a qualitative guide, in practice the data for each study will have to be carefully examined to determine how much excess sharing there is. Ideally information on the geographic origin of the samples could be used, although this is not always available. Even so, excess sharing and the corresponding effect on association studies has been observed in real data. Babron *et al.* (2012) found that in the WTCCC dataset (Wellcome Trust Case Control Consortium, 2007), there is excess sharing of rare variants within geographic regions and that this led to the effect described here, with more confounding of rare variants in both simulated and real data (type 2 diabetes case-control data), which was corrected neither by genomic control nor mixed models.

The second question about whether these simulations are realistic is to what extent the sharp spatial distributions of risk are likely to occur. There are three ways in which nongenetic risk might show sharply defined boundaries of the type for which we have shown differential inflation. First, localised environmental exposure could be highly patchy, for example, associated with urban areas. Second, there could be systematic measurement bias at a single recruitment centre. Third and more subtly, there could be local variation in recruitment policy or rates of misclassification (the effect of which can be thought of as changing the background disease risk). Although we have simulated quantitative trait data, case-control studies are subject to the same issues of population structure, and a case-control study that randomly misclassifies cases and controls will bias effect-size estimates (Copeland *et al.*, 1977). When this misclassification is restricted to a particular spatial area, for example, a single recruitment centre in a large study, it will produce the effects described here. In fact, if we added additional disconnected small areas of risk of the same size as the first, the inflation in P value had the same distribution with respect to frequency (Figure B.3), and this observation can be extended to the case where multiple collection

centers are making biased measurements or random misclassifications. Because the extent and clustering of nongenetic risk will differ between phenotypes and study designs, it is not possible to predict any general influence of differential stratification. The principal problem with trying to account for known nongenetic risk factors (by including them as covariates within the analysis) is that, although information about broad-scale risk factors may be available, it is typical that the more localised a risk factor is, the less we are likely to know about it and the greater effect this lack of knowledge will have on rare variants.

Given this, we expect that problems of extreme confounding in rare variants will become a serious issue when dealing with sequence data. What approaches can be taken to guard against its effects? As we mentioned earlier, one approach is to use methods that are robust with respect to stratification (although at a cost to power and ease of experimental design), such as family-based association, perhaps only for replication. Another is to adapt existing methods to work better with rare variants. For example, although PCA with rare variants did not effectively control inflation if we linearly corrected using the top components, in principle, more sophisticated methods for selecting nonlinear functions of components could correct appropriately. Both Listgarten *et al.* (2013) and Sul and Eskin (2013) report such methods, which appear to be effective, although they lose power to detect variants that are spatially structured, which most causal variants are likely to be. Alternatively, we might look to the development of new measures of relatedness that are more sensitive to recent ancestry and fine-scale structure. Whichever approach is taken, it is likely to require fine-grained information about the geographic origin and recruitment path of each sample. The collection of such information must be an important consideration in the design of future studies.

In this chapter, we made the observation that individuals who share rare genetic variant are, in an equilibrium lattice migration model, likely to come from spatially close locations. However, this is really a compound of two statements. First, individuals who share rare genetic variants are likely to be genealogically closely related. Second, in this particular model, individuals who are genealogically closely related are likely to come from spatially close locations. The first statement is not specific to any particular model of population structure and applies very generally. A natural question is whether we could then invert the argument made in this chapter and use allele sharing (particularly rare allele sharing), to make inference about population structure and history. Rare alleles should be highly informative about recent ancestry, and recent population structure, and sharing of alleles at different frequencies should allow us to make inference about changes in structure over time. We turn to these problems in the next chapter.

Chapter 4

Rare variant sharing, history and structure

In this chapter we develop methods for making inferences about population history and structure, based on patterns of rare variant sharing. In contrast to the previous chapters, we will generally not make strong assumptions about the underlying model. We are particularly interested in population history, expressed as changes in population structure over time. As we discussed in the previous chapter, rare variant sharing is highly informative about recent ancestry, which suggests that this approach will be more useful for making inference about recent history than it will for ancient history.

One of the main motivations for this work is the impending availability of extremely large datasets, on the scale of tens or hundreds of thousands of samples. Even the most efficient existing methods will struggle to deal with datasets of this size. By focusing on the most informative variants, methods like those we describe here are able to make accurate inference, while remaining computationally feasible. In particular, looking at rare variants allows us to make algorithmic simplifications which increase speed and, even more usefully, to reduce memory requirements by using much sparser datasets.

First we provide a brief description of the 1000 Genomes Project and data release, which we will use to illustrate our methods. We begin the chapter proper by investigating the effect of different demographies and population structure on patterns of allele sharing at all frequencies. We demonstrate that these patterns can identify population structure and, qualitatively, assign meaning to it. With the observation that rare variants are much easier to model than common variants, we then focus on patterns of doubleton sharing within the 1000 Genomes data, and compare these to other descriptions of structure like IBD sharing and PCA. Just like these statistics, doubleton sharing can be used to define populations of individuals within the data (i.e. clustering), but with the advantage that it is very easy to assign meaning to these populations in terms of coalescence rates. Finally, we describe

both a method for making inference about the distribution of coalescence times from allele sharing patterns, and a method for making inference about migration rates within a general migration model.

4.1 The 1000 Genomes Project

The 1000 Genomes Project (1000 Genomes Project Consortium, 2010, 2012) has, at the time of writing, collected low coverage whole-genome sequence, high coverage exome sequence, and array genotype data on over 2,500 individuals from 26 populations, with the aim of providing a substantially complete catalog of common variation (MAF>1%) in humans. It provides the only large publicly available dataset of sequenced genomes. In this chapter we will use the phase 1 data release. This is a callset which integrates all three data sources for 1,092 individuals from 14 populations, containing almost 40 million SNPs, indels and large deletions. We will also be using separately the array genotyping data for these individuals, which contains genotypes for around 2 million SNPs. Power to detect common variants is high (over 99% for variants at 1% frequency), but lower for rare variants. For example, power to detect SNPs which occur only once (singletons) is around 30% and and around 66% for SNPs occurring twice (doubletons).

The populations in this dataset are described in Table 4.1. We make the assumption that it makes sense to think about these samples as taken from groups which are, in some sense, distinct from each other. Later we will consider what exactly we mean here by the word "population". It is worth noticing that, while some of these populations look like samples from relatively isolated groups (for example FIN), some of them (for example ASW, and all the American populations) are clearly very recently admixed. This can clearly be seen in the first two principal components of the data which we plotted in Figure 1.5b.

4.2 Allele sharing and structure

We begin by investigating the effect that different population structures have on allele sharing probability. Throughout we will assume that all the variants we look at are selectively neutral and further, that the minor allele is the derived allele[†]. Since the age of an allele is correlated with its frequency[‡], we expect that sharing at different frequencies will reveal

[†]For an allele occurring n times in a sample of N chromosomes, the probability that the minor allele is the derived allele is $1 - \frac{n}{N}$ so for rare alleles in a large sample, is very close to 1. If this assumption proved problematic, we could also estimate the derived allele by looking at an outgroup species, probably chimpanzee.

[‡]An allele at frequency f has expected age $\frac{-2f}{1-f}\log(f)$ (Kimura and Ota, 1973)

population structure at different times in the past. Suppose we have a sample of N chromosomes (or haploid individuals), divided into K distinct populations with population k of size N_k. These are genotyped at C sites and $X_{ic} \in \{0, 1\}$ is the genotype of individual i at site c. Define the pairwise sharing probability p_{ij}^f between two chromosomes i and j as frequency f as $p_{ij}^f = \frac{\sum_{c \in C_f} \mathbf{1}\{X_{ic} \text{ and } X_{jc}\}}{\sum_{c \in C_f} \mathbf{1}\{X_{ic} \text{ or } X_{jc}\}}$ where $C_f = \{c \in C : X_{\bullet c} = f\}$ is the set of variants with frequency f. That is, the sharing probability between two chromosomes is the number of variants at frequency f in the population that are in both chromosomes, divided by the number that are in either. Also define the pairwise sharing probability between two populations at frequency f to be the average of the sharing probabilities for all pairs of individuals in the two populations.

We investigated the behaviour of this quantity under a variety of demographic scenarios by simulating data using the program *ms* (Hudson, 2002). The results (in terms of the population sharing probabilities) are shown in Figure 4.1. In a single, panmictic population, the allele sharing probability at frequency f is simply given by $\frac{f^2}{2f-f^2}$ (Figure 4.1a). Suppose we have two populations but there is migration between them. Then two chromosomes picked from within the same population are more likely to share a variant than two chromosomes from different populations. This is true at all frequencies, but the effect is greater for rarer variants (Figure 4.1b). Decreasing the migration rate reduces the between-population sharing probability (Figure 4.1c). In principle, we could detect demographic events like changes in migration rates by looking for changes in the rate of allele sharing at different frequencies.

Code	Name	Description
ASW	African-American SW	African Ancestry in Southwest US
YRI	Yoruba	Yoruba in Ibadan, Nigeria
LWK	Luhya	Luhya in Webuye, Kenya
MXL	Mexican-American	Mexican Ancestry in Los Angeles, California
PUR	Puerto Rican	Puerto Rican in Puerto Rico
CLM	Colombian	Colombian in Medellin, Colombia
CHB	Han Chinese	Han Chinese in Beijing, China
JPT	Japanese	Japanese in Tokyo, Japan
CHS	Southern Han Chinese	Han Chinese South
CEU	CEPH	Utah residents with Northern and Western European ancestry
TSI	Tuscan	Toscani in Italy
GBR	British	British in England and Scotland
FIN	Finnish	Finnish in Finland
IBS	Spanish	Iberian populations in Spain

Table 4.1: Abbreviations and descriptions for the populations in the 1000 Genomes Project phase 1 data release. These are further grouped into 4 continental-scale metapopulations, corresponding to Africa, America, Asia and Europe. Taken from www.1000genomes.org.

However, one of the problems in using this, or related statistics, is that the sharing probability distribution depends on many factors like the size of the samples and the diversity in each sample at different frequencies. While some of these are probably known, others are not and would be difficult to incorporate into inference. For example, in a simple two population migration model, changing the migration rate changes the allele frequency spectrum (Figure 4.1d). With a small number of populations, this can be modelled. For example, Gutenkunst *et al.* (2009) fit parametric models using the joint allele-frequency spectrum, which is similar. But with many populations, it would quickly become difficult to take account of the distorting effects of this dependency. Events which occurred at the same time historically would appear at different points in the frequency spectrum, depending on the sample sizes and parameters in a complicated and probably non-unique way.

Nonetheless, these statistics can reveal population structure, and provide at least a qualitative description of the process giving rise to the structure, which provides an advantage over PCA. For example, in Figure 4.2 we show an example where two different demographic histories lead to samples with very similar principal components. This is not a surprise. As we discussed in Chapter 1, the principal components can be expressed as a function of the expected coalescence times and therefore any models which give the same expected coalescence times will have the same PC's. However (Figures 4.2c and d), the population sharing statistics for these models have very different distributions and it is clear that the two samples have different histories. In particular, the population split event is clear in Figure 4.2d. In principle then, looking at the sharing statistics provides more information than simply looking at the principal components although, as we discussed above, it is difficult to make quantitative inference using this information without specifying a parametric model, which we are trying to avoid here.

We demonstrated in principal that we can assign some meaning to populations in terms of the relationship between them. In fact we can invert this argument to provide a definition for what we mean by populations. Generally grouping individuals into populations implies some form of exchangeability between individuals in the same population, either in a statistical or a model-based sense (see Lawson (2013) for a discussion of this). When looking at the sharing statistics, we we think of two individuals as being in the same population if they have the same allele sharing probabilities with all other individuals, at all frequencies. This is a consequence of the definition we really want to use, which is that two individuals are in the same population if they have the same distribution of coalescence times with all the other individuals in the sample.

Arising from this then, is the question of whether we can group individuals into populations by looking at the individual sharing statistics and grouping them together. We

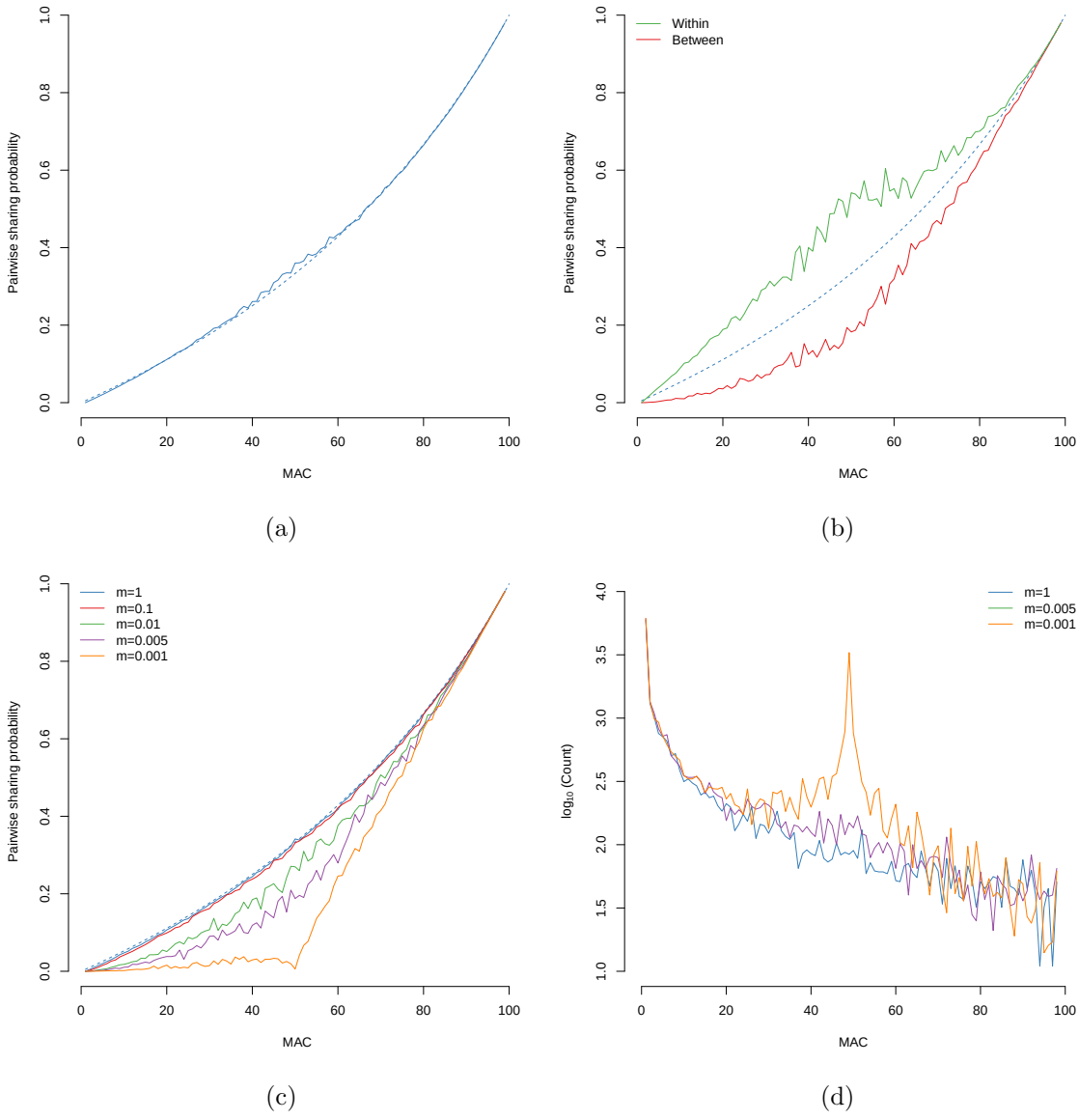


Figure 4.1: The effect of structure on allele sharing probability, as a function of minor allele count (MAC). **a-c**: Expected pairwise allele sharing probabilities under different scenarios, using simulated haplotype data. The blue dashed line is the theoretical prediction for a single population. **a**: A single, panmictic population. **b**: Two islands with constant migration rate m=0.05, comparing pairs within and between populations. **c**: Two-island models, comparing between-population sharing probabilities for different migration rates. **d**: Allele count spectrum for a two-island model for different migration rates. All simulations performed with *ms*, with 100 individuals split into two groups of 50 in **b-d**, 100 regions of length 1000 and $\theta = 5$ (mutations occur on the tree as a Poisson process with rate θ. To convert to a real mutation rate, $\theta = 4N_e\mu$ where μ is the per-base, per-generation mutation rate).

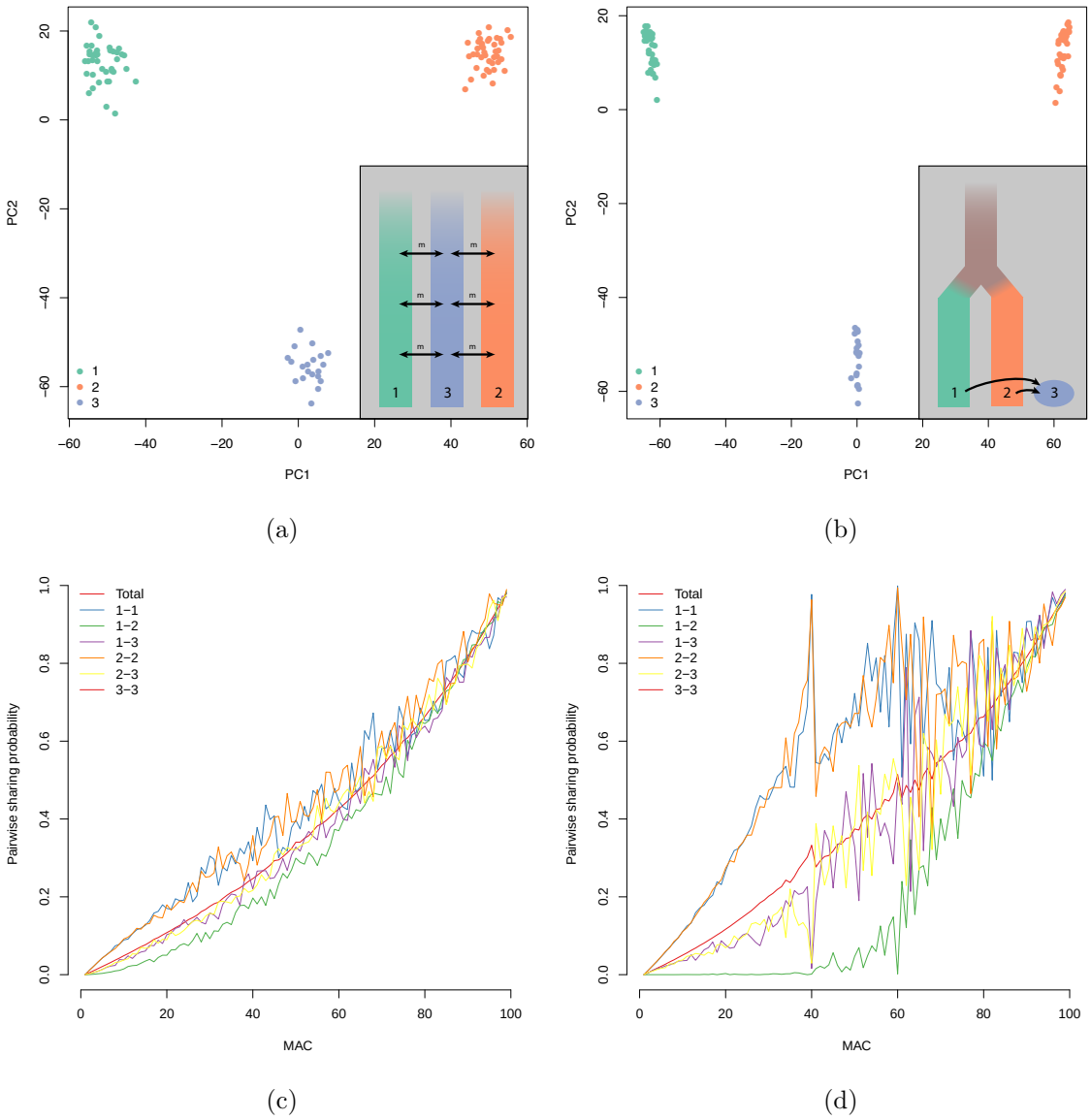


Figure 4.2: Here we compare the principal components of the genotypes (**a**-**b**) and the pairwise sharing statistics (**c**-**d**) for data simulated from two demographic models of three populations (shown in the insets in **a** and **b**). In **a** and **c**, there are three populations arranged in a linear stepping stone model. In **b** and **d**, two of the populations split at time 0.8 (in coalescent time units of $2N_e$) in the past, and the third is an admixture of the two other populations. We simulated data from these models this using *ms* and simulated the admixture by creating new individuals by combining randomly broken chromosomes from the other two populations. **a**-**b**: The first two principal components of the genotypes. Note that the PC's from the two models are very similar, in particular, the centres of the clusters are in the same place. **c**-**d**: The population sharing statistics, organised by population (so "1-2" means the average of all pairs where one member is in population 1 and the other is in population 2). Note that in this case we can clearly tell the difference between the populations and we can assign meaning to them. For example fact that populations 1 and 2 split with no further migration is clear by the lack of sharing at low frequencies.

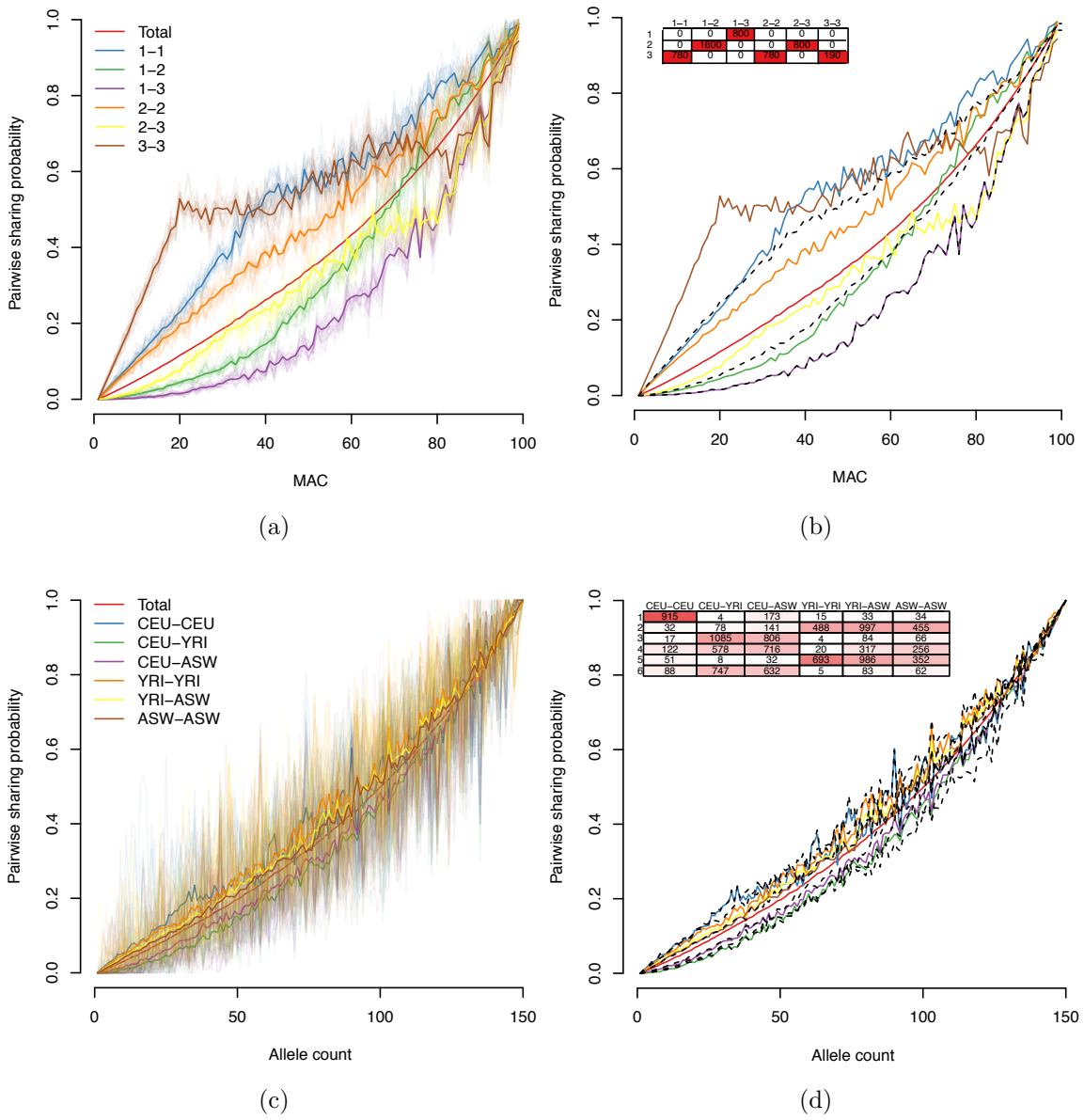


Figure 4.3: Clustering population histories based on allele sharing. **a**-**b**: Simulated data from a model where populations 1, 2 and 3 split at time 1, and there is subsequent migration between populations 1-2 and 2-3. **a**: Population allele sharing, solid lines show the means for population pairs and shaded lines are samples of individual histories to show the variation around the mean. **b**: K-means clustering of the histories, with $K = 3$. The dashed lines show the means, and the inset grid shows the cluster assignments, counting the number of pairwise histories (columns) assigned to each cluster (rows). In this case, each set of histories is assigned a single cluster, though some different sets are assigned to the same cluster. For example, all the within group pairs, 1-1, 1-2 and 1-3 are assigned to cluster 3. **c**-**d**: The same analysis but on 1000 Genomes data. We used 10Mb of Chromosome 1 from 50 chromosomes each of ASW, CEU and YRI and set $k = 6$. While the cluster assignment is not as clean as in **b**, we still largely distinguish between sets of histories, so cluster 1 contains mostly CEU-CEU histories whereas clusters 2,3 and 5 contain CEU-YRI and CEU-ASW clusters, for example.

demonstrate one idea for doing this in Figure 4.3. Here, we simply treat the sharing probabilities as vectors and apply K-means clustering. This groups the pairs of individuals into clusters. It would then be easy to group the individuals themselves into clusters by grouping individuals with similar pairwise relationships. This works reasonably well, though suffers from all the disadvantages of K-means clustering described earlier. A more sophisticated approach is to model the sharing explicitly. That is, let T_{ij}^f be the total number of variants at frequency f that either individual i or individual j have, and let S_{ij}^f be the number they share, so that $p_{ij}^f = \frac{S_{ij}^f}{T_{ij}^f}$. Then we model S_{ij}^f as binomially distributed,

$$S_{ij}^f \sim B\left(T_{ij}^f, \theta_{\phi(i,j)}^f\right) \tag{4.1}$$

where $\theta_1^f \dots \theta_k^f$ are a set of k maps from f to $[0,1]$ which we call sharing trajectories, and $\phi : \{1 \dots N\} \times \{1 \dots N\} \to \{1 \dots k\}$ is a function that maps pairs of individuals to trajectories. This binomial mixture model can be fitted using an EM algorithm. Though we need to specify k, we can attempt to find the optimal value by fitting for different values of k and using some criterion like AIC or BIC to distinguish between them[†]. This model is implemented in the *flexmix* (Leisch, 2004; Grün and Leisch, 2008) package in *R*, and works well on small examples though it is too time and memory intensive to run on large datasets so we did not pursue this approach further. However, slightly more approximate methods such as Gaussian mixture models would probably be much faster and might provide a promising approach.

The analyses we've described could be extended, but they are computationally intensive and it would be difficult to extend them to large genome-wide datasets. However we know, both theoretically and from the results of the previous chapter, that rare variants are highly informative about ancestry. As well as this, looking at rare variants removes some of the issues we have analysing common variants, reduces the amount of data required, and allows approximations and simplifications which would not be possible if we tried to consider the whole frequency spectrum. It is reasonable to think, therefore, that if we looked only at rare variants, we would be able to reconstruct most of the structure in the data, but with simpler and less costly algorithms. Therefore we focus for the rest of the chapter on analyses based on sharing of rare variants. In fact, we will only consider doubletons, since this allows us to make even more simplifications in the analysis. We begin by looking at doubleton sharing in the 1000 Genomes dataset, and compare with other measures of relatedness.

[†]These are methods for choosing among models with different numbers of parameters by penalising the likelihood. Akaike's information criterion (Akaike, 1974, AIC) penalises by the number of parameters and the Bayesian information criterion (Schwarz, 1978, BIC) penalises by the number of parameters times half the log of the sample size.

4.3 f_2 sharing in the 1000 Genomes

Taking the integrated callset from the 1000 Genomes dataset (filtered to include only bialleleic autosomal SNPs), we counted the number of doubleton (or f_2) variants shared by each pair of individuals. Let d_{ij} be the number of f_2 variants shared by individuals i and j, so they both have genotype 1 (heterozygous) if $i \neq j$ or genotype 2 (homozygous derived allele) if $i = j$. We also calculated a statistic normalised by the total number of f_2 variants, $\tilde{d}_{ij} = \log_{10}\left(\frac{d_{ij}}{d_{\bullet j}}\right)$. Figure 4.4 shows this quantity, normalised to have zero mean and unit variance.

We also called IBD chunks on the same samples, but this time using only the array genotype data rather than the sequence data. We used the *fastIBD* method of *Beagle* (Browning and Browning, 2011) and, as suggested, ran ten times with different random seeds, combined the results, and kept only high confidence chunks larger than 1Mb. Let l_{ij} be the total length of shared chunks between individuals i and j. Then, as for the f_2 variants, we calculated a normalised statistic, $\tilde{l}_{ij} = \log_{10}\left(\frac{l_{ij}}{l_{\bullet j}}\right)$. Normalised, this is shown in Figure 4.5.

For comparison, we also show in Figure 4.6 the covariance matrix of genotype vectors which would be used in the calculation of PCA, for example. The first thing we notice about this matrix is that it seems to show significantly less fine detail than the other two measures.

There are a number of interesting features which are visible in both the f_2 and IBD sharing matrices (and, to some extent, in the covariance matrix). The large scale continental structure is obvious, with Asian, African, European and American blocks appearing. The American block is close to both the European and African blocks, reflecting its admixed origin. Similarly, the African American (ASW) block shows both European and American ancestry. On a smaller scale, individual populations are generally quite distinct, with the possible exception of the two Chinese populations, CHS and CHB. Within Europe, the CEU population is very close to some of the British (GBR) samples, but the Tuscans (TSI), Spanish (IBS) and particularly the Finns (FIN) are all quite distinct. The Spanish clearly share more ancestry with the American populations than other Europeans, which is not surprising given the history of these populations. The American populations are all quite distinct, and we can see that the Puerto Ricans (PUR) tend to have more African ancestry than the Columbians (CLM) or Mexicans (MXL), and that the Mexicans have more of what looks like Asian ancestry, probably reflects ancestral Native American ancestry.

Some of the populations show even finer scale structure, most notably the British (GBR) and one of the Chinese populations (CHB). This is probably a result of the sampling strate-

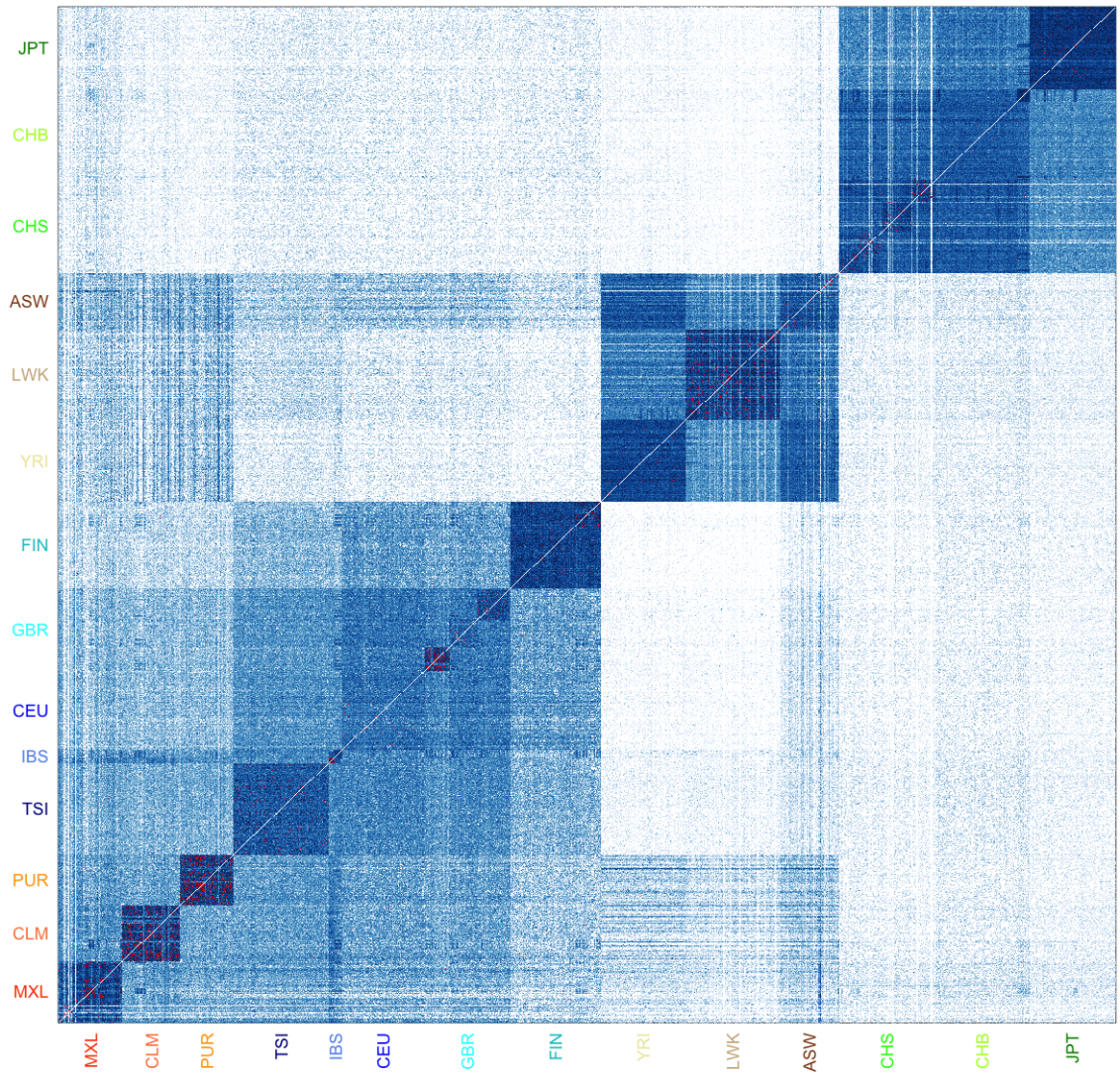


Figure 4.4: Matrix of normalised f_2 variant sharing, as described in main text. The matrix is 1092×1092, with one cell for each pair of individuals. Values have been normalised to have mean 0 and variance 1, and the scale runs from < -2 (white), through 2 (darkest blue) to > 2 (red). Populations are labeled along the edges (see Table 4.1 for definitions). We have ordered individuals by population but not otherwise within population.

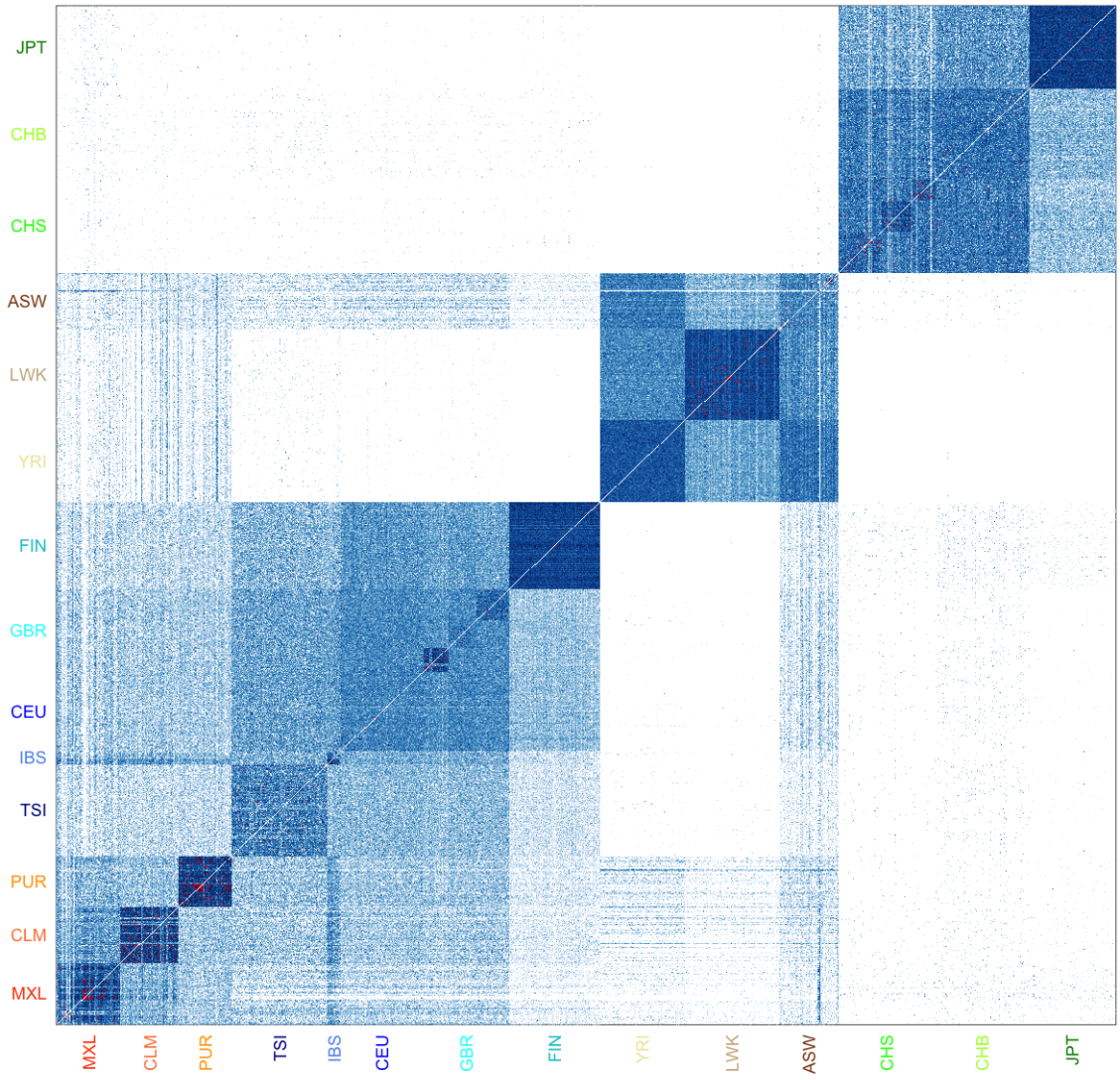


Figure 4.5: Matrix of normalised IBD sharing, as described in main text. The matrix has been normalised as to be comparable with Figure 4.4.

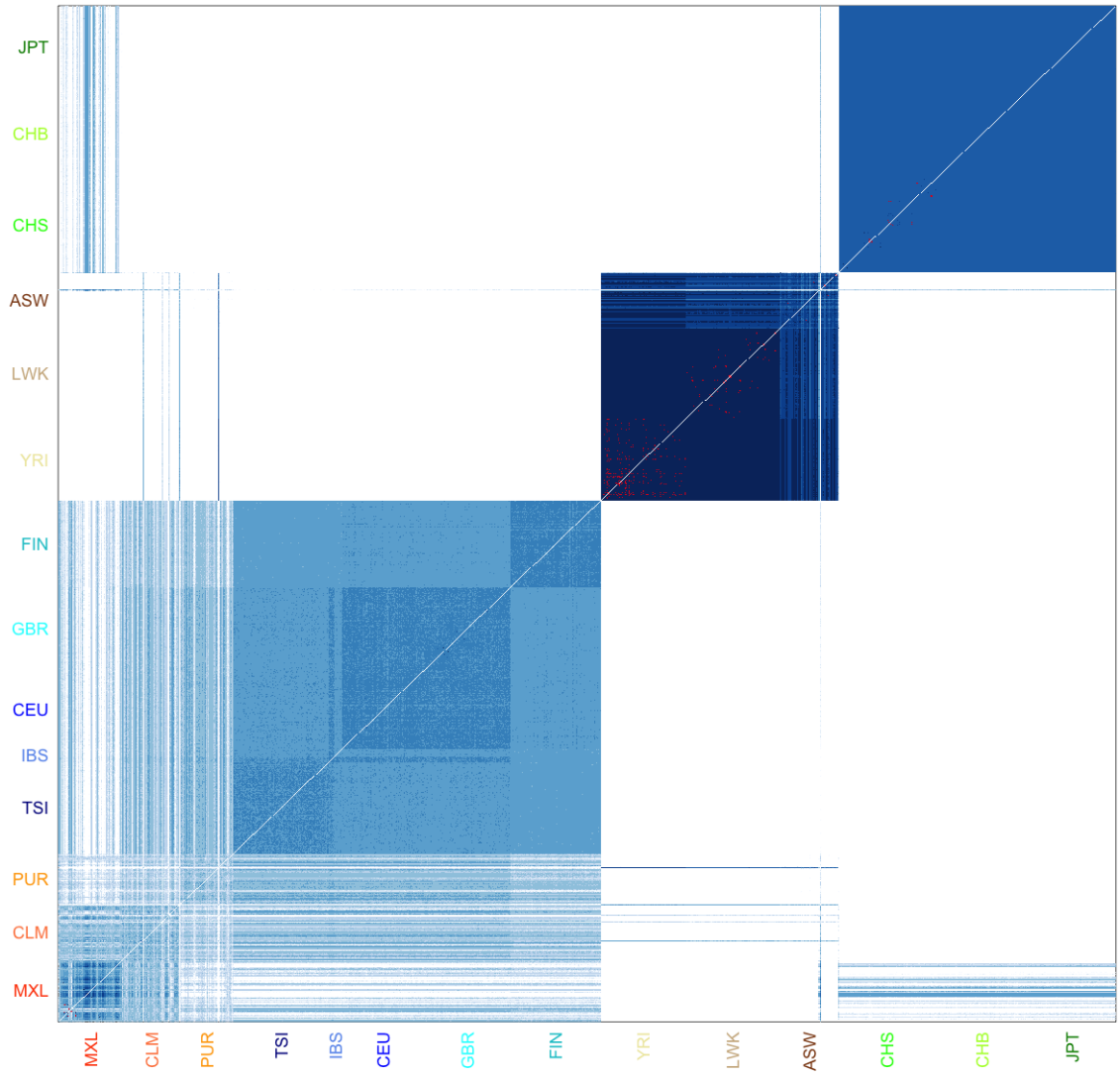


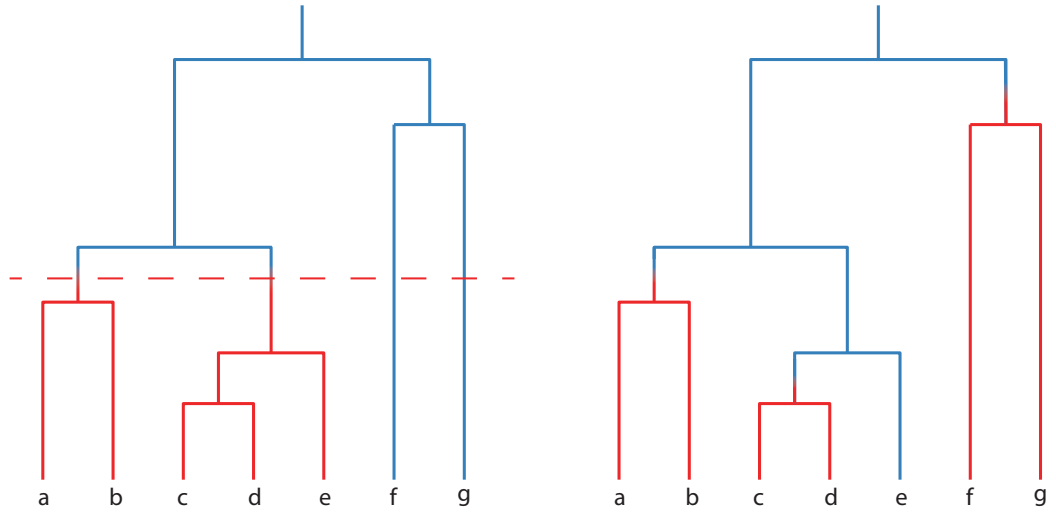
Figure 4.6: Genotypic covariance matrix. Again, the matrix has been normalised and the colour scheme chosen so that it is comparable with Figures 4.4, and 4.5.

gies employed for these populations. In fact we know that at least some of the GBR samples are from the Orkney islands and it seems likely that one of the three subclusters within GBR is made up of these individuals. There also seems to be substructure within the American populations, though this is probably more continuous and reflects variation in admixture proportions. There appears to be a sub-block in the Chinese samples in the f_2 but not the IBD matrix. This may be due to systematic sequencing or variant calling errors on a subset of samples. Finally, on an individual level, we can see that some individuals appear to be genetically closer to other populations than the ones to which they actually belong. For example there is one African American sample which looks Mexican and one Puerto Rican sample which looks African American (both of these are clearly visible in the PCA plot in Figure 1.5b).

Just as interesting as these similarities are the differences between the matrices. The most obvious difference between the IBD and f_2 matrices is that, relative to IBD sharing, there is more f_2 sharing in the Asian-European, African-European, and American-Asian areas. This has a simple genealogical interpretation. The i, j^{th} entry of the f_2 sharing matrix can be interpreted as being proportional to the matrix of probabilities that the first coalescence of both lineage i and j is with each other, whereas the i, j^{th} entry of the IBD sharing matrix is the probability that lineages i and j have coalesced by a specified time (Figure 4.7). As a consequence, f_2 sharing can "see" older coalescence events, whereas IBD ignores older events. The areas where there is more f_2 sharing represent older ancestry. On the other hand, IBD sharing can detect events which are recent but not the most recent (lineage e in Figure 4.7), which f_2 sharing doesn't, though we could see these events by looking at variants at higher frequency. Thinking about f_2 sharing in this way as detecting genealogical nearest neighbours is similar conceptually to the haplotype copying model of Li and Stephens (2003) and its recent implementation by Lawson *et al.* (2012).

In practice, the relatedness matrices constructed from IBD and f_2 sharing are similar (Figure 4.8a, the correlation between entries is 0.58). Part of the difference is due to the effect described above where they are measuring different things, and part is due just to random variation. The distribution of IBD tract lengths is roughly exponential, at least for chunks larger than 1Mb (Figure 4.8b)[†], and we would expect that longer tracts contain more f_2 variants. Figure 4.8c plots the probability that tracts of a given length contain f_2 variants. In general, the probability that a tract contains any f_2 variant is high (red line). However the probability that a tract contains a variant which is shared between the two individuals who are IBD is relatively low (blue line). This is partly because of the IBD tracts where more than two individuals are IBD at the same place, and therefore cannot

[†]The true theoretical distribution is complex, see Palamara *et al.* (2012).



(a) IBD (b) f_2

Figure 4.7: Contrasting genealogical interpretations of IBD and f_2 sharing. **a**: IBD sharing. If two lineages coalesce before a specified time (red dashed line), then they are IBD. So here, a and b are IBD, and c,d and e are IBD. The height is not generally specified directly, but implicitly by specifying chunk size. Very approximately, a 1Mb chunk corresponds to a time on the order of 100 generations ago. **b**: f_2 sharing identifies nearest neighbours. So here, each of pairs a-b, c-d, and f-g might all carry an f_2 variant.

all share f_2 variants. If we restrict to tracts where there are only two individuals IBD, the probability of finding a consistent variant is much higher (green line), about 0.35 for 1Mb tracts, rising to over 0.8 for 3Mb tracts. One final interesting observation is to compare for each individual the total number of f_2 variants and the total length of IBD tracts (i.e. $d_{i\bullet}$ and $l_{i\bullet}$ in our original notation). This shows clear inter-population differences which we can relate to their history. Having extensive IBD indicates a small recent population size or bottleneck (in this case, roughly in the past 100 generations). Conversely, having many f_2 variants shows a population with high diversity, suggesting a large ancestral population size. The Finns (FIN) are clear outliers, with a lot of IBD. On average, each Finnish individual has 300% of their genome covered by IBD chunks, but they have relatively few rare variants, consistent with a strong founder bottleneck. The Luhya (LWK) have both a lot of IBD and many rare variants, indicating a large ancient population and a small recent population. In contrast, the two Chinese populations have little IBD and few rare variants, suggesting a small ancient and large recent population size[†].

So far, we've looked at rare variant sharing in a fairly qualitative way, identifying structure and comparing with the most other popular sharing measure, IBD. We could certainly

[†]The fact that there are no populations which have both a large ancient and and a large modern population is something of a comment on the rise and fall of human civilisations.

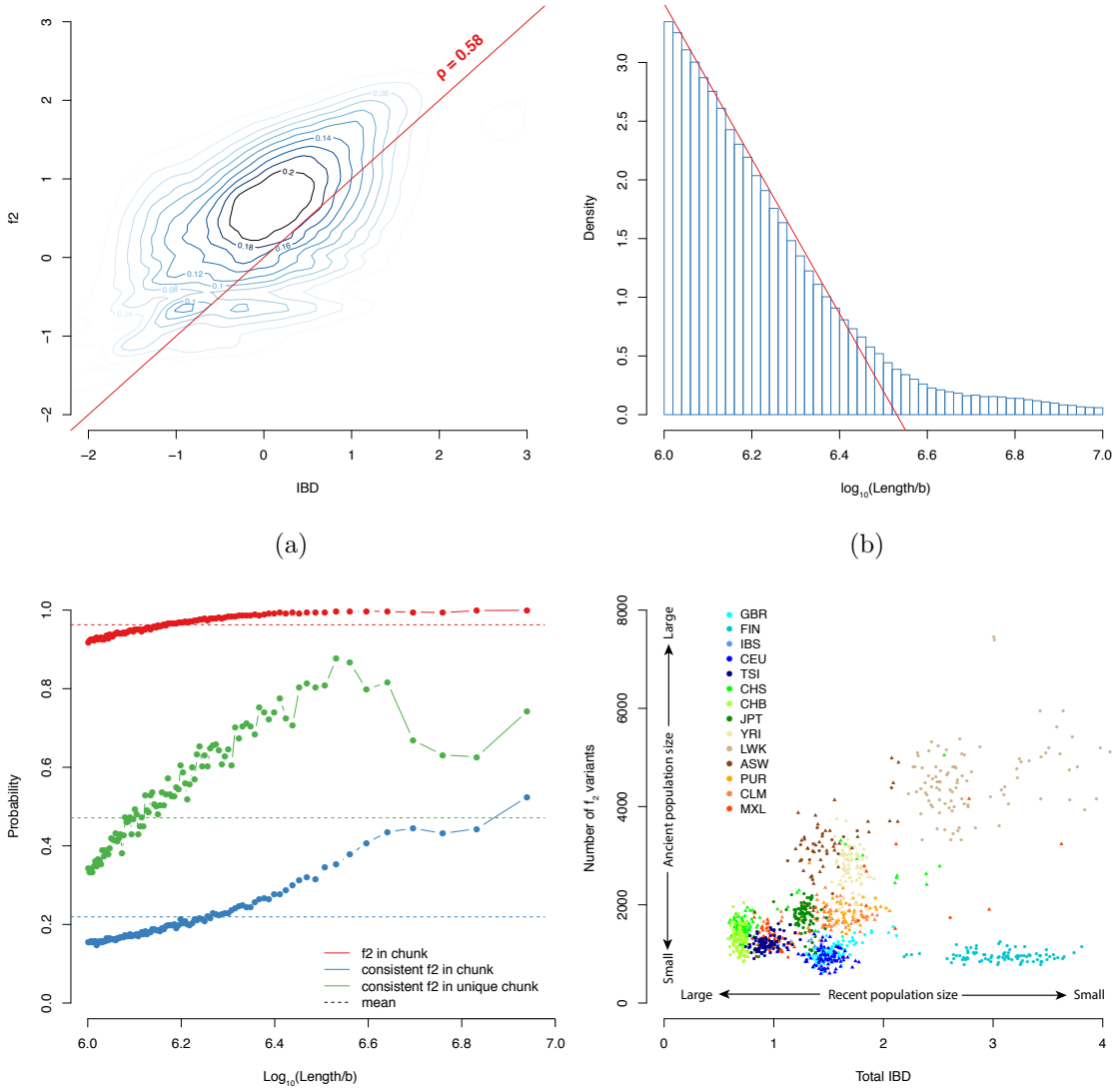


Figure 4.8: Comparisons of IBD and f_2 sharing. **a**: Kernel density plot of the joint distribution of the normalised values in Figures 4.4 and 4.5. The red line is the line $y = x$. **b**: The distribution of the length of the IBD tracts between 1 and 10 Mb. **c**: Probability that an IBD tract of a specified length contains an f_2 variant anywhere (red line), an f_2 variant which is consistent with the IBD tract in the sense that the individuals sharing the variant are the ones which are IBD (blue line) and (green line) a consistent variant, conditional on a tract where only two individuals are IBD. Each point is a bin containing an equal number of IBD tracts. **d**: Relationship between total length of IBD tracts and total number of f_2 variants. IBD is expressed in units of total autosome length ($\approx$ 2.8Gb). For each individual we summed the length of all the chunks shared with any other individual, and divided by 2.8Gb.

use this information in clustering, to identify structure within a population, and the comparison to IBD sharing suggests it will perform well. However, a more interesting use of this information is to make explicit inference about the relatedness between populations and their relative histories, which is the subject of the next section. In order to do this, we use f_2 sharing to find shared haplotypes, and then estimate their age. The ages of haplotypes shared between populations then tell us about the historical coalescence rates of individuals in those populations.

4.4 Estimating haplotype ages from f_2 sharing

In this section we describe a method for estimating the age of f_2 variants, by finding and dating shared haplotypes between the two genomes. Given a sample of individuals, we could use this to cluster individuals into populations, based on all individuals in the population sharing the same distribution of coalescent times with the rest of the samples. Alternatively, if the populations are known *a priori*, we can use this method to infer the historical relationships between the samples. We then apply this method to the 1000 Genomes data. In practice this is similar to the method of Ralph and Coop (2013) which gets the same information by dating IBD chunks found with *Beagle*. However there are several difference between the methods. First, Ralph and Coop (2013) restrict themselves to coalescent events on the order of less than 100 generations ago. In principal, by using rare variants, we can find much older haplotypes, providing they carry unique mutations. Second, we do not need to phase the genotype data, removing one possible source of error. Finally, we incorporate an additional source of information in the form of the number of mutations on the haplotype, which is informative about the time to common ancestry. Our approach is similar in spirit to that of *Chromopainter* (Lawson *et al.*, 2012), in that it finds chunks where two sequences are genealogical nearest neighbours.

4.4.1 Method

Definitions

Suppose we have a sample of genotypes from N individuals from a single contiguous genetic region. Define an f_2 variant to be one which occurs exactly twice in the sample in different individuals. That is, either two individuals have genotype 1 and all the others have genotype 0, or two individuals have genotype 1 and the others have genotype 2. As before, we assume that the minor allele is the derived allele. Throughout we assume that the effective population size N_e is known and constant.

Define an f_2 haplotype shared between chromosomes a and b to be a region satisfying the following two conditions: 1) The time to most recent common ancestor (TMRCA) of a and b does not change over the region. 2) At one or more sites in the region, a and b coalesce with each other before either of them coalesce with any other chromosome. In other words, they are unique genealogical nearest neighbours. Additionally, we say that individuals i and j ($i \neq j$) share an f_2 haplotype if a is one of i's two chromosomes and b is one of j's two chromosomes.

The problem we solve is to determine the age of the f_2 haplotypes or equivalently, the coalescence time of a and b which we call t in units of generations ($\tau = \frac{t}{2N_e}$ in coalescent time). Since each f_2 variant must lie in an f_2 haplotpye, the variants provide a simple way of detecting the haplotypes. We first describe a simple algorithm to find regions which are strictly larger than the f_2 haplotypes. Then the next problem is to determine the likelihood of the age. We describe how to do this exactly in a simplified case, and then an approximate method for the full case.

Detecting haplotypes

Given a sample of individual genotypes vectors, we first find all f_2 variants. That is, variants which have exactly two copies (in different individuals) in the sample. Suppose individuals i and j share an f_2 variant at genomic position x. Then we know that one chromosome from i must be the unique genealogical nearest neighbour of one chromosome from j at that point and therefore lies on an f_2 haploype. Then for each variant in turn, we scan left and right along the genome from x, until we reach points where i and j have inconsistent homozygote genotypes. These points define a region which must be larger than the f_2 haplotype. We record the genetic and physical lengths (L_g in Morgans and L_p in bases) of this region, and the number of singletons (variants occurring exactly once in the sample) S that each individual carries in this region (Figures 4.9 and 4.10).

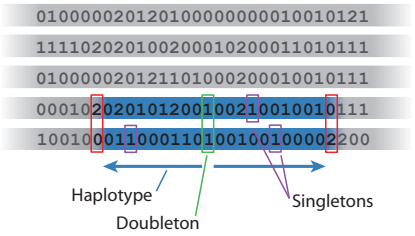


Figure 4.9: Detecting haplotypes from genotypes, using f_2 sharing.

Exact case

Suppose we knew the exact lengths of and number of singletons on the f_2 haplotype, rather than the region defined above. Call these exact quantities L_g^*, L_p^* and S^*. Let the age of this haplotype be t generations, or τ in coalescent time ($\tau = \frac{t}{2N_e}$). Then, for a randomly chosen

haplotype, L_g^* has an exponential distribution with parameter $4N_e\tau$[†] and S^* has a Poisson distribution with parameter $\theta L_p^* \tau$ where $\theta = 4N_e\mu$ and μ is the per-base per-generation mutation rate which is approximately 1.2×10^{-8} for humans (Scally and Durbin, 2012). Therefore (ignoring a constant term), the log-likelihood of τ given L_g^*, L_p^* and S^* is

$$\ell\left(\tau; L_g^*, L_p^*, S^*\right) = (1 + S^*)\log(\tau) - 4N_e\tau L_g^* - \theta L_p \tau$$

and the maximum likelihood estimate of t is therefore

$$\hat{t} = \frac{1 + S^*}{2\left(L_g^* + \mu L_p^*\right)}. \tag{4.2}$$

Note that this does not depend on N_e, implying that in the full case, we will depend on N_e only through the error terms, and therefore that mis-specifying N_e will not have a large impact on the results.

Approximate likelihood for genetic length

We make two corrections to the likelihood for genetic length. The first relates to the ascertainment process of the haplotypes, and the second to the overestimate in the length due to the way we detect the endpoints.

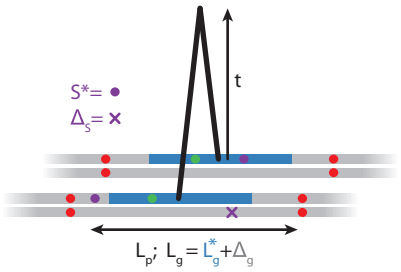


Figure 4.10: Model used to estimate the age of f_2 haplotypes.

The ascertainment problem is as follows. Suppose we pick a haplotype at random, then its length is exponentially distributed (i.e. gamma with shape parameter 1). However, if we pick a point on the sequence at random then the distribution of the length of the haplotype in which it falls is gamma distributed with shape parameter 2. This is an example of the "inspection paradox" and it is because in the second case, we are sampling haplotypes effectively weighted by their length. In our case, we detect haplotypes if they contain one or more f_2 variants. Therefore the probability that we find a haplotype is increasing with its physical length (because longer haplotypes are more likely to carry f_2 variants, but non-linearly. The probability is also increasing with genetic length, but in a complex way that depends on the variation of recombination and mutation rate along the genome, and on the age of the haplotype (older haplotypes are likely to have longer branches above them, and therefore to have more f_2 mutations). Rather than trying to take all of these effects into account, we made the simplifying assumption that we could model the genetic length L_g^* as a gamma

[†]To see why, note that the probability that a short segment of length L_g^* has no recombination in t generations is $\left(1 - L_g^*\right)^{4N_e\tau} \sim e^{-L_g^* 4N_e\tau}$ when $L_g^* 4N_e\tau$ is small.

distribution with shape parameter k where $1 < k < 2$ and rate $4N_e\tau$. Simulations suggested that k around 1.7 was optimal, and we used this value throughout.

The second correction involves correcting for the overestimate of genetic length. We tried to detect the ends of the haplotype by looking for inconsistent homozygote genotypes, but of course in practice, after the end of the f_2 haplotype, there will be some distance before reaching such a site. This (genetic) distance Δ_g is the amount by which we overestimate the length of the haplotype. We estimate the distribution of Δ_g for a given sample by sampling pairs of genotype vectors, then sampling sites at random and computing the sum of genetic distance to the first inconsistent homozygote site on either side. We then fit a gamma distribution with (shape, rate) parameters (k_e, λ_e) to this distribution. The likelihood of τ is then given by the convolution density of L_g^* and Δ_g,

$$L(\tau; l_g) = \int_0^{l_g} f_\gamma\left(x; (1.7, 4N_e\tau)\right) f_\gamma\left(l_g - x; (k_e, \lambda_e)\right) dx \tag{4.3}$$

where $f_\gamma\left(x; (k, \lambda)\right) = \frac{1}{\Gamma(k)}\lambda^k x^{k-1} e^{-kx}$ is the density of a gamma distribution with (shape, rate) parameters (k, λ).

Finally, note that the rate at which recombination events occur on the branch connecting the two shared haplotypes is $4N_e\tau$. We have assumed that the first such event marks the end of the haplotype. However, there is a non-zero probability that a recombination event occurring on this branch does not change the TMRCA of a and b. Simulations suggest that for large numbers of chromosomes, this probability is extremely small and so we assume it is 0. In practice, for small samples, this might be a non-negligible effect.

Approximate likelihood for singleton count

Recall that the physical length of the shared haplotype is L_p bases. We assume that we estimated this exactly. Then assuming a constant mutation rate μ per base per generation, the sum of the number of singletons on the shared haplotypes, S^* has a Poisson distribution with parameter $\theta L_P \tau$, where $\theta = 4N_e\mu$.

Now consider the distribution of singletons on the unshared haplotypes. We make the following two assumptions: 1) There is no recombination on the unshared haplotypes over the region. 2) The distribution of the time to first coalescence of the unshared haplotype is exponential with parameter N (Recall that N is the number of sampled individuals). In fact the true distribution is a mixture of exponentials but the exponential distribution at least matches the correct mean and variance, $\frac{1}{N}$ and $\frac{1}{N^2}$ respectively (Blum and François, 2005).

Conditional on the time to first coalescence, τ_1 say, the number of mutations on each of the haplotypes is Poisson with parameter $\theta L_P \tau_1$ and so, using the assumptions above,

the unconditional distribution is geometric (on $0, 1 \ldots$) with parameter $\frac{1}{1+\frac{\theta L_p}{2N}}$ and the distribution of the number of mutations on both haplotypes, Δ_S, is the sum of two geometric distributions which is negative binomial with parameters $\left(2, \frac{\theta L_p}{\theta L_p + 2N}\right)$. Therefore the density of the total number of singletons, S is the convolution of these two densities

$$L(\tau; L_p, S) = \sum_{x=0}^{s} f_{Po}\left(x; \theta L_P \tau\right) f_{NB}\left(s - x; \left(2, \frac{\theta L_p}{\theta L_p + 2N}\right)\right) \tag{4.4}$$

where $f_{Po}\left(x, \lambda\right) = \frac{\lambda^x e^{-\lambda}}{x!}$ is the density of a Poisson distribution with parameter λ and $f_{NB}\left(x, (n, p)\right) = \binom{x+n-1}{x}(1-p)^n p^x$ is the density of a negative binomial distribution with parameters (n, p). In practice we estimate θ separately for each individual by counting singletons and then use the average of these values in Equation 4.4. A more accurate approach would be to compute the likelihood as a double convolution over the distribution of both haplotypes with different values for θ. An extension would be to estimate θ separately for different regions of the genome.

Approximate full likelihood

We can now write the approximate log-likelihood for τ as the sum of the logs of Equations 4.3 and 4.4,

$$\ell(\tau; L_g, S) = \log\left[L(\tau; l_g)\right] + \log\left[L(\tau; L_p, S)\right]. \tag{4.5}$$

In practice computing this requires a numerical integral for the first term, and a summation for the second. We maximise it numerically with respect to τ in order to find the maximum likelihood estimate (MLE).

Density estimation

As well as computing the MLE of the age of each haplotype, we also estimate the density of the distribution of ages, using the following procedure. First, we discretise the possible values of τ so that $\tau \in \{g_1 \ldots g_K\}$. We use a uniform grid in log-space. Then, suppose we have detected a total of M haplotypes, let the age and likelihood function of the m^{th} haplotype be τ_m and $L_m(\tau)$ respectively for $m = 1 \ldots M$. We specify a prior on τ, $p(\tau)$. We also specify a parameter $\alpha \in [0, 1]$ which defines how certain we are about the prior. Then starting from iteration $i = 0$, we sample values of τ_m, writing τ_m^i for the sampled value of τ_m at the i^{th} iteration.

- *For i=0:* Sample τ_m^0 such that $\mathbf{P}\left(\tau_m^0 = g_k\right) \propto L_m(g_k)p(g_k)$.
- *For i>0:* Write $M_k^i = \sum_{m=1}^{M} \mathbf{1}\left\{\tau_m^i = {}_k\right\}$. Then sample τ_m^i such that $\mathbf{P}\left(\tau_m^i = g_k\right) \propto \alpha L_m(g_k)p(g_k) + (1-\alpha)L_m(g_k)\frac{M_k^{i-1}}{M}$.

- *Estimation:* We run the algorithm for some initial number N_b of burn-in iterations, then for an additional N_T total iterations. We thin with parameter h so every $\frac{N_T}{h}$ iterations we record the M_k^i and then finally set our estimate q_k of $\mathbf{P}(\tau = g_k)$ to be the mean $q_k = \bar{M}_k^i$, where the average is with respect to i.

In all the applications described here, we set $\alpha = 0.05$ and used a prior such that $\log_{10}(t)$ had a normal distribution with mean 2 and variance 4 (recall that $t = 2N_e\tau$). We used $N_b = 100$, $N_T = 10,000$ and $h = 100$.

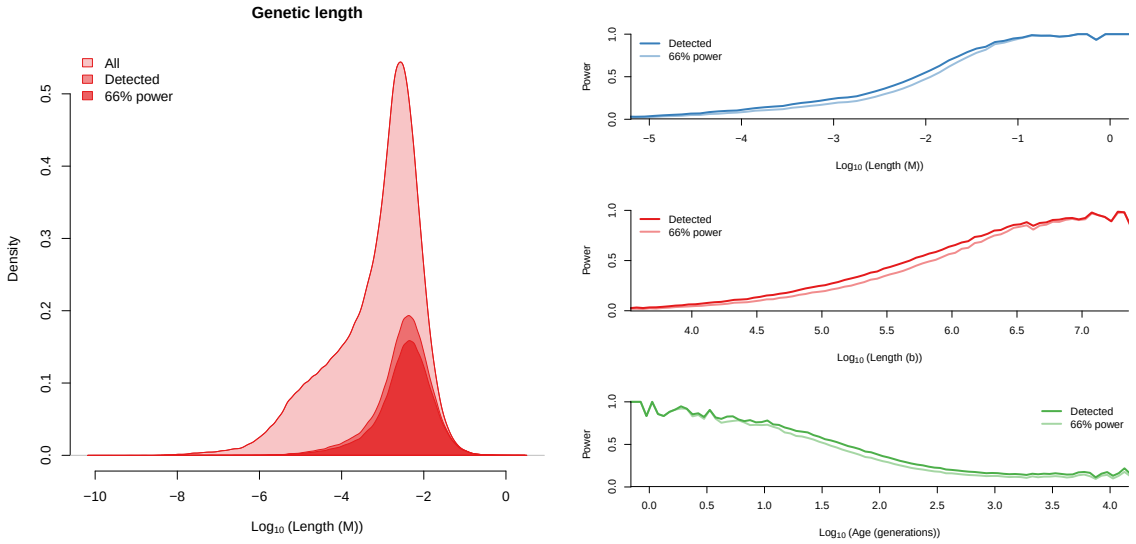
4.4.2 Results: Simulated data

In order to test our method, and calibrate the correction to L_g, we simulated whole-genome data for 100 individuals (i.e. 200 chromosomes) using *MaCS* (Chen *et al.*, 2009). This outputs the marginal coalescent trees along the chromosome so we can check how well we detect and estimate the age of the shared haplotypes (Figure 4.11). We had an overall power of 26% though this varies with length and age (Figure 4.11a-b). Unsurprisingly we have more power to detect very long haplotypes, but we detected many small haplotypes as well. 19% of our total had true genetic length, L_g^*, smaller than 0.1cM. Having imperfect power to detect doubletons does not have a large effect on our power to detect f_2 haplotypes since most haplotypes carry more than one f_2 variant. We have higher power for more recent haplotypes because they are typically longer, but this effect is cancelled to some extent for older haplotypes because the branches above them tend to be longer and therefore more likely to carry mutations. We can estimate the overall distribution of ages fairly well, though our density estimates are skewed to the left (Figure 4.11c). There is high uncertainly in the age of any individual haplotpye, but the approximate confidence intervals are well calibrated. Most of the information about the ages comes from the genetic length, and the main advantage of the singleton information is for very old haplotypes which otherwise have their age overestimated (Figure 4.11d).

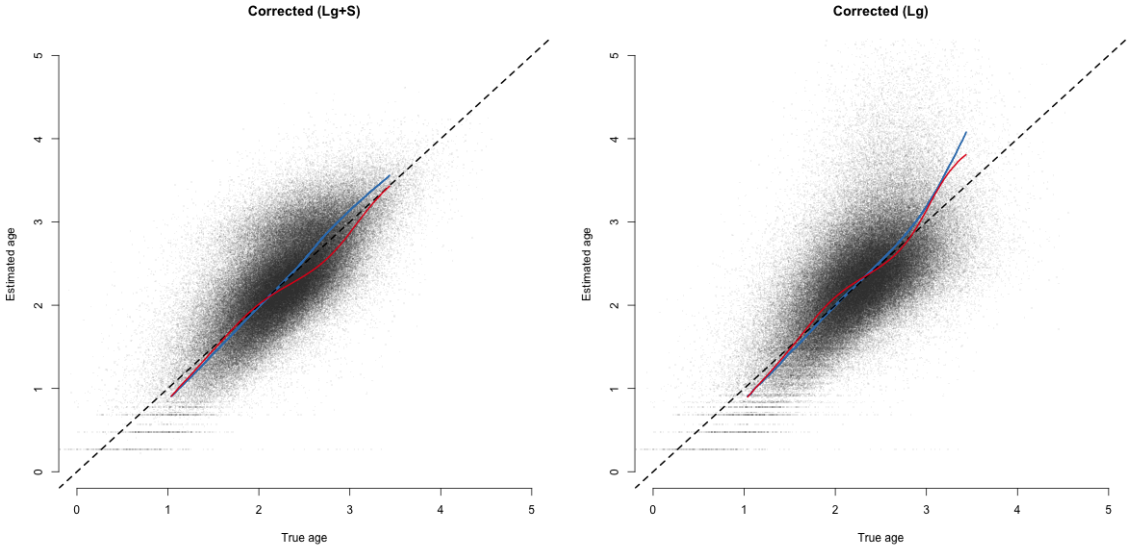
4.4.3 Results: 1000 Genomes data

After inspection of 4,066,530 f_2 variants, we detected 1,909,756 shared haplotypes totalling 2,049Gb in physical length and 20,240M in genetic length. This implies that we detect haplotypes covering an average total of 3.7Gb and 37M in each individual or approximately 130% and 106% of the physical and genetic length of the autosomal genome respectively[†]. The number of haplotypes is less than the number of f_2 variants because some haplotypes contain more than one variant. Table 4.2 shows the number of haplotypes shared between

[†]Each individual has two haplotypes, so f_2 haplotypes could cover up to 200% of the genome if they overlapped



(a) (b)



(c) (d)

Figure 4.11: Results from simulated data. Estimating f_2 age with simulated data. We simulated whole genomes for 100 individuals (200 chromosomes), with $N_e = 14,000$, $\mu = 1.2 \times 10^{-8}$ and HapMap 2 recombination rates. **a**: Distribution of the genetic lengths of f_2 haplotypes (lightest colour), and of the lengths of the detected haplotypes. The medium colour shows the haplotypes detected with 100% power to find f_w variants, and the darkest shows the haplotypes detected with 66% power to detect them. **b**: Power to detect f_2 haplotypes as a function of genetic length (top), physical length (middle) and haplotype age (bottom); in each case the darker line represents the power to detect f_2 haplotype with 100% power to detect f_2 variants, and the lighter line the power with 66% power to detect variants. **c**: Estimated age against true age ($\log_{10}$ scale). The grey dots are the MLEs for each detected haplotype. The blue line is a quantile-quantile (qq) plot for the MLEs (from the 1^{st} to 99^{th} percentile). The red line is the qq plot for the density estimate. **d**: As **c**, but without using the singleton information.

	MXL	CLM	PUR	TSI	IBS	CEU	GBR	FIN	YRI	LWK	ASW	CHS	CHB	JPT	
MXL	31,355	17,389	12,491	19,097	4,240	12,645	13,639	7,730	8,072	9,647	10,067	4,118	4,386	3,995	MXL
CLM		22,748	12,620	20,010	5,038	13,587	15,225	8,076	11,276	11,929	11,231	2,605	2,900	2,590	CLM
PUR			16,325	19,476	3,831	12,841	13,759	7,123	12,315	12,358	11,553	1,897	2,073	1,816	PUR
TSI				43,627	6,972	36,944	36,015	19,315	4,394	8,929	8,781	3,979	4,689	3,698	TSI
IBS					1,522	5,131	6,396	3,105	800	1,403	1,603	582	668	585	IBS
CEU						21,763	42,228	24,449	2,587	4,337	7,856	2,521	2,898	2,455	CEU
GBR							27,718	24,678	2,849	4,561	8,460	2,821	3,014	2,758	GBR
FIN								51,838	1,944	3,095	4,976	2,571	3,424	2,985	FIN
YRI									101,993	89,069	96,739	3,520	3,426	3,261	YRI
LWK										137,226	69,188	5,055	4,959	4,690	LWK
ASW											33,403	3,739	3,714	3,251	ASW
CHS												70,201	114,518	53,016	CHS
CHB													60,918	65,618	CHB
JPT														110,255	JPT
AMR	112,928														
EUR	176,317	351,701													
AFR	98,448	66,575	527,618												
ASN	26,380	39,648	35,615	474,526											
	AMR	EUR	AFR	ASN											

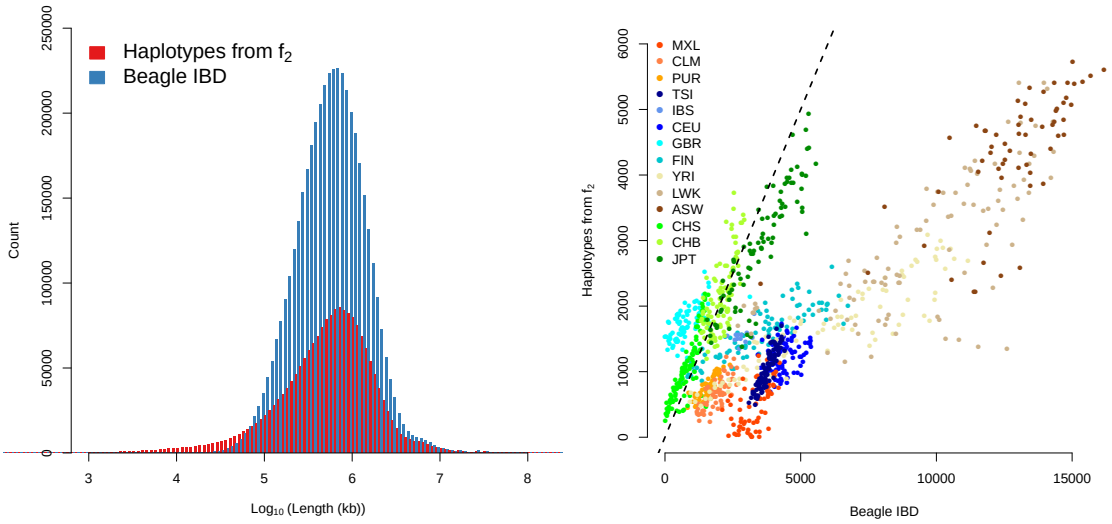
Table 4.2: Number of haplotypes shared between populations. **Upper right**: Haplotype counts by pair of populations. Populations described in Table 4.1 **Lower left**: Haplotype counts by continental grouping, Abbreviations: AMR, American; EUR, European; AFR, African; ASN, Asian.

each pair of populations, and aggregated over continental groups. These haplotypes contained a total of 10,020,877 singletons[†].

For comparison, the *fastIBD* method of *Beagle* (run on the array genotype data) detected 4,421,936 IBD chunks in the same dataset. *Beagle* detected more chunks than we found shared haplotypes at almost all lengths, with the exception of very short haplotypes (<65kb), though the distribution of lengths was similar. In both cases the modal length of chunk was around 1Mb (Figure 4.12a). The reduction in the number of f_2-detected haplotypes relative to IBD chunks was not uniform across populations (Figure 4.12b). In particular, there was virtually no difference in the number of chunks by the two methods for the Asian populations and, rather strangely, for GBR the f_2 method detected more chunks than the IBD method. The f_2 method should detect just over 20% of the IBD chunks, because only this proportion have a consistent f_2 variant (Figure 4.8c). This implies that around 53% of the f_2-based haplotypes are not concordant with any IBD chunk.

To estimate the age distributions of the haplotypes, we used the combined recombination rate map from HapMap 2 to determine genetic lengths, and assumed $N_e = 18,000$, and a mutation rate of 0.4×10^{-8} per-base per-generation, reflecting a true mutation rate of 1.2×10^{-8} Scally and Durbin (2012) and a power of $\frac{1}{3}$ to detect singletons (1000 Genomes Project Consortium, 2012). . We computed density estimates for the ages of all the f_2 haplotypes

[†]There are only 7,970,795 singletons in the dataset, but we are counting some of them twice in overlapping haplotypes.



(a) (b)

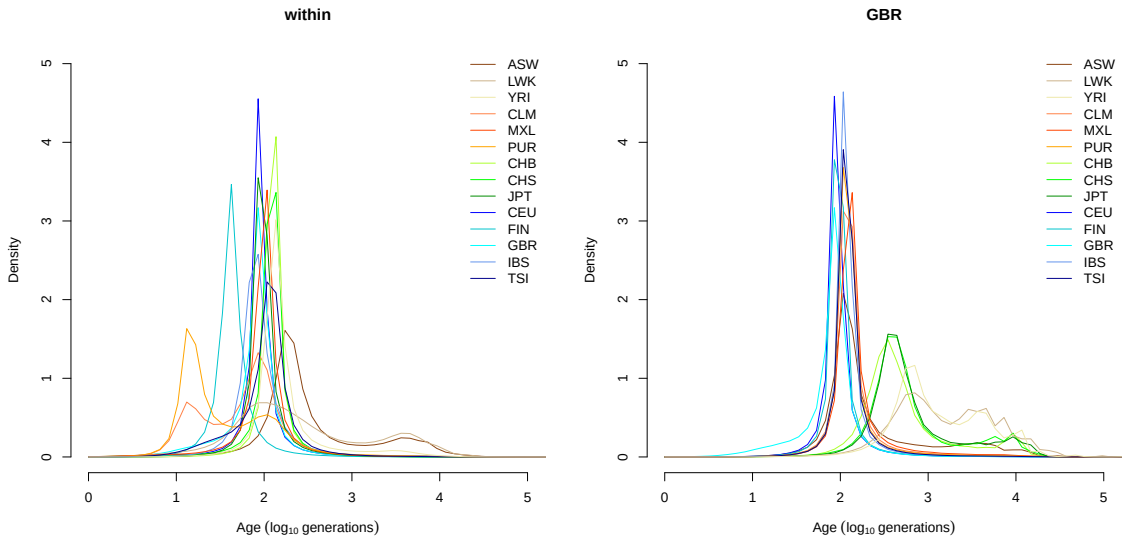
Figure 4.12: Comparing the number of shared haplotypes found using the f_2 sharing method described here, and the *fastIBD* IBD detection method of *Beagle*. **a**: Histogram of the number of haplotypes/IBD chunks, as a function of length. **b**: The number of haplotypes/IBD chunks detected per individual. Compare with Figure 4.8d.

---

shared between every pair of populations (Figure 4.13, Appendix C). These estimates largely reflect the expected demographic histories of these populations. For haplotypes shared within populations (Figure 4.13a), most of the European and Asian populations are tightly clustered around 100 generations ago. For example the median age of GBR-GBR haplotypes is 84 generations and the upper 95% quantile is 198 generations. PUR, CLM and FIN show some much more recent haplotypes (on the order of 10 generations ago), presumably representing expansion following recent population bottlenecks. The African populations have many recent haplotypes but also a long tail. For example the median age of LWK-LWK haplotypes is 184 generations, but the 95% quantile is 6,714 generations.

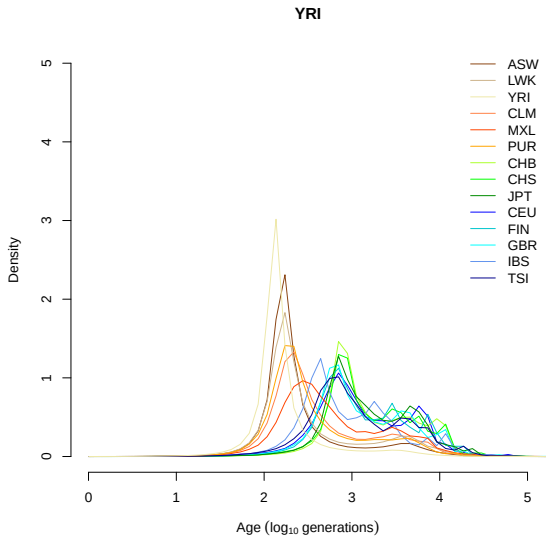
Similarly, between-population sharing is largely consistent with the relative histories of these populations (Figure 4.13b-d). European-Asian sharing peaks between 400 and 1,000 generations ago and [European/Asian]-African sharing rather earlier. GBR-LWK haplotypes have a median age of 1,355 generations, for example. Interestingly both Asian-European and Asian-African haplotypes appear to have a bimodal distribution with a local minimum around 2,000-4,000 generations ago, which it is tempting to see as a sign of a population split some 30-60 thousand years ago (Fenner, 2005, estimates the mean human generation time to be 30 years).

Admixed populations have age distributions which look largely like linear combinations

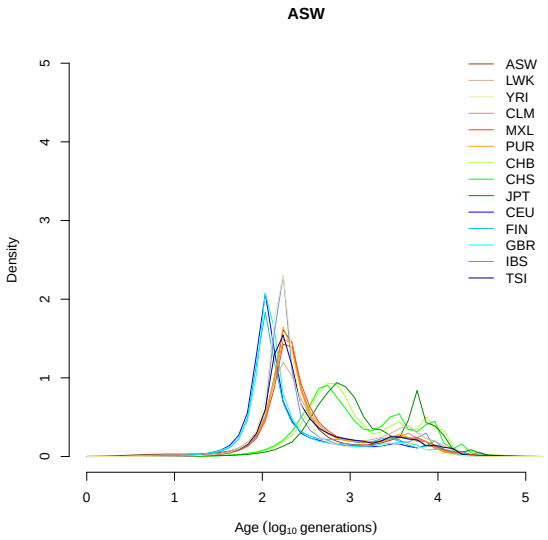


(a) All populations

(b) GBR



(c) YRI



(d) ASW

Figure 4.13: Density of age distributions for **a**: Haplotypes shared within populations **b**: Haplotypes shared by Great British (GBR) samples **c**: Yoruban samples (YRI), **d** African American (ASW). In **b-d** there are separate densities for haplotypes shared with each other population, so the dark green line in **b** represents the distribution of GBR-JPT haplotypes, for example. Equivalent plots for all other populations are shown in Appendix C

of the admixing populations (Figure 4.13d, for example). Even in these populations we can see signs of more subtle history. For example, GBR-American haplotypes have a similar age distribution to GBR-IBS/TSI haplotypes, presumably representing the fact that the major contribution to European admixture in these populations is from southern Europe.

4.4.4 Discussion

We have described a method for extracting information about the age of shared haplotypes from a combination of sequence and array genotype data. The age distributions we obtain for different populations are plausible, although we need to take some care in interpreting these distributions. We've estimated the age distribution of f_2 variants, and it's not unreasonable to think of these distributions as the distribution of the TMRCA of two sequences, conditional on the fact that they are unique genealogical nearest neighbours (in the sense that the the two sequences coalesce with each other before either of them coalesce with difference sequences). We would like to extend this to a statement about the distribution of coalescence times across the sequence, although to do this we would need to take the varying power to detect differently sized haplotypes into account.

There are several other methods which perform this kind of inference, for example PSMC (Li and Durbin, 2011), or the approach of Ralph and Coop (2013) which uses the genetic length of IBD chunks to estimate the date of recent ancestry. These use slightly different information and answer questions over different timescales. One way to classify them is in terms of the number of coalescent events they use, and the timescales over which they observe these events. For example, when we applied our approach to the 1000 Genomes data, we observed approximately 1,600 coalescences per genome, with a median age of 110 generations. In contrast, we would expect PSMC to see on the order of 100,000 coalescences per genome, but they would me much older, with a median age of $2N_e \log 2 \approx 25,000$ generations. When we ran *Beagle* on the same data, we detected 4.4 million IBD chunks, equivalent to 4,050 coalescences per genome. The length distribution of these IBD chunks is similar to our f_2 chunks, suggesting that they have roughly the same age distribution, although the f_2 method finds smaller chunks. IBD-based approaches might use more information therefore, but observe a narrower time range. Further, the f_2 approach has the advantage that, since every f_2 variant must lie in an f_2 haplotpye, the only source of false positives are genotyping errors, in contrast to IBD approaches which can have high false positive rates particularly for short (and therefore old) chunks, so we would expect that IBD approaches become more inaccurate for older haplotypes, while the f_2 approach does not. Finally, the f_2 method is significantly faster to run. Detecting haplotypes is mostly a process of scanning through array data, and does not require large amounts of

data to be loaded at once, making it time, and memory efficient. We found and dated all the haplotypes in the 1000 Genomes data on a desktop computer in a few hours. Running *Beagle* to find the IBD chunks took tens of machine-days, running on a cluster. This would be a big advantage when scaling to datasets of tens or hundreds of thousands of samples.

One area for improvement is to incorporate variable mutation rates between populations and across the genome. It is simple to estimate θ by counting the number of singletons, and so we could simply estimate θ as a function of population and position. There are several improvements still to make to do with the models for length and number of singletons discussed in Section 4.4.1. In particular we do not model the overestimate in the physical length L_p and the distribution of the number of singletons S^* is overdispersed compared to our model because we have assumed no recombination on the unshared chromosomes.

Finally, another source of extra data would be to look at variants at higher frequencies. This should allow us to detect even more shared haplotypes, although they are no longer obviously interpretable in terms of genealogical nearest neighbours.

4.5 Estimating migration rates from f_2 sharing

In the previous section we used rates of f_2 sharing to estimate the distribution of coalescence times, a very generic property of the population, and related to no specific parametric model. We now describe a similar approach using the same information, but this time to estimate parameters, specifically the migration rates between populations. Again, we will use the f_2 variants to detect nearest genealogical neighbours, and then use the fact that if these are in different populations, they must result from migration events (or less likely, ancient shared ancestry), and therefore the relative numbers of shared f_2 variants within and between populations are informative about the rate of migration. Once again, we demonstrate this approach on the 1000 Genomes dataset.

4.5.1 Method

The ideal approach would be to compute the likelihood of the observations under the structured coalescent and use the likelihood with respect to the migration rate to make inferences. This likelihood is intractable however, and we take the approach of considering simplified approximations of the structured coalescent, computing the expected rates of rare allele sharing analytically under these models as a function of migration rates, and then matching moments by choosing the rates to match the observed sharing. We consider three different simplified models.

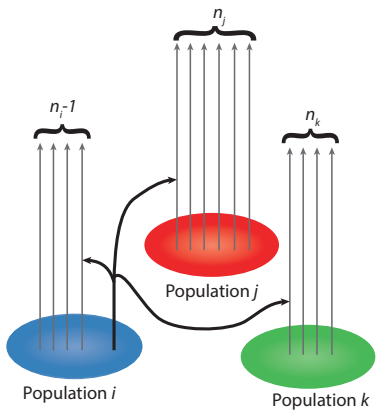


Figure 4.14: Model 1. A lineage in deme i either coalesces in deme i, or migrates to another deme. ≤ 1 migration event, and trunk genealogy.

Model 1 is based on the conditional sampling distribution using a trunk genealogy (a genealogy with no coalescence, Figure 4.14). Suppose we have K demes, and we have sampled $n_i - 1$ lineages from deme i, and n_j lineages from all the other demes $j \neq i$. Consider a new sequence sampled from deme i. The only way it can carry an f_2 variant is if its lineage is the first to coalesce with some other lineage. This other lineage can either be from deme i, or from another deme, in which case the f_2 is shared within deme i, or between deme i and another deme.

In order to simplify this model, we make the following five assumptions:

1. Only the newly sampled lineage can coalesce, the other lineages do not coalesce.
2. If the newly sampled lineage migrates to another deme, it *must* coalesce in that deme, so it can only migrate once.
3. Only the newly sampled lineage can migrate.
4. The effective population size, and thus the time scaling factor, is the same in all subpopulations.
5. The probability of seeing an f_2 mutation given that two lineages coalesce with each other before any other lineages is independent of the demes from which those lineages came.

These assumptions allow us to express the expected f_2 sharing probabilities for each pair of demes, $\bar{P}^1_{ij}$, explicitly in terms of the number of lineages sampled from each deme, n_i for deme i, and the migration rate from deme i to deme j, m_{ij}. Specifically,

$$\bar{P}^1_{ij} = \begin{cases} \frac{n_i - 1}{n_i - 1 + m_{i\bullet}} & \text{if } i = j \\ \frac{m_{ij}}{n_i - 1 + m_{i\bullet}} & \text{if } i \neq j. \end{cases} \tag{4.6}$$

Where the $\bullet$ symbol denotes summation over indices. Now consider our assumptions, in reverse order. The last assumption is reasonable if mutation rates are high, or at least constant across demes. We probably have some idea of whether the second to last assumption is reasonable from other sources. We'll discuss later the effect of deviating from this. For assumption 3, migration of the other lineages acts to change the number of lineages

in the other demes back in time, and also means we assign sharing to the wrong demes. However, if all the m_{ij} and n_i are equal these effects will cancel out, since an equal number of lineages will move between every pair of demes. It follows therefore, that the magnitude of this effect depends on the difference between the true model and the uniform isotropic model, so we assume that it is small, provided that the migration is not strongly directional and the sample sizes not too different.

The first two assumptions are more problematic and our two other models are each derived by relaxing one of these assumptions. Suppose we relax assumption 2 and allow the sampled lineage to migrate more than once. We call this model 2 (Figure 4.15). Now the movement of the sampled lineage follows a continuous time Markov chain on $2K$ states, where K is the number of demes. States $1 \dots K$ are the states representing the lineage being in the corresponding deme, without having coalesced, and states $(K+1) \dots 2K$ are the absorbing states where the lineage has coalesced in deme $1 \dots K$. Then the rate matrix for this chain is given by

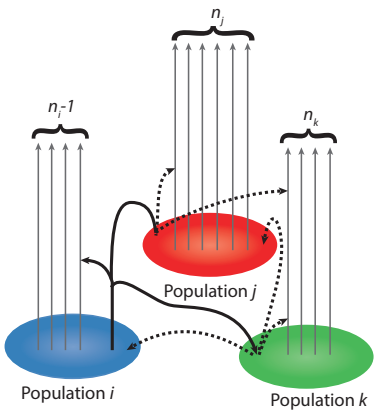


Figure 4.15: Model 2. Still a trunk genealogy, but the sampled lineage can migrate multiple times.

$$Q = \left(\begin{array}{c:c} Q_{11} & Q_{12} \\ \hdashline 0 & 0 \end{array}\right) \tag{4.7}$$

where

$$Q_{11} = \begin{pmatrix} -n_1 - m_{1\bullet} & m_{12} & \cdots & m_{1K} \\ m_{21} & -n_1 - m_{2\bullet} & \cdots & m_{2K} \\ & \ddots & \vdots & \\ m_{k1} & & & -n_k - m_{k\bullet} \end{pmatrix} \text{ and } Q_{12} = \begin{pmatrix} n_1 & \cdots & 0 \\ \vdots & n_2 & \\ 0 & & \ddots \quad \vdots \\ & & \cdots \quad n_k \end{pmatrix}$$

and therefore the transition matrix of the embedded discrete Markov chain is given by the rate matrix, normalised so it is stochastic (the rows sum to 1) and with zeros on the diagonal.

$$T = \left(\begin{array}{c:c} T_{11} & T_{12} \\ \hdashline 0 & I \end{array}\right) \tag{4.8}$$

where

$$T_{11} = \begin{pmatrix} 0 & \frac{m_{12}}{n_1+m_{1\bullet}} & \cdots & \frac{m_{1k}}{n_1+m_{1\bullet}} \\ \frac{m_{21}}{n_2+m_{2\bullet}} & 0 & \cdots & \frac{m_{2k}}{n_2+m_{2\bullet}} \\ \vdots & & \ddots & \vdots \\ \frac{m_{k1}}{n_k+m_{k\bullet}} & & & 0 \end{pmatrix} \text{ and } T_{12} = \begin{pmatrix} \frac{n_1}{n_1+m_{1\bullet}} & \cdots & & 0 \\ \vdots & \frac{n_2}{n_2+m_{2\bullet}} & \vdots & \\ & & \ddots & \vdots \\ 0 & & \cdots & \frac{n_k}{n_k+m_{k\bullet}} \end{pmatrix}.$$

The probability that the sampled lineage is absorbed in state j, conditional on starting in state i, is given by the i, j^{th} entry of the matrix $(I - T_{11})^{-1} T_{12}$ (Norris, 1998), therefore in this model our expected sharing probabilities are given by

$$\bar{P}^2_{ij} = \left[(I - T_{11})^{-1} T_{12}\right]_{ij}. \tag{4.9}$$

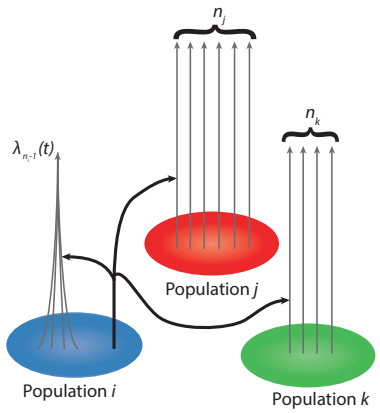


Figure 4.16: Model 3. The number of lineages in deme i decreases with time.

Finally, consider a model where we reinstate the second assumption, and relax the first. We assumed that the number of lineages in all demes remained constant, in particular the number of lineages in deme i. In fact, the number of lineages in deme i decreases back in time as other lineages coalesce, so assuming that it stays constant means that we overestimate the probability of coalescence relative to migration, which leads us to suspect that models 1 and 2 will overestimate the migration rate. Instead, set the rate of coalescence in deme i to be proportional to the expected number of lineages at time t (Figure 4.16). The expected number of lineages $\lambda_n(t)$ at time t, given n lineages at time 0 is given by Equation 5.11 of Tavaré (1984),

$$\lambda_n(t) = \sum_{i=1}^{n} \frac{n\,(n-1)\ldots(n-i+1)}{n\,(n+1)\ldots(n+i-1)}\,(2i-1)\,e^{-\frac{ti(i-1)}{2}}. \tag{4.10}$$

We also define the integrated expected number of lineages $\Lambda_n(t)$

$$\begin{aligned}\Lambda_n(t) &= \int_0^t \lambda_n(\tau)d\tau \\ &= t + \sum_{i=2}^{n} \frac{n\,(n-1)\ldots(n-i+1)}{n\,(n+1)\ldots(n+i-1)} \frac{2\,(2i-1)}{i\,(i-1)} \left(1 - e^{-\frac{ti(i-1)}{2}}\right).\end{aligned} \tag{4.11}$$

Coalescence of the sampled lineage in deme i therefore occurs with time-dependent rate $\lambda_{n_i-1}(t)$ while migration out of the deme occurs with constant rate $m_{i\bullet}$. If the lineage migrates at time t, the probability that the lineages have coalesced first is $\left(1 - e^{-\int_0^t \lambda_{n_i-1}(\tau)d\tau}\right)$. Therefore, conditioning on the first migration, the probability that the sampled lineage coalesces in deme i rather than migrating is

$$\begin{aligned}c_i &= \int_0^\infty \left(1 - e^{-\int_0^t \lambda_{n_i-1}(\tau)d\tau}\right) m_{i\bullet} e^{-m_{i\bullet}t} dt \\ &= 1 - m_{i\bullet} \int_0^\infty e^{-\int_0^t \left(\lambda_{n_i-1}(\tau) + m_{i\bullet}\right)d\tau} dt \\ &= 1 - m_{i\bullet} \int_0^\infty e^{\left(\Lambda_{n_i-1}(t) + m_{i\bullet}t\right)} dt\end{aligned} \tag{4.12}$$

and it follows that the expected allele sharing probabilities in this model are given by

$$\bar{P}^3_{ij} = \begin{cases} c_i & \text{if } i = j \\ (1 - c_i)\,\frac{m_{ij}}{m_{i\bullet}} & \text{if } i \neq j. \end{cases} \tag{4.13}$$

Each of $\bar{P}^1_{ij}, \bar{P}^2_{ij}$ and $\bar{P}^3_{ij}$ can be easily computed as a function of the n_i, which are known, and the m_{ij} (Equations 4.6, 4.9 and 4.13). Therefore, we can find values of m_{ij} which best fit the observations by minimising the function $\sum_{i,j}\left(\bar{P}^m_{ij} - P_{ij}\right)^2$ with respect to m_{ij}[†], where P_{ij} are the observed sharing probabilities and $m = 1, 2$ or 3 is the model we are using.

4.5.2 Results

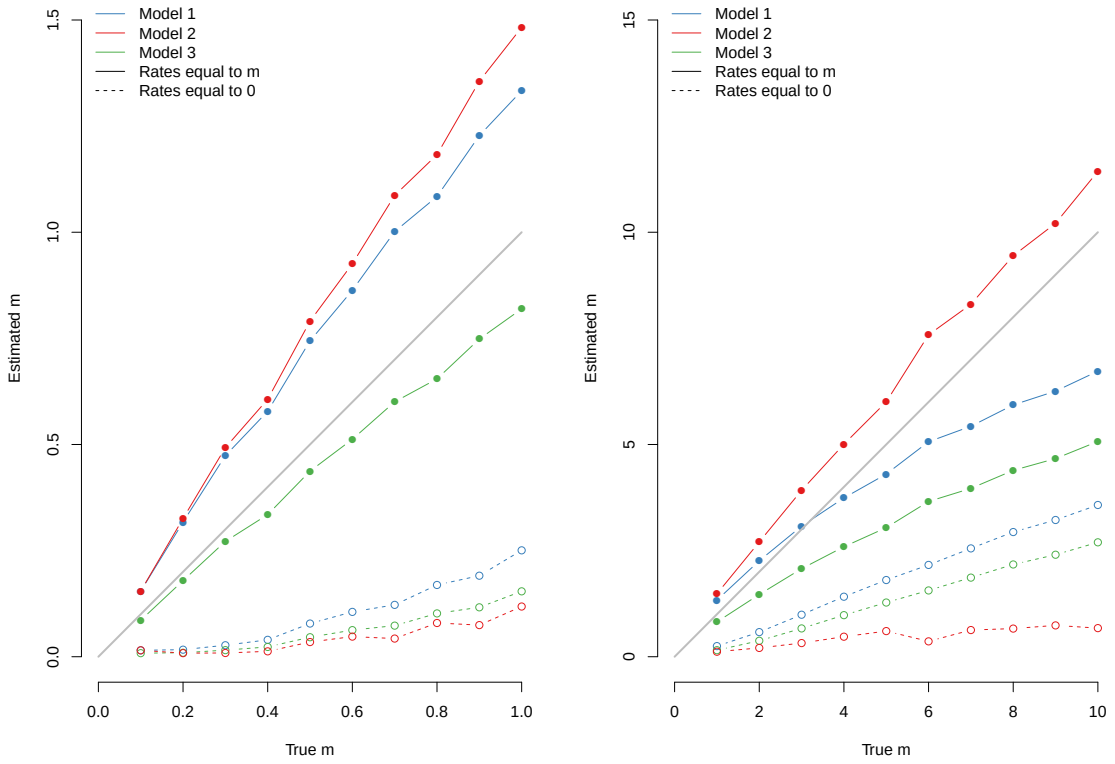
To test the different models, we simulated data using our code from Chapter 3. We simulated in a model of 4 demes arranged in a 2×2 grid, with migration at rate m between adjacent demes so that some of the m_{ij} were equal to m and some were equal to 0. That is, the migration rate matrix is given by

$$[m_{ij}] = \begin{pmatrix} - & m & m & 0 \\ m & - & 0 & m \\ m & 0 & - & m \\ 0 & m & m & - \end{pmatrix}. \tag{4.14}$$

We let m range from 0 to 10, and sampled 50 lineages from each deme. The results are shown in Figure 4.17. Interestingly, the best performing model is different for different values of the parameters. For small values of m, model 3 performs best. We suggest this is because, when m is small (Figure 4.17a) the rate of coalescence in the original deme, which model 3 treats correctly, is the most important parameter for the sharing probability. Models 1 and 2 overestimate the migration rate because they overestimate the probability of coalescence in the original deme, so fit a higher migration rate to compensate. Interestingly, model 1 performs better than model 2, at least for demes where the rate is m, although it overestimates the zero entries more.

In contrast, when m is higher (Figure 4.17b), model 2 performs better than the others. Presumably for high m, multiple migration events are more common, and model 2 takes this into account. Models 1 and 3 overestimate the zero entries by a large amount here. These results are reasonable, given the approximations. Model 2 has a more accurate treatment of migration and so we might expect it to perform better when migration is high, for example. However, none of these models are really ideal, and we should probably be cautious about extending these results to different migration matrices.

[†]We use the *L-BFGS-B* algorithm implemented in the base R function *optim*. This is a gradient-based optimisation routine, allowing for linear constraints, so we can constrain the m_{ij} to be non-negative.



(a) $0 < m \leq 1$ (b) $1 \leq m \leq 10$

Figure 4.17: Estimating migration rates for simulated data. We show the migration rates estimated using the three models described in the text for two parameter regimes **a**: $0 < m \leq 1$ and **b**: $1 < m \leq 10$. For each value of m, we simulated 50 sequences from each of four demes at 1,000 sites at each of 10,000 independent genealogies, then computed the f_2 sharing and found the optimal value of the m_{ij} under each model as described in the main text. Since the demes are arranged in a 2×2 square, the migration matrix has eight entries equal to m and two equal to 0 (Equation 4.14), and we show separately the estimates averaged over these two sets of entries (solid and dashed lines).

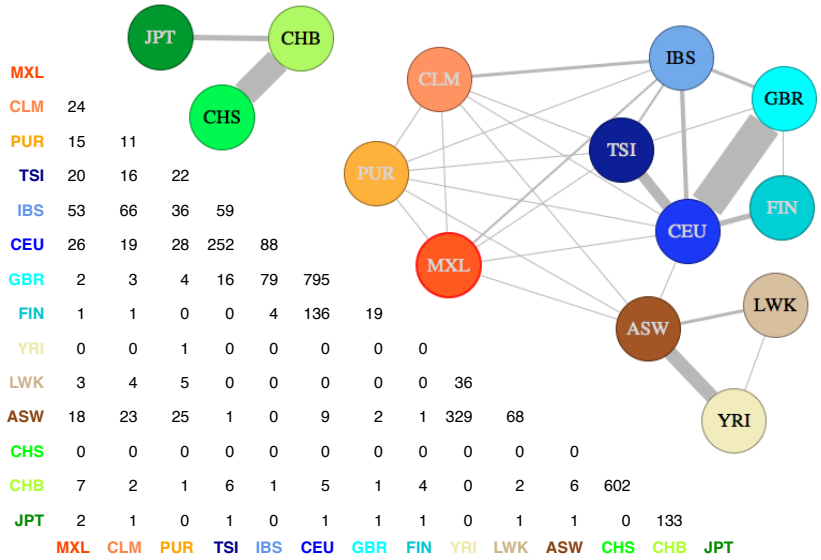


Figure 4.18: Estimated migration rates between 1000 Genomes populations. **Lower left**: Estimated symmetric migration rates (averaged over lower and upper diagonal parts, rounded to nearest integer). **Upper right**: Major migration routes. Network showing all migration routes with $m_{ij} > 10$.

Nonetheless, as an illustration we estimated migration rates for the 1000 Genomes data (Figure 4.18). We used model 2 since migration rates are high. We also constructed a symmetric migration matrix by averaging over the upper and lower diagonal matrices. We discuss why the matrix is asymmetric later. The first thing we notice is that some of the estimated rates are extremely high. $m_{CEU-GBR} = 795$, for example. This is not surprising. In fact, the probability that a CEU f_2 variant is shared with CEU is 22%, only slightly higher than the 20% probability that it is shared with GBR. On a large scale, the results are consistent with what we expect from the relationships between the populations (network shown in Figure 4.18). Note that since these migration rates are derived from f_2 sharing, they reflect rates over the timescales of hundreds or thousands of generations, in fact exactly over the timescales estimated in the previous section, and shown in Figure 4.13.

4.5.3 Discussion

There are several alternative methods for estimating migration rates which we did not consider here. MCMC approaches like that of Beerli and Felsenstein (2001) are likely to be too computationally intensive for large samples. On the other hand, approximate methods which collapse individuals over demes might be tractable. A similar approach

Population	YRI	MXL	CLM	PUR	TSI	IBS	CEU	GBR	FIN	LWK	ASW	CHS	CHB	JPT
Size	1.00	1.02	1.18	1.31	1.04	1.07	0.80	0.88	0.78	1.39	1.18	0.99	0.80	1.10

Table 4.3: Population sizes estimated from the 1000 Genomes migration matrix. All sizes given relative to the YRI size.

to the one we described here could be implemented using the method of Nagylaki (1998), which gives explicit formulas for mean pairwise coalescent times under arbitrary migration models. However, we are not aware of an equivalent formula for the expected coalescence time conditional on the first coalescence (equivalently, an f_2 variant), so for the purpose of estimating recent coalescence rates, an extension of the method described here might be more appropriate.

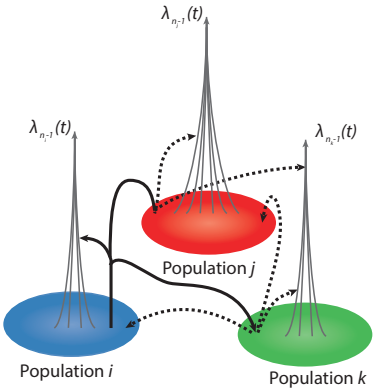


Figure 4.19: Model 4. Incorporating the multiple migrations of model 2 and the variable coalescence rates of model 3.

In this section we attempted to make parametric inference about migration rates, based on a series of models designed as approximations to the structured coalescent. Our simulations make it clear that none of the simplified models we considered are wholly satisfactory. Model 2 works well when migration is high, and model 3 when migration is low. The obvious solution would be a model which incorporates both multiple migrations and variable coalescence rates (Figure 4.19). The problem with this model is that it is not as tractable as the other two. Because the transition rates are not homogeneous, there is no simple analogue to Equation 4.9. The absorption probabilities can be represented as an infinite series of nested integrals, in a similar spirit to Equation 4.12, but conditioning on the time of every possible move. However, even if we truncate this series and limit the number of jumps, computing $\bar{P}_{ij}$ this way would probably be numerically unstable and difficult to optimise numerically. However, since we can easily simulate from this model, an importance sampling approach might be practical. We could simulate from the full model, record where each path is absorbed and accept the paths in proportion to the observed sharing probabilities. Then we can use the sampled paths as an empirical sample from the distribution of all paths, and estimate the migration rates by counting the observed transitions.

Another question that we have avoided so far is what it means when the estimated migration matrix is asymmetric. There are actually two effects here. The first is that asymmetric migration itself, and the second is variation in effective population size across demes, which mean that the m_{ij} represent rates on different timescales. An *ad hoc* method

for extracting the population sizes is as follows. Suppose $M = [m_{ij}]$ is the estimated migration matrix and $N = (1, N_2 \dots N_K)$ is a vector of effective population sizes where we set N_1 to 1. Then we define $\tilde{M}$ by $\left[\tilde{M}\right]_{ij} = m_{ij}/N_j$ and choose N to make $\tilde{M}$ as close to diagonal as possible by minimising $||\tilde{M} - \tilde{M}^t||^2$. This works well on small simulated datasets, and the estimated population sizes for the 1000 Genomes data are shown in Table 4.3. Like the migration rates, these are effective population sizes on the timescale of hundreds of generations. We would like to interpret the remaining asymmetry in $\tilde{M}$ as evidence of directional migration, though it's likely there are other things we've ignored which affect this. In particular, while some of the directions look correct (for example we see migration from YRI to ASW, and from ASW to PUR, MXL and CLM), some of them do not. We see migration from all populations into IBS, which seems unlikely. IBS has a much smaller sample size than the other populations and it's likely that this causes us to mis-estimate its parameters.

4.6 Conclusion: Rare variant sharing

In this chapter we investigated sharing of rare variants, and particularly doubletons, in the 1000 Genomes data. Doubletons, or f_2 variants, are highly informative because they indicate the first coalescent events of two lineages. We developed two methods for inference using these information and applied them to the 1000 Genomes data. In the first case we estimated coalescence times, using the data to estimate a distribution. In the second case we estimated migration rates in a parametric model. Both of these methods seem to work reasonably well and, importantly, are fast and efficient to apply. For example, finding the shared haplotypes and making inference about times in Section 4.4 was substantially faster than running *Beagle* to find IBD chunks, and the two methods provide roughly equivalent information.

Much of the gain in speed is because looking only at doubletons greatly simplifies calculations. However by doing this we are ignoring large amounts of data. A natural extension to this work would be to make these kind of methods work with rare variants at higher frequency in order to use more of the data, but trying to keep the simplifications which make these methods faster. For example, it would be easy to consider f_3 variants in our migration model, by looking at patterns of sharing, conditional on sharing an f_2.

Another interesting extension would be to combine the two inference problems we investigated here and jointly estimate migration rates and coalescent times. A simple way of doing this would be to stratify variants by their age and then estimate migration rates

separately for each age category to see how migration rates have varied over time. For example, we'd expect to see higher migration rates between Asian and American populations for variants older than around 14,000 years (approximately 450 generations), compared to more recent variants.

At the moment, these methods are useful, although there are plenty of practical alternatives. They may come into their own when large, fully sequenced, datasets are available. Though the 1000 Genomes is the largest publicly available dataset at the time of writing, several large-scale projects are being planned and in the next decade we expect the amount of data available to increase dramatically. This will allow us to make inference about human populations at a much finer scale or, equivalently, over more recent time periods. Methods like the ones described in this chapter which scale well both in terms of computational time and memory usage may then become invaluable.

Chapter 5

Conclusions

In this thesis, we approached the problems of spatial structure in several different ways, with varying degrees of success. To finish, we discuss the conclusions that can be drawn from our results and how our approaches fit into the existing literature. Predicting the future is both difficult and foolish, but nonetheless we attempt to predict which approaches will be useful for the next generation of genomic data, and which questions we might then be able to answer.

5.1 Models: beyond the stepping stone

Generally, when specifying models, we have worked with the stepping-stone model, mainly because it is easy to implement, well characterised and extremely useful for understanding the qualitative behaviour of structured populations. For example in Chapter 3, we used this model to illustrate the relatively subtle variation in the effect of population structure at different allele frequencies. In Chapter 2, we extended the model to include spatially variable selection. In this case we were able to make explicit parametric inference about the selection coefficient. However, our main observation was that the model fitted poorly to real data. One conclusion we might draw from this is that even the simplest natural population is likely to have dynamics significantly more complicated than those contained in the stepping-stone models we considered. A natural response to this is to build ever-more complicated models, to the point at which inference and interpretation become difficult. Perhaps a more promising approach would be something like the one we took when using data from both ancient and modern samples in Section 2.6, when we combined a relatively constrained parametric model with a more complex but generic model. This meant that we were able to combine different sources of information, and provided a simple check on the robustness of the conclusions, since we could compare the results of the two methods.

Nonetheless, modern genomic datasets do provide the ability to fit complex parametric models. Models involving splits, population size changes and admixture events can be fitted to human datasets (Gutenkunst *et al.*, 2009; Gravel *et al.*, 2011; Harris and Nielsen, 2013). However, these models typically require us to define the structure of the model in a way which effectively imposes a strong and inexplicit prior on the histories. In Chapter 4, we developed inference methods which make very few assumptions about the structure of the underlying model. These approaches are complementary in the sense that there is a trade off between prior assumptions and ease of interpretability. This suggests a two-step paradigm, where we first infer the large scale structure of the model using fast and unconstrained approaches, and then infer parameters within these models using more highly specified models. In this approach, we would first infer which populations appeared to have migration, and roughly over what timescales, as well as qualitative splits and size changes. We then can then fit a parametric model of migration rates, split times, and size changes, but much more easily because we have significantly constrained the model. For example, using f_2 sharing, we can quickly infer the network in Figure 4.18. Then, using all the data we can fit a slower but more accurate model, but with a strong prior on the rates, and with many structural zeros. Alternatively the unconstrained models can be used to provide sanity checks on inference from other approaches.

Of course, unconstrained models can allow parametric inferences. Indeed we showed in Chapter 4 that in some cases we could make relatively accurate inference using approximate models and data summaries. However, these can be difficult to interpret. It is not always easy or possible to move from inference about, say, coalescent times to inference about actual quantities of interest like migration rates or population size. In this sense our approach is rather like the *PSMC* of Li and Durbin (2011) which infers the distribution of coalescence times across the genome and uses these to make inference about historical population size. In general there is no direct relationship between the N_e estimated by *PSMC* and the population size, although clearly the two are correlated.

Most analysis of human structure makes, explicitly or implicitly, an assumption of discrete populations, despite debate about the appropriateness of this assumption (Serre and Pääbo, 2004; Rosenberg *et al.*, 2005, for example). It is probably not unreasonable to assume that major geographic features can lead to discontinuities in patterns of genetic variation, say on a continental scale. Then the question becomes to what extent there is continuous variation on a smaller scale, as opposed to a series of fairly homogenous groups. Probably the current limitation on investigating this question is the lack of samples collected in a spatially continuous manner. Still, models which explicitly incorporate continuous structure

and can be fitted to genome-scale data have the potential to characterise spatial structure on much smaller scales than current approaches.

Another important feature which has been missing from our models is the ability to include anything but the simplest time-varying structure and parameters. To model real populations we really want to be able to model processes like migration and selection which are both anisotropic and time-varying. We did model spatially-variable selection in Chapter 2, but time-varying selection is also likely to be part of the story as well as time- and directionally-varying migration. These add an extra dimension of complexity to the models, but the approach of Chapter 4 shows how we might begin to approach this. By looking at variants of different frequency we can investigate structure on different timescales. Again this would allow us to infer the qualitative structure of the model in a relatively assumption-free way, and we could then fit parameters of this, more constrained model.

None of these approaches can, in isolation, give us the whole picture and there is certainly room for all of them to be developed further. Perhaps the biggest constraint on which approaches will be useful in the future is not the statistical or technical difficulties of modelling, but the type of data which will be available.

5.2 Data: more, and more ancient

It is quite unsatisfying to develop models for which no suitable data exist, so we should look at what data are likely to become available in the next few years before designing our modelling approach. Specifically medical projects will probably continue to provide large datasets, though large genotyping based association-style studies may be less common than smaller sequencing studies. However obtaining these data for the purpose of population genetic research is always difficult in practice and even then, it usually comes with little metadata about the samples, limiting its usefulness.

One alternative source of large datasets are national cohort studies such as the UK Biobank (www.ukbiobank.ac.uk), which combine genetic data with extensive phenotype and metadata. These potentially provide sample sizes in the hundreds of thousands, so it will be important to develop models which can scale to this size. Scalability in terms of running time is obviously important but, as with many "big data" applications, memory use and storage requirements provide the harder limit. The development of methods which can operate on these scales will probably be an important area of research in the next few years. One class of examples of scalable models are the models we developed in Chapter 4. Summary statistics like the f_2 sharing matrix are efficient in the sense that they contain the most information, for the least storage and processing cost. Indeed, in some sense these scale arbitrarily, since the expected number of f_2 variants in a sample does not change

with the sample size. By analogy to the historical inference problems described above, fast methods operating on these reduced datasets can provide initial solutions which can then be refined with slower, more highly specified techniques. Another approach to national cohorts would be to take smaller samples, but carefully choose the sampling strategy to maximise the information available. Projects like the People of the British Isles (POBI; www.peopleofthebritishisles.org) are currently taking this approach, with interesting results though perhaps of limited transferability.

At the end of Chapter 2, we discussed how the analysis of ancient and modern DNA could be integrated. This sort of analysis is likely to become increasingly important. Sequence data from ancient DNA will probably be sparse for the foreseeable future and is fundamentally limited by the rapid degradation of DNA in ancient samples. However, genotyping at specific loci is much easier and it is possible that we could obtain relatively large samples with this information. This has the potential to be informative, not just about selection, as we demonstrated, but also about both modern and ancient population structure. Imagine that we had identified and dated f_2 variants in a large modern sample and used them to make inferences about migration as we did in Chapter 4. If we typed those f_2 variants in ancient samples, then since we know the age and location of the ancient samples, we could locate the origins of modern shared haplotypes much more accurately in space and time, and resolve specific migration events. For example, suppose we observed some relatively recent haplotypes shared between modern European and modern African populations, suggesting post-split migration. By typing ancient samples from both Europe and Africa at the f_2 variants which effectively tag this shared haplotype, we could determine the direction of the migration, which would not be possible from modern data alone. Similarly, if we observe substructure in a modern population (for example, the GBR samples in Figure 4.4), looking for the relevant f_2 variants in ancient samples would tell us whether this structure is persistent ancient structure, or a result of recent population movements. In general, the point here is that the informative rare alleles from modern data are the ones that we should target in ancient samples, to maximise the amount of information obtained from limited material.

5.3 Humans: when and where

Given the likely availability of datasets which are large or ancient (though probably not both), what interesting questions will we be able to answer? Human population structure can only really be understood in terms of history, and it is likely that much research will focus on understanding the historical processes which have led to the structure we see today, and their relative importance. With the ability to analyse data from ancient populations we

can answer not only questions about specific historical events, but more general questions about the way in which populations have developed, For example,

- To what extent does modern population structure reflect ancient population structure? We see relatively continuous structure across Europe, for example, but we do not know how much of this is due to recent isolation by distance rather than to ancient structure, or directional migration. Similarly, to what extent have the large-scale population movements of history involved the replacement of populations, and how much mixture has there been?

- Similarly, how important has selection been, relative to structure, for creating the phenotypic differences between populations which we see today. Given that classical selective sweeps appear not to be an important feature of recent human evolution, and evidence from GWAS that many interesting phenotypes have a genetic architecture which is close to infinitesimal[†], it is a reasonable hypothesis that much recent selection has acted polygenically. Ancient DNA will allow us to determine when and where this selection occurred, and to relate it to environmental or cultural factors such as changes in climate or diet.

- Some of the most important selective pressures on humans, which vary greatly in space and time, are those from infectious disease. Investigating this is particularly difficult and interesting because we must take into account not only the population structure of humans, but also the population structure of the pathogen (and, potentially, of its non-human hosts). This sort of modelling, of diseases like malaria, is obviously of great practical interest and has the potential to greatly improve our treatment and control of these diseases.

Of course, the models we have worked on are relevant in many settings other than human populations. Ecologists have driven development of spatially explicit models in many areas, and the growing availability of genetic data in that field has the potential to greatly improve population management and conservation. Cancer geneticists are starting to investigate within-tumour variation, and it will soon be practical to explicitly model the spatial structure of developing tumours - perhaps one area in which three-dimensional models will be important. Understanding population structure will continue to be an important part of association studies, perhaps even more important as people try to conduct cross-population studies. Indeed, the further development of structured population models will be driven by the need to analyse large and complex datasets, and to integrate that analysis with diverse sources of non-genetic data.

[†]In the sense that the genetic risk is controlled by a large number of loci, each with very small effect.

Appendix A

Local selection in subdivided populations

In Chapter 2, we considered allele frequency trajectories in a lattice model where we allowed selection coefficients to vary spatially. We pointed out that this implies that if some of the selection coefficients were of different signs, it would possible for the allele frequencies to reach an internal equilibrium, which would look like balancing selection due to an overdominant allele if the subdivisions were not observed and we treated all the demes as a single population. This observation was first made by Levene (1953) and these systems have been heavily studied since. The exact value of this internal equilibrium depends on the relative values of the selection coefficients and the migration rates. Here, we gain some intuition for the result in the lattice model, by considering the two-deme case which can be solved exactly.

This model is the simplest possible migration model. We have two demes, both of equal population size N_e, each evolving according to a Wright-Fisher process with bidirectional migration at constant rate $m = \frac{M}{N_e}$ per generation, just as in the lattice model in Chapter 2. let the selection coefficients in demes 1 and 2 be s^1 and s^2 respectively and let the corresponding allele frequencies at time t be f_t^1 and f_t^2 respectively. Also write $f_t = \frac{f_t^1 + f_t^2}{2}$. If there is are unique equilibria for the frequencies we denote them by $\bar{f}^1$, $\bar{f}^2$ and $\bar{f}$. We assume without loss of generality that $s^1 > 0$ and $s^2 < 0$. Then, as long as N_e is sufficiently large, we can model the changes in frequency as normally distributed:

$$f_t^1 - f_{t-1}^1 \sim N\left(s^1(1-m)f_t^1(1-f_t^1) + m(f_t^2 - f_t^1)\ ,\ \frac{f_t^1(1-f_t^1)}{2N_e}\right) \tag{A.1}$$

$$f_t^2 - f_{t-1}^2 \sim N\left(s^1(1-m)f_t^1(1-f_t^1) + m(f_t^2 - f_t^1)\ ,\ \frac{f_t^1(1-f_t^1)}{2N_e}\right). \tag{A.2}$$

We can approximate the rate of change of the expected frequency at time t, $f^1(t)$ and $f^2(t)$

by the differential equations:

$$\frac{df^1}{dt} = s^1(1-m)f_t^1(1-f_t^1) + m(f_t^2 - f_t^1) \tag{A.3}$$

$$\frac{df^2}{dt} = s^2(1-m)f_t^2(1-f_t^2) + m(f_t^1 - f_t^2). \tag{A.4}$$

Setting these equal to zero and solving, we obtain the equilibrium values for the frequencies:

$$\bar{f}^1 = \frac{s^1 - m(2+s^1) + \frac{\sqrt{s^1}\sqrt{-4m^2+(m-1)^2 s^1 s^2}}{\sqrt{s^2}}}{2(m-1)s^1} \tag{A.5}$$

$$\bar{f}^2 = \frac{s^2 - m(2+s^2) - \frac{\sqrt{s^2}\sqrt{-4m^2+(m-1)^2 s^1 s^2}}{\sqrt{s^1}}}{2(m-1)s^2}. \tag{A.6}$$

These solutions are shown in Figure A.1a. The equilibrium value for the average frequency,

$$\bar{f} = \frac{1}{2} - \frac{m}{1-m}\frac{s^1+s^2}{2s^1 s^2} \tag{A.7}$$

is plotted in Figures A.1b-c. In order for there to be an internal equilibrium with $0 < \bar{f} < 1$, we must have

$$\begin{aligned} \frac{1}{2} &> \left| \frac{m}{1-m}\frac{s^1+s^2}{2s^1 s^2} \right| \\ m &< \frac{\left|s^1 s^2\right|}{|s^1 s^2| + |s^1 + s^2|} \end{aligned} \tag{A.8}$$

so there is only an equilibrium if m is sufficiently small. In general, the smaller in magnitude s^1 and s^2 are, the smaller the maximum value of m to give an equilibrium (Figure A.1d). However, note that there is an equilibrium, for some range of m whatever the value of s^1 and s^2, as long as they are of opposite signs.

We have demonstrated that a subdivided population with internal variation in selection coefficients can reach an internal allele frequency equilibrium. However, this can also be a signature of overdominant selection (i.e. $h > 1$ in our notation). The equilibrium value $\bar{f}$ and h are related by $h = \frac{\bar{f}}{2\bar{f}-1}$. Suppose the selection coefficient is s, then the trajectory of the expected allele frequency can be approximated with

$$\frac{df}{dt} = sf_t(1-f_t)(1-h(2-f_t)). \tag{A.9}$$

Therefore for any allele reaching equilibrium in our subdivided population, we can find an equivalent panmictic population, with an allele under overdominant selection, that will reach the same equilibrium value. Can we distinguish these effects? In principle yes. At the equilibrium value, the variance of the average allele frequency in the subdivided model is smaller than in the overdominant model (since the function $f(1-f)$ is concave). If we are

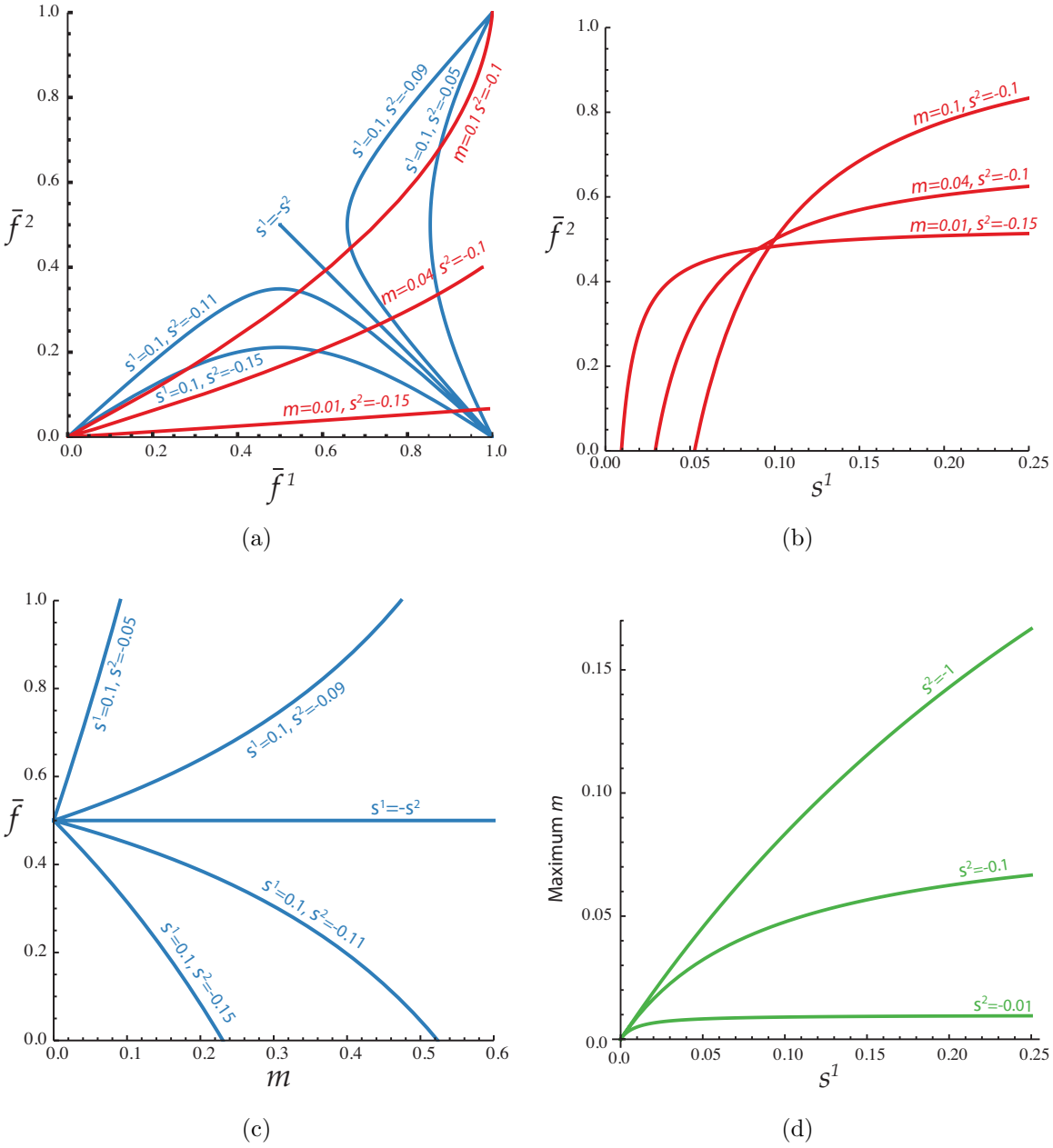


Figure A.1: Equilibrium frequencies in the two deme model.**a**: $\bar{f}^1$ and $\bar{f}^2$ as a function of the parameters. The blue lines show s^1 and s^2 held constant and the equilibrium value moves along the line as m varies. The red lines show m and s^2 held constant and the equilibrium value moves along the line as s^1 varies. This fills in the lower right corner of the quadrant because we assumed that $s^1 > 0 > s^2$. There is a corresponding solution in the upper left if $s^1 < 0 < s^2$. **b** Overall equilibrium value $\bar{f}$ as a function of s^1 with m and s^2 held constant. These parameter values for these lines correspond to the red lines in **a**. **b** Overall equilibrium value $\bar{f}$ as a function of m with s^1 and s^2 held constant.Lines correspond to the blue lines in **a**. **d**: Maximum value of m which has an equilibrium, as a function of s^1, with different lines showing different values of s^2.

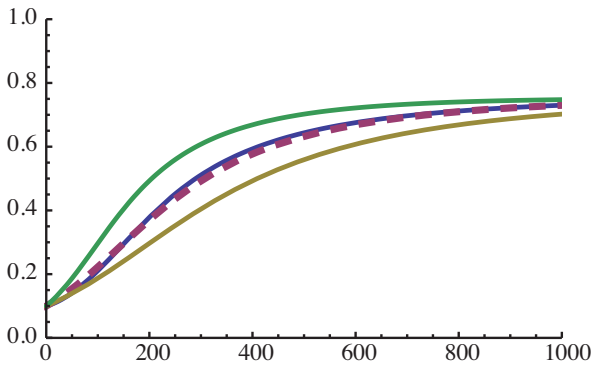


Figure A.2: Expected allele frequency trajectories for the subdivided and overdominant models (obtained by numerically solving the ODEs in Equations A.3, A.4 and A.9). The purple dashed line is the expected trajectory of f_t in the subdivided model with $s^1 = 0.02$, $s^2 = -0.01$, $m = 0.01$ and $f_0^1 = f_0^2 = 0.1$. The solid lines are expected trajectories of f_t in an overdominant model with $h = 1.49$, chosen to have the same equilibrium as the subdivided model, $f_0 = 0.1$, with the three lines representing $s = 0.006$, 0.08 and 0.12. The line with $s = 0.08$ is virtually indistinguishable from, though not identical to, the subdivided model.

not at the equilibrium then in general, the dynamics of the expected allele frequency are not the same. For example, in the subdivided model, we cannot write $\frac{df_t}{dt} = \frac{1}{2}\left(\frac{df_t^1}{dt} + \frac{df_t^2}{dt}\right)$ in terms of f_t. Theoretically then, if we saw the trajectory in an infinitely large population, we could tell the difference. In practice however, this is almost certainly impossible. Random fluctuations in the allele frequency will obscure the difference in mean value and there are always parameters which make the expected trajectory virtually indistinguishable, if not mathematically identical (Figure A.2). On the other hand, the effects of these two types of selection on linked neutral loci can be distinguished (Charlesworth *et al.*, 1997). Local selection results in more between-deme diversity than balancing selection. Therefore if additional sequence data is available, it would probably be possible to distinguish the two relatively easily.

What intuition can we gain for the lattice model from this analysis? We expect that there will be an internal equilibrium if the migration rate is low. Equation A.8 suggests that the maximum migration rate for an equilibrium to exist is determined by the absolute value of the sum of the selection coefficients in neighbouring demes and that the equilibrium exists for the largest possible m (i.e. 1) when neighbouring demes have equal but opposite selection coefficients[†]. If we just saw the whole population, the allele frequency trajectory would most likely be effectively indistinguishable from the trajectory of an allele in a single population under overdominant selection.

[†]Like a chessboard where the selection coefficient is s in black squares and $-s$ in white squares.

Appendix B

Further examples of the effect of spatial risk on association studies

In this appendix we present some examples of systematic investigation of the effect of different patterns of spatial risk on association studies, to support the conclusions made in Chapter 3. For Figures B.1 to B.3, the left and right hand columns correspond to Figures 3.8a and 3.8c respectively and all parameters are the same as in Figure 3.8 unless otherwise stated. For Figures B.4 and B.5, we show only the qq plots.

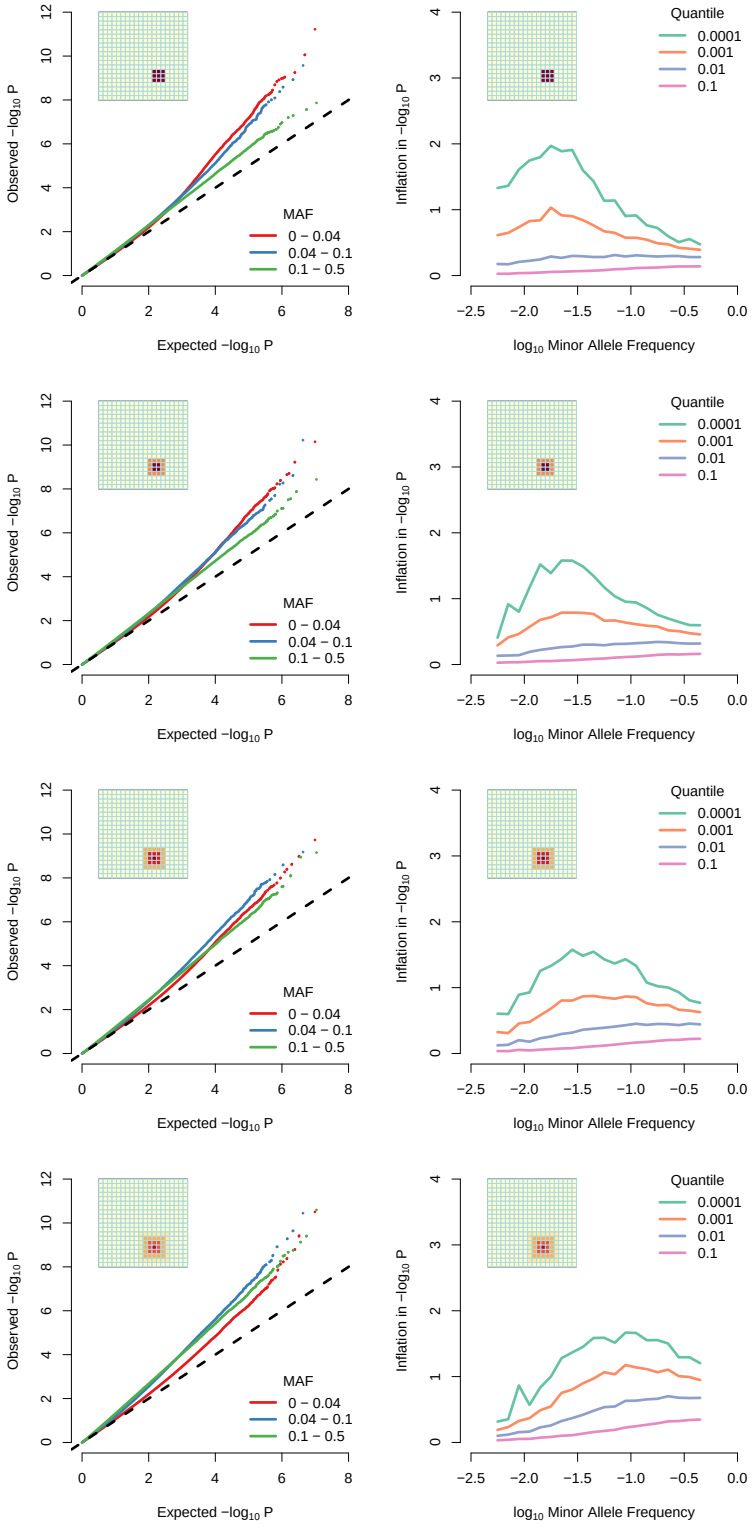


Figure B.1: Inflation in P-values as the smoothness of the risk patch increases. As the risk gets smoother, the effect of the inflation in rare variants becomes less pronounced.

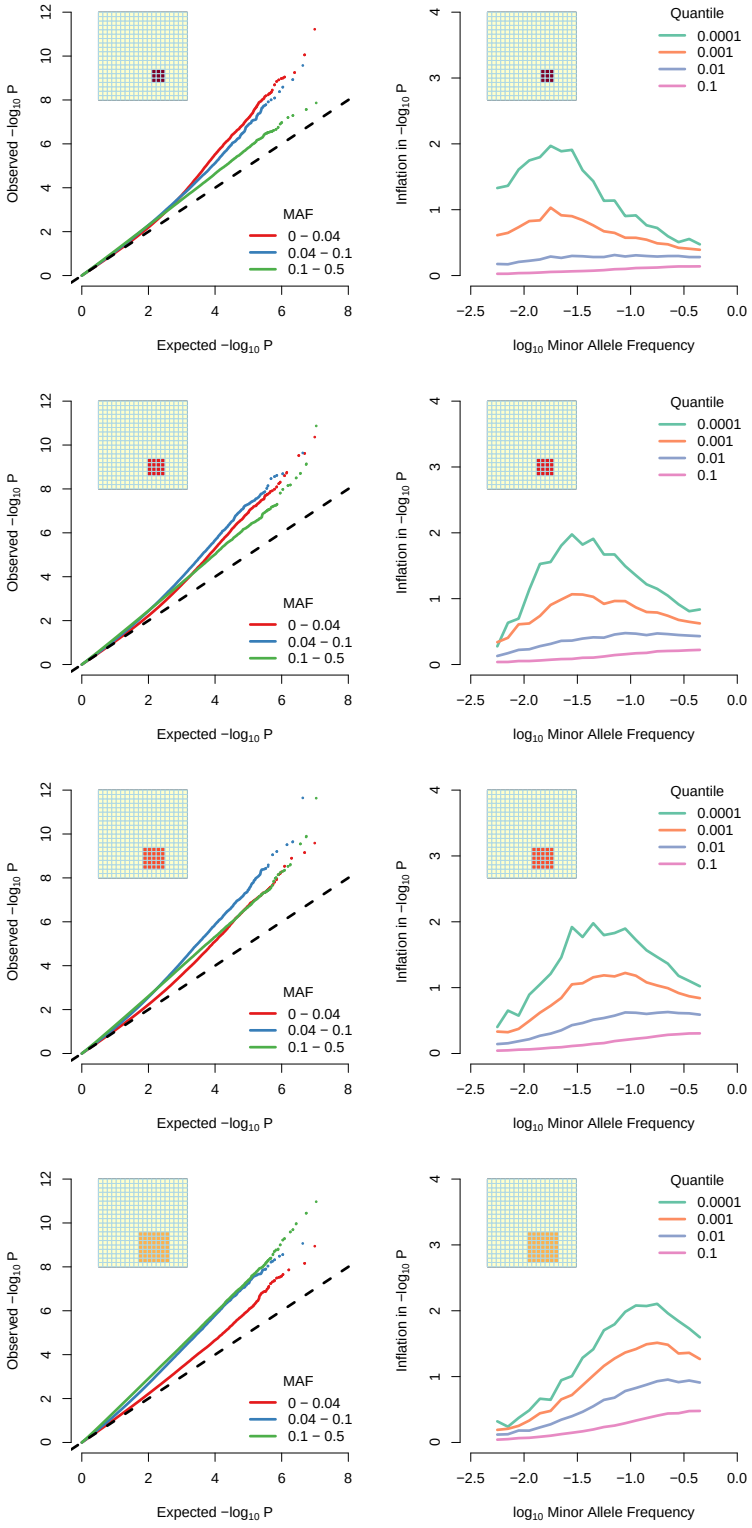


Figure B.2: Inflation in P-values as the size of the risk patch increases. As in Figure B.1, as the size of the patch of risk increases, the inflation in rare variants becomes less pronounced.

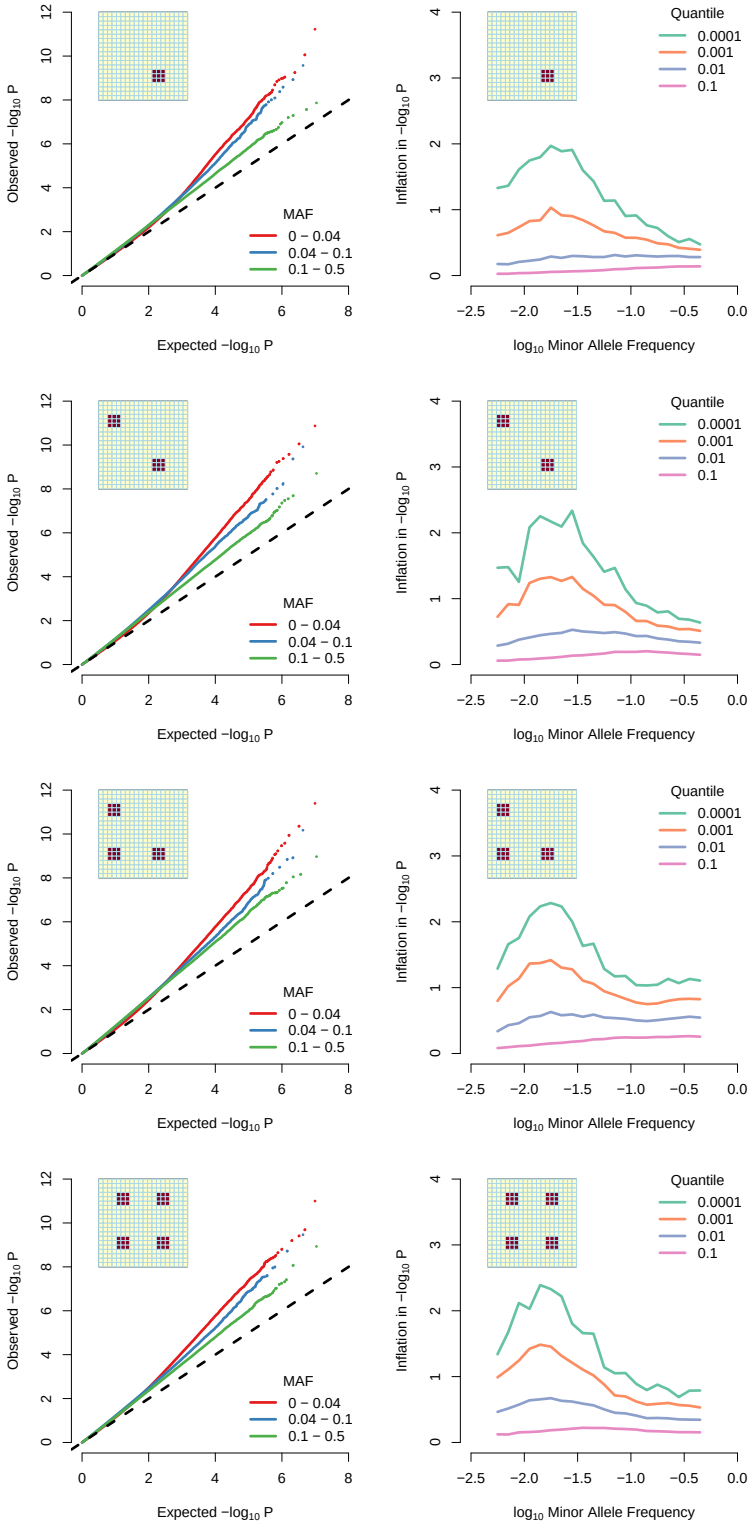


Figure B.3: Inflation in P-values as small disconnected patches of risk are added. As more patches of the same size are added, the behaviour of does not change significantly.

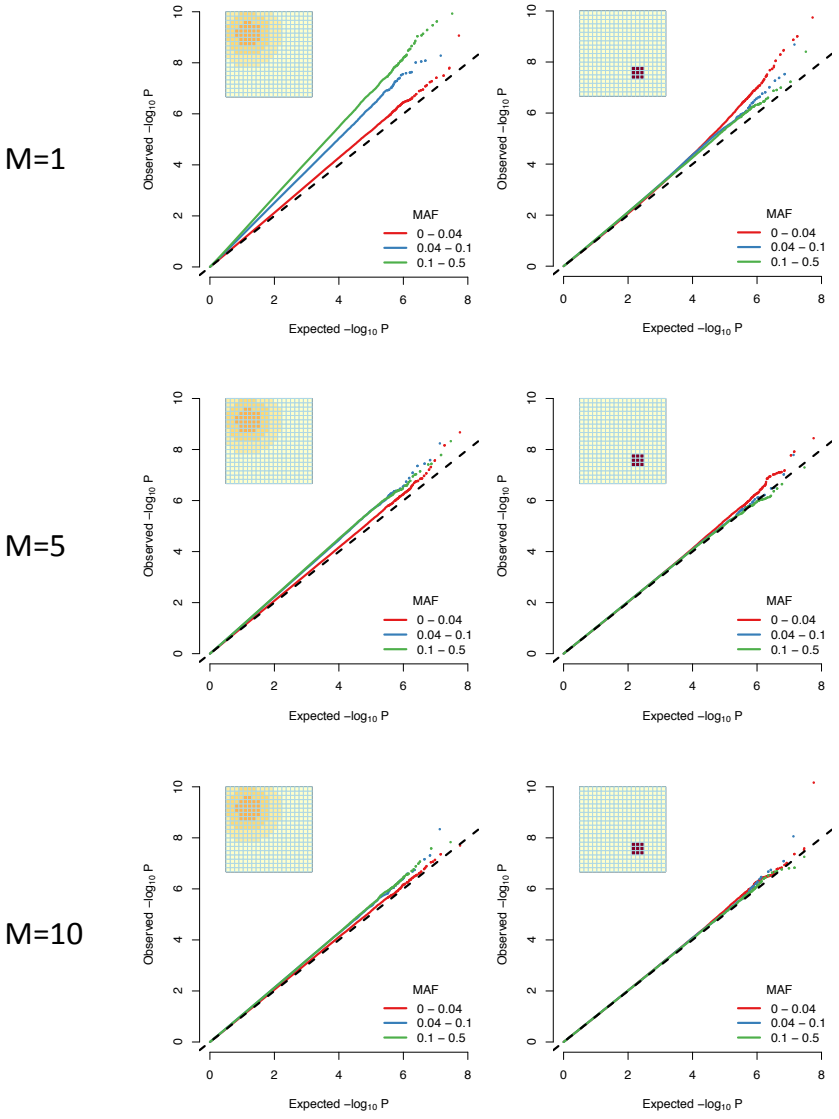


Figure B.4: Inflation in P-values as migration rate is increased. As migration rate increases there is much less inflation overall, but rare variants are still noticably more inflated than common ones.

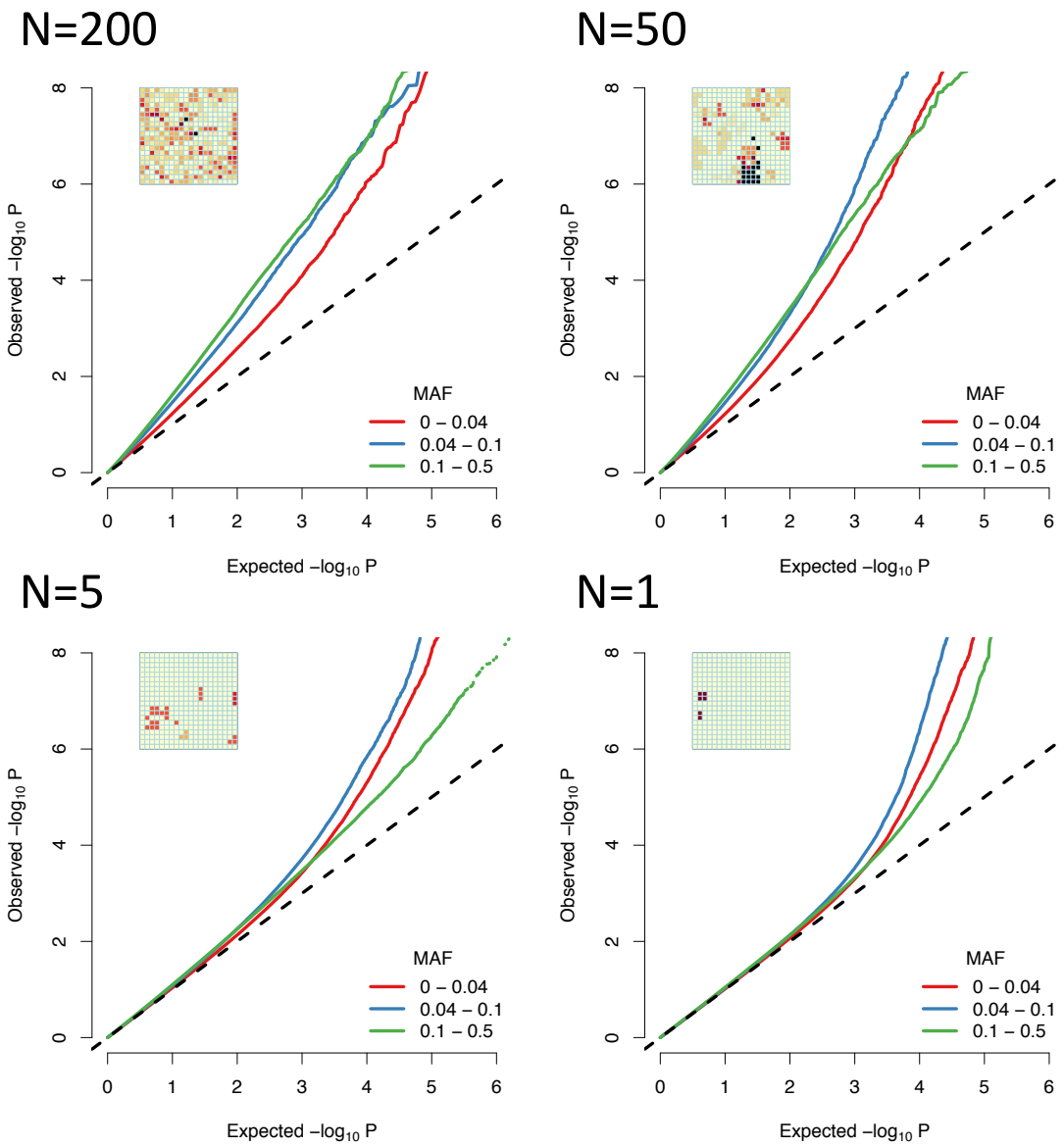
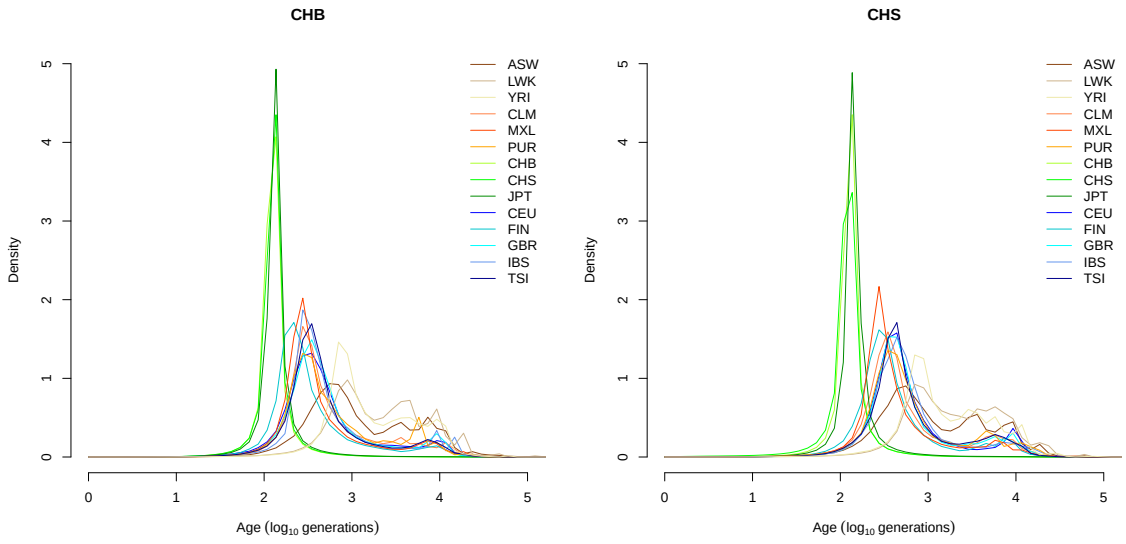


Figure B.5: Illustration of the effect of background genetic risk from rare variants. Each of these plots shows the qq plot for a model where the phenotype is influenced by N rare variants at independent loci, not tested for association. The effect of each rare variant is drawn from a normal distribution with mean 0 and standard deviation ranging from 0.2 in the N=200 case to 1 in the N=1 case. The grids in the top left corners of each plot show examples of the resulting spatial pattern of (absolute) phenotypic mean, which was resimulated for each genealogy. Other parameters are the same as in Figure 3.8.

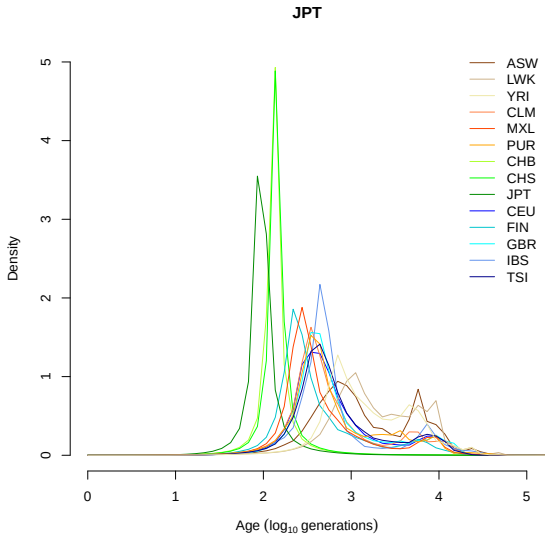
Appendix C

Distributions of 1000 Genomes haplotype ages

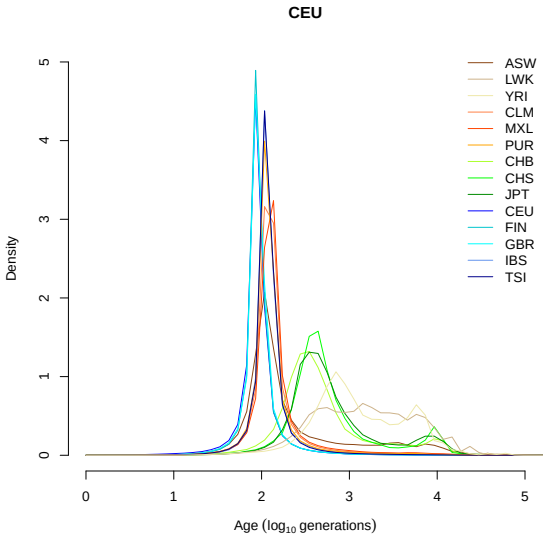


(a) CHB (b) CHS

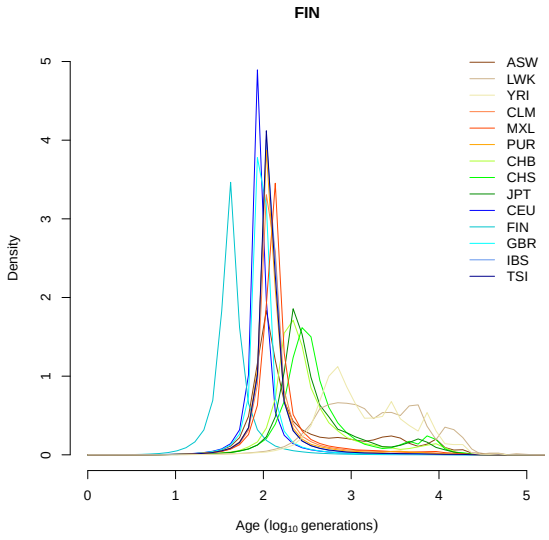
Figure C.1: Haplotype age distributions for ages less than 2000 generations, as in Figure 4.13, for all 1000 Genomes populations.



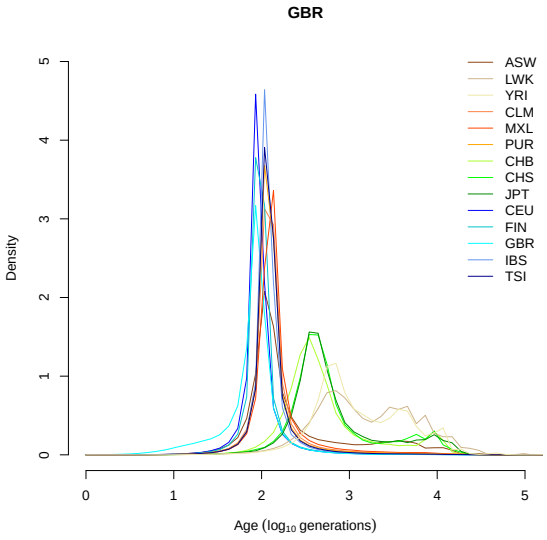
(c) JPT



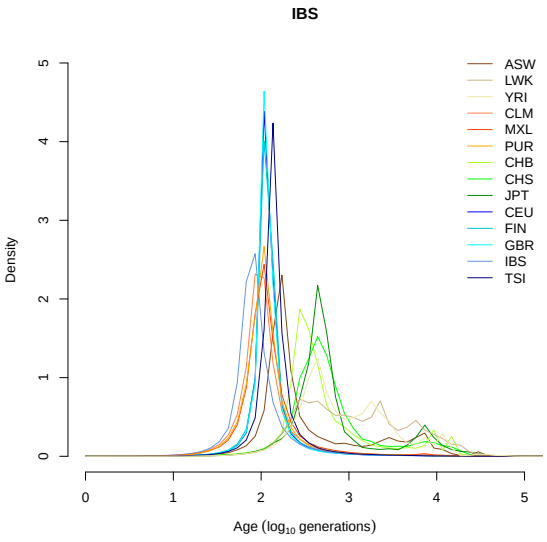
(d) CEU



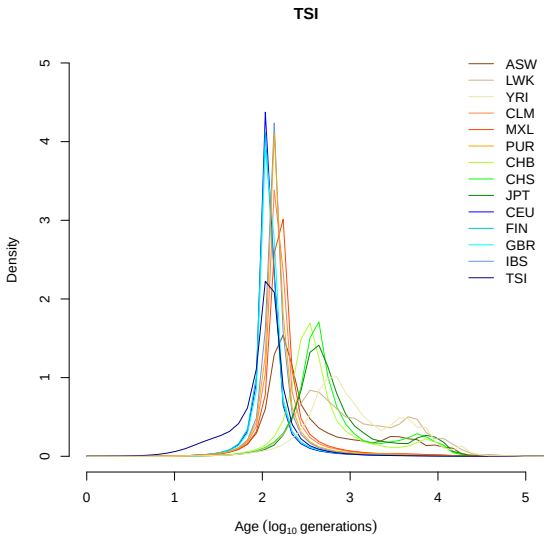
(e) FIN



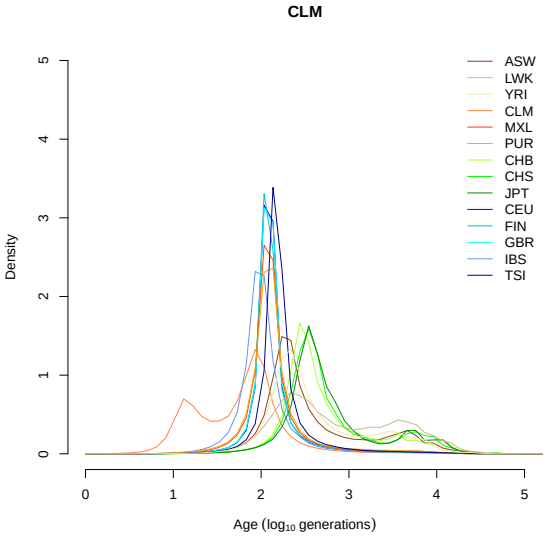
(f) GBR



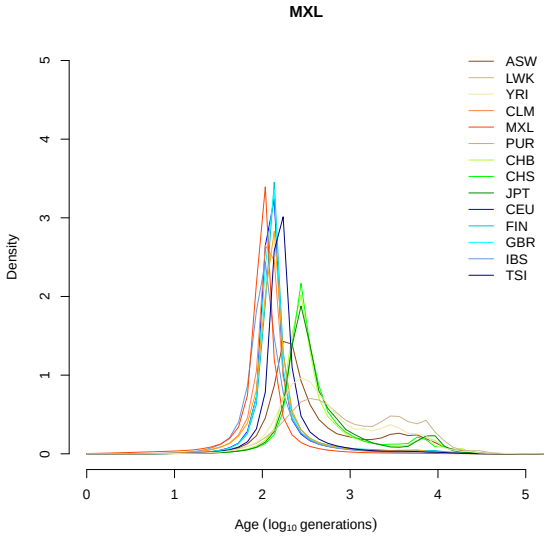
(g) IBS



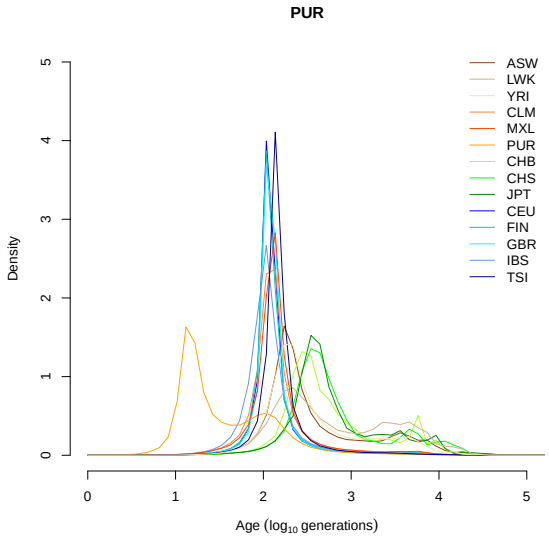
(h) TSI



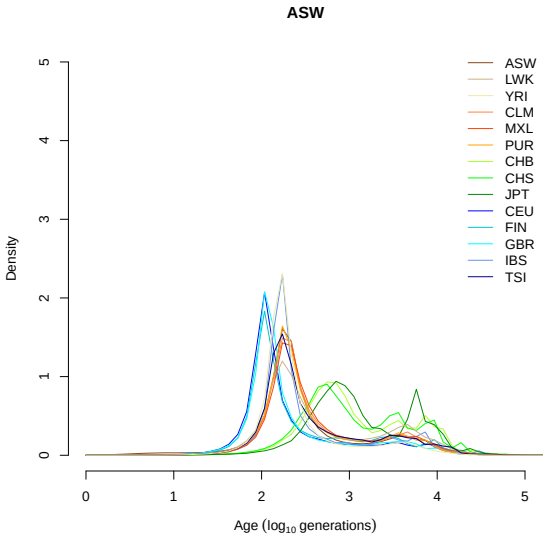
(i) CLM



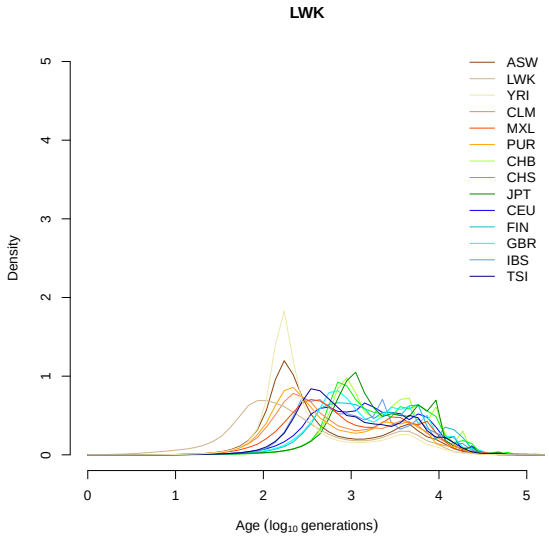
(j) MXL



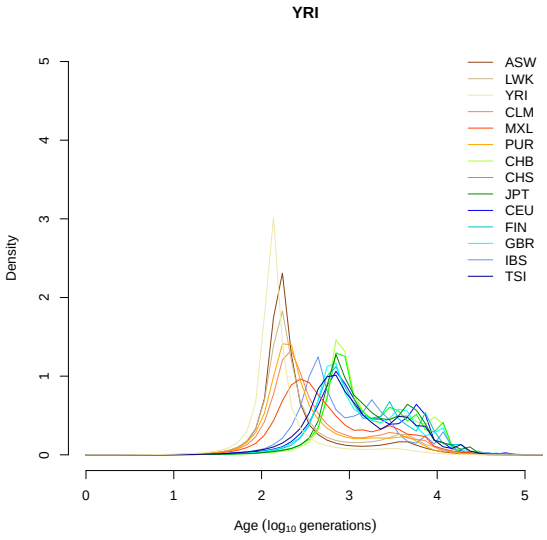
(k) PUR



(l) ASW



(m) LWK



(n) YRI

Appendix D

A note on tools

Most of the code used for simulation, data anlysis and visualisation in this thesis was written in R (www.r-project.org). Useful packages not mentioned directly in the main text included *MASS*, *multicore*, *ggplot2* (Wickham, 2009) and *SNPrelate* (Zheng *et al.*, 2012). Some additional code was written in C and python (www.python.org). *Mathematica* was used to solve some of the equations and check solutions in Section 2.2.2 and Appendix A. Data figures were drawn in R and descriptive figures prepared in *Adobe Illustrator* or using the LaTeX package *tikz*. All coding and writing was done in the *Aquamacs* (www.aquamacs.org) IDE, which is a fork of *emacs* (www.gnu.org/s/emacs/). Typeset in LaTeX.

References

1000 Genomes Project Consortium, 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Abbott, E. A. *Flatland*. Seely.

Akaike, H., 1974. A new look at the statistical model identification. *Automatic Control* **19**: 716–723.

Alexander, D. H., J. Novembre, and K. Lange, 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**: 1655–1664.

Anderson, E. C., E. G. Williamson, and E. A. Thompson, 2000. Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics* **156**: 2109–2118.

Austerlitz, F., B. Jung-Muller, and B. Godelle, 1997. Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Genetics* **51**: 148–164.

Babron, M., M. de Tayrac, D. N. Rutledge, E. Zeggini, and E. Génin, 2012. Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PLoS One* **7**: e46519.

Bacanu, S. and B. Devlin, 2000. The power of genomic control. *The American Journal of Human Genetics* **66**: 1933–1944.

Bahlo, M. and R. C. Griffiths, 2000. Inference from gene trees in a subdivided population. *Theoretical Population Biology* **57**: 79–95.

Barton, N. H., F. Depaulis, and A. M. Etheridge, 2002. Neutral evolution in spatially continuous populations. *Theoretical Population Biology* **61**: 31–48.

Barton, N. H., A. M. Etheridge, and A. Véber, 2010. A new model for evolution in a spatial continuum. *Electronic Journal of Probability* **15**: 162–216.

Beerli, P. and J. Felsenstein, 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**: 763–773.

Beerli, P. and J. Felsenstein, 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences* **98**: 4563–4568.

Berry, A. J., J. W. Ajioka, and M. Kreitman, 1991. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. *Genetics* **129**: 1111–1117.

Bersaglieri, T., P. C. Sabeti, and N. Patterson, 2004. Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74**: 1111–1120.

Bignell, G. R., C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, et al., 2010. Signatures of mutation and selection in the cancer genome. *Nature* **463**: 893–898.

Billera, L. J., S. P. Holmes, and K. Vogtmann, 2001. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics* **27**: 733–767.

Bishop, J., L. M. Cook, and J. Muggleton, 1978. The response of two species of moths to industrialization in northwest England. I. Polymorphisms for melanism. *Philosophical Transactions of the Royal Society of London Series B* **281**: 489–515.

Blum, M. G. B. and O. François, 2005 Minimal clade size and external branch length under the neutral coalescent. *Advances in Applied Probability* **37**: 647–662.

Bollback, J. P., T. L. York, and R. Nielsen, 2008. Estimation of $2N_e s$ from temporal allele frequency data. *Genetics* **179**: 497–502.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, and E. Minch, 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455–457.

Browning, B. L. and S. R. Browning, 2011. A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics* **88**: 173–182.

Burger, J., M. Kirchner, B. Bramanti, W. Haak, and M. G. Thomas, 2007 Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proceedings of the National Academy of Sciences* **104**: 3736–3741.

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

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, et al., 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437**: 1153–1157.

Caliński, T. and J. Harabasz, 1974. A dendrite method for cluster analysis. *Communications in Statistics- Theory and Methods* **3**: 1–27.

Cavalli-Sforza, L. L., 1966. Population structure and human evolution. *Proceedings of the Royal Society of London Series B* **164**: 362–379.

Cavalli-Sforza, L. L. and I. Barrai, 1964. Analysis of human evolution under random genetic drift. *Cold Spring Harbor Symposia on Quantitative Biology* **29**: 9–20.

Charlesworth, B., 1998. Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution* **15**: 538–543.

Charlesworth, B., M. Nordborg, and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research* **70**: 155–174.

Chen, C., E. Durand, and F. Forbes, 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology* **7**: 747–756.

Chen, G. K., P. Marjoram, and J. D. Wall, 2009. Fast and flexible simulation of DNA sequence data. *Genome Research* **19**: 136–142.

Chen, J., H. Zheng, J. X. Bei, L. Sun, W. Jia, et al., 2009. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *The American Journal of Human Genetics* **85**: 775–785.

Clarke, C. A., B. S. Grant, D. F. Owen, and T. Asami, 1994. A long term assessment of *Biston betularia* (L.) in one UK locality (Caldy Common near West Kirby, Wirral), 1959-1993, and glimpses elsewhere. *Linnean* **10**: 18–26.

Cockerham, C. C., 1969. Variance of gene frequencies. *Evolution* **23**: 72–84.

Cook, L. M., 2003. The rise and fall of the *carbonaria* form of the peppered moth. *Quarterly Review of Biology* **78**: 399–417.

Cook, L. M., R. L. H. Dennis, and G. S. Mani, 1999. Melanic morph frequency in the peppered moth in the Manchester area. *Proceedings of the Royal Society of London Series B* **266**: 293–297.

Cook, L. M., B. S. Grant, I. J. Saccheri, and J. Mallet, 2012 Selective bird predation on the peppered moth: the last experiment of Michael Majerus. *Biology letters* **8**: 609–612.

Cook, L. M. and D. A. Jones, 1996. The *medionigra* gene in the moth *Panaxia dominula*: The case for selection. *Philosophical Transactions of the Royal Society of London Series B* **351**: 1623–1634.

Cook, L. M., A. M. Riley, and I. P. Woiwod, 2002. Melanic frequencies in three species of moths in post industrial Britain. *Biological Journal of the Linnean Society* **75**: 475–482.

Cook, L. M., S. L. Sutton, and T. J. Crawford, 2005. Melanic moth frequencies in Yorkshire, an old English industrial hot spot. *Journal of Heredity* **96**: 522–528.

Cook, L. M. and J. R. G. Turner, 2008. Decline of melanism in two British moths: spatial, temporal and inter-specific variation. *Heredity* **101**: 483–489.

Copeland, K. T., H. Checkoway, A. J. McMichael, and R. H. Holbrook, 1977. Bias due to misclassification in estimation of relative risk. *American Journal of Epidemiology* **105**: 488–495.

Cox, J. T. and R. Durrett, 2002. The stepping stone model: New formulas expose old myths. *The Annals of Applied Probability* **12**: 1348–1377.

DeGiorgio, M. and N. A. Rosenberg, 2013. Geographic sampling scheme as a determinant of the major axis of genetic variation in principal components analysis. *Molecular Biology and Evolution* **30**: 480–488.

Devlin, B. and K. Roeder, 1999 Genomic control for association studies. *Biometrics* **55**: 997–1004.

Diggle, P. and R. Gratton, 1984. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society Series* **46**: 193–227.

DuMouchel, W. H. and W. W. Anderson, 1968 The Analysis of Selection in Experimental Populations. *Genetics* **58**: 435–449.

Dunne, J., R. P. Evershed, M. Salque, L. Cramp, and S. Bruni, 2012. First dairying in green Saharan Africa in the fifth millennium BC. *Nature* **486**: 390–394.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge: Cambridge University Press.

Enattah, N. S., T. Jensen, M. Nielsen, and R. Lewinski, 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *The American Journal of Human Genetics* **82**: 57–72.

Enattah, N. S., T. Sahi, E. Savilahti, and J. D. Terwilliger, 2002. Identification of a variant associated with adult-type hypolactasia. *Nature* **30**: 233–237.

Ewens, W. *Mathematical population genetics*. New York: Springer.

Falush, D., M. Stephens, and J. K. Pritchard, 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Felsenstein, J., 1975. A pain in the torus: some difficulties with models of isolation by distance. *The American Naturalist* **109**: 359–368.

Fenner, J. N., 2005. Crosscultural estimation of the human generation interval for use in geneticsbased population divergence studies. *American Journal of Physical Anthropology* **128**: 415–423.

Fisher, R., 1937. The wave of advance of advantageous genes. *Annals of Eugenics* **7**: 353–367.

Fisher, R. A. and E. B. Ford, 1947. The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. *Heredity* **1**: 143–174.

Flatz, G., 1984 Gene-dosage effect on intestinal lactase activity demonstrated in vivo. *American Journal of Human Genetics* **36**: 306–310.

François, O., S. Ancelet, and G. Guillot, 2006. Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**: 805–816.

Fumagalli, M., M. Sironi, U. Pozzoli, A. Ferrer-Admettla, L. Pattini, et al., 2011 Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics* **7**: e1002355.

Gibson, G., 2011 Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13**: 135–145.

Grant, B. S., A. D. Cook, C. A. Clarke, and D. F. Owen, 1998. Geographic and temporal variation in the incidence of melanism in peppered moth populations in America and Britain. *Journal of Heredity* **89**: 465–471.

Grant, B. S., D. F. Owen, and C. A. Clarke, 1996. Parallel rise and fall of melanic peppered moths in America and Britain. *Journal of Heredity* **87**: 351–357.

Gravel, S., B. M. Henn, and R. N. Gutenkunst, 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* **108**: 11983–11988.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, and U. Stenzel, 2010. A draft sequence of the Neandertal genome. *Science* **328**: 710–722.

Grün, B. and F. Leisch, 2008. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* **28**: 1–35.

Guillot, G., A. Estoup, F. Mortier, and J. F. Cosson, 2005. A spatial statistical model for landscape genetics. *Genetics* **170**: 1261–1280.

Gutenkunst, R. N., R. D. Hernandez, and S. H. Williamson, 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**: e1000695.

Haldane, J., 1956. The theory of selection for melanism in Lepidoptera. In *Proceedings of the Royal Society of London Series B*, pp. 303–306.

Haldane, J. B. S., 1954 An exact test for randomness of mating. *Journal of Genetics* **52**: 631–635.

Hancock, A. M., D. B. Witonsky, G. Alkorta-Aranburu, C. M. Beall, A. Gebremedhin, et al., 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genetics* **7**: e1001375.

Harpending, H. and T. Jenkins, 1973. Genetic distance among southern African populations. In M. Crawford and P. Workman (Eds.), *Methods and Theories of Anthropological Genetics*. University of New Mexico Press.

Harris, K. and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics* **9**: e1003521.

Henderson, C. R. *Applications of linear models in animal breeding*. University of Guelph Press.

Herbots, H., 1997 The structured coalescent. In P. Donnelly and S. Tavaré (Eds.), *Progress in population genetics and human evolution*. New York: Springer.

Hermisson, J. and P. S. Pennings, 2005. Soft sweeps molecular population genetics of adaptation from standing genetic variation. *Genetics* **14**: 2335–2352.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, et al., 2011. Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920–924.

Hindorff, L. A., H. A. Junkins, P. N. Hall, J. P. Mehta, and T. A. Manolio, A catalogue of published genome-wide association studies (www.genome.gov/gwastudies).

Ho, M. W., S. Povey, and D. Swallow, 1982 Lactase polymorphism in adult British natives: estimating allele frequencies by enzyme assays in autopsy samples. *American Journal of Human Genetics* **34**: 650–657.

Holsinger, K. E. and B. S. Weir, 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics* **10**: 639–650.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *The Journal of Educational Psychology* **23**: 498–520.

Hudson, R. R., 1990. Gene genealogies and the coalescent process. In D. Futuyama and J. Antonovics (Eds.), *Oxford surveys in evolutionary biology volume 7*. Oxford University Press.

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

Hudson, R. R., M. Slatkin, and W. P. Maddison, 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.

Illingworth, C. J. and V. Mustonen, 2011. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics* **189**: 989–1000.

Ingram, C., C. A. Mulcare, Y. Itan, and M. G. Thomas, 2009. Lactose digestion and the evolutionary genetics of lactase persistence. *Human Genetics* **124**: 579–591.

International HapMap Consortium, 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**: 52–58.

Ionita-Laza, I., J. D. Buxbaum, N. M. Laird, and C. Lange, 2011. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics* **7**: e1001289.

Itan, Y., A. Powell, M. A. Beaumont, and J. Burger, 2009. The origins of lactase persistence in Europe. *PLoS Computational Biology* **5**: e1000491.

Jablonski, N. G. and G. Chaplin, 2010 Colloquium paper: human skin pigmentation as an adaptation to UV radiation. *Proceedings of the National Academy of Sciences* **107**: 8962–8968.

Jones, D. A., 2000. Temperatures in the Cothill habitat of *Panaxia* (*Callimorpha*) *dominula* L. (the scarlet tiger moth). *Heredity* **84**: 578–586.

Jost, L., 2008. G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015–4026.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, et al., 2008. Efficient control of population structure in model organism association mapping. *Genetics* **178**: 1709–1723.

Kettlewell, H. B. D., 1955. Selection experiments on industrial melanism in the *Lepidoptera*. *Heredity* **9**: 323–342.

Kettlewell, H. B. D., 1956. Further selection experiments on industrial melanism in the *Lepidoptera*. *Heredity* **10**: 287–301.

Kettlewell, H. B. D., 1958. A survey of the frequencies of *Biston betularia* (L.) (Lep.) and its melanic forms in Great Britain. *Heredity* **12**: 51–72.

Kimura, M., 1953 "Stepping stone" model of population. *Annual Report of the National Institute of Genetics* **3**: 63–65.

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

Kimura, M. and G. H. Weiss, 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**: 561–576.

Klein, R. J., 2005 Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**: 385–389.

Knowler, W. C., R. C. Williams, D. J. Pettitt, and A. G. Steinberg, 1988. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics* **43**: 520–526.

Kuhner, M. K., 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**: 768–770.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, et al., 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**: 832–838.

Lao, O., T. T. Lu, M. Nothnagel, O. Junge, and S. Freitag-Wolf, 2008. Correlation between genetic and geographic structure in Europe. *Current Biology* **18**: 1241–1248.

Lawson, D. J., 2013. Populations in statistical genetic modelling and inference. *arXiv* **arXiv**: 1306.0701v1.

Lawson, D. J. and D. Falush, 2012. Population identification using genetic data. *Annual Review of Genomics and Human Genetics* **13**: 337–361.

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

Le Corre, V. and A. Kremer, 1998. Cumulative effects of founding events during colonisation on genetic diversity and differentiation in an island and stepping stone model. *Journal of Evolutionary Biology* **11**: 495–512.

Lees, D. R. and E. R. Creed, 1975. Industrial melanism in *Biston betularia* - Role of selective predation. *Journal of Animal Ecology* **44**: 67–83.

Lees, D. R. and E. R. Creed, 1977. The genetics of the *insularia* forms of the peppered moth, *Biston betularia*. *Heredity* **39**: 67–73.

Leffler, E. M., Z. Gao, S. Pfeifer, L. Ségurel, and A. Auton, 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**: 1578–1582.

Leisch, F., 2004. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* **86**: 1–16.

Leuenberger, C. and D. Wegmann, 2010. Bayesian computation and model selection without likelihoods. *Genetics* **184**: 243–252.

Levene, H., 1953. Genetic equilibrium when more than one ecological niche is available. *The American Naturalist* **87**: 331–333.

Lewinsky, R. H., T. Jensen, and J. Møller, 2005. T 13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. *Human Molecular Genetics* **14**: 3945–3953.

Li, B. and S. M. Leal, 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics* **83**: 311–321.

Li, H. and R. Durbin, 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493–496.

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

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, et al., 2011 FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**: 833–835.

Listgarten, J., C. Lippert, and D. Heckerman, 2013. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics* **45**: 470–471.

MacHugh, D. E., R. T. Loftus, and P. Cunningham, 1998. Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. *Animal Genetics* **29**: 333–340.

Madsen, B. E. and S. R. Browning, 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* **5**: e1000384.

Mahajan, M., P. Nimbhorkar, and K. Varadarajan, 2012. The planar k-means problem is NP-hard. *Theoretical Computer Science* **442**: 13–21.

Malaspinas, A. S., O. Malaspinas, S. N. Evans, and M. Slatkin, 2012. Estimating allele age and selection coefficient from time-serial data. *Genetics* **192**: 599–607.

Malécot, G. *Les mathèmatiques de l'hèrèditè*. Paris.

Malécot, G., 1959 Les modèles stochastiques en génétique de population. *Institut de Statistique de l'Université de Paris* **8**: 173–210.

Malmström, H., A. Linderholm, and K. Lidén, 2010. High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern Europe. *BMC Evolutionary Biology* **10**: 89.

Mani, G. S. and M. E. N. Majerus, 1993. Peppered moth revisited - Analysis of recent decreases in melanic frequency and predictions for the future. *Biological Journal of the Linnean Society* **48**: 157–165.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, et al., 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.

Marjoram, P. and P. Donnelly, 1994. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* **136**: 673–683.

Maruyama, T., 1970 Effective number of alleles in a subdivided population. *Theoretical Population Biology* **1**: 273–306.

Maruyama, T., 1971a. An invariant property of a structured population. *Genetical Research* **18**: 81–84.

Maruyama, T., 1971b. Analysis of population structure: II. Twodimensional stepping sone models of finite length and other geographically structured populations*. *Annals of Human Genetics* **35**: 179–196.

Maruyama, T. and M. Kimura, 1971. Some methods for treating continuous stochastic processes in population genetics. *Japanese Journal of Genetics* **46**: 407–410.

Maruyama, T. and M. Kimura, 1975. Moments for sum of an arbitrary function of gene frequency along a stochastic path of gene frequency change. *Proceedings of the National Academy of Sciences* **72**: 1602–1604.

Mathieson, I. and G. McVean, 2012 Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics* **44**: 243–246.

Mathieson, I. and G. McVean, 2013 Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**: 973–984.

Maynard Smith, J. and J. Haigh, 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* **23**: 23–35.

McVean, G., 2009. A genealogical interpretation of principal components analysis. *PLoS Genetics* **5**: e1000686.

Melnick, D. J. and K. K. Kidd, 1985. Genetic and evolutionary relationships among Asian macaques. *International journal of primatology* **6**: 123–160.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza, 1978. Synthetic maps of human gene frequencies in Europeans. *Science* **201**: 786–792.

Meyer, M., M. Kircher, M. T. Gansauge, H. Li, and F. Racimo, 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**: 222–226.

Morris, A. P. and E. Zeggini, 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* **34**: 188–193.

Nagy, D., G. Tömöry, and B. Csányi, 2011. Comparison of lactase persistence polymorphism in ancient and presentday Hungarian populations. *American Journal of Human Genetics* **145**: 262–269.

Nagylaki, T., 1982. Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**: 159–172.

Nagylaki, T., 1990. Models and approximations for random genetic drift. *Theoretical Population Biology* **37**: 192–212.

Nagylaki, T., 1998 The expected number of heterozygous sites in a subdivided population. *Genetics* **149**: 1599–1604.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, et al., 2011. Testing for an unusual distribution of rare variants. *PLoS Genetics* **7**: e1001322.

Nei, M., 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70**: 3321–3323.

Nelis, M., T. Esko, R. Magi, F. Zimprich, A. Zimprich, et al., 2009. Genetic structure of Europeans: a view from the North-East. *PLoS One* **4**: e5472.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, et al., 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* **3**: e170.

Nordborg, M., 1997. Structured coalescent processes on different time scales. *Genetics* **146**: 1501–1514.

Norris, J. R. *Markov chains*. Cambridge University Press.

Notohara, M., 1990. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology* **29**: 59–75.

Novembre, J. and A. Di Rienzo, 2009. Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics* **10**: 745–755.

Novembre, J., T. Johnson, K. Bryc, and Z. Kutalik, 2008. Genes mirror geography within Europe. *Nature* **456**: 98–101.

Novembre, J. and M. Stephens, 2008 Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* **40**: 646–649.

Nye, T., 2011. Principal components analysis in the space of phylogenetic trees. *Annals of Statistics* **39**: 2716–2739.

O'hara, R. B., 2005. Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth. *Proceedings of the Royal Society of London Series B* **272**: 211–217.

Palamara, P. F., T. Lencz, A. Darvasi, and I. Pe'er, 2012 Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics* **91**: 809–822.

Pannell, J. R., 2003. Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution* **57**: 949–961.

Patterson, N., P. Moorjani, Y. Luo, and S. Mallick, 2012. Ancient admixture in human history. *Genetics* **192**: 1065–1093.

Patterson, N., A. L. Price, and D. Reich, 2006. Population structure and eigenanalysis. *PLoS Genetics* **2**: e190.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**: 559–572.

Peter, B. M., E. Huerta-Sanchez, and R. Nielsen, 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genetics* **8**: e1003011.

Piazza, A. and P. Menozzi, 1981. Synthetic gene frequency maps of man and selective effects of climate. *Proceedings of the National Academy of Sciences* **78**: 2638–2642.

Pickrell, J. K., N. Patterson, C. Barbieri, and F. Berthold, 2012. The genetic prehistory of southern Africa. *Nature* **3**: 1143.

Pickrell, J. K. and J. K. Pritchard, 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**: e1002967.

Pirinen, M., P. Donnelly, and C. C. A. Spencer, 2012 Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Annals of Applied Statistics* **7**: 369–390.

Plantinga, T. S., S. Alonso, N. Izagirre, M. Hervella, R. Fregel, et al., 2012 Low prevalence of lactase persistence in Neolithic South-West Europe. *European Journal of Human Genetics* **20**: 778–782.

Poulter, M., E. Hollox, and C. B. Harvey, 2003. The causal element for the lactase persistence/nonpersistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Annals of Human Genetics* **67**: 298–311.

Price, A. L., A. Helgason, S. Palsson, and H. Stefansson, 2009. The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genetics* **5**: e1000505.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, et al., 2006 Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**: 904–909.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, and N. Rafaels, 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* **5**: e1000519.

Pritchard, J. K. and N. A. Rosenberg, 1999 Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65**: 220–228.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell, S., S. S. Cherny, and P. C. Sham, 2003 Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**: 149–150.

Purcell, S., B. Neale, K. Todd-Brown, and L. Thomas, 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559–575.

Ralph, P. and G. Coop, 2010. Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics* **186**: 647–668.

Ralph, P. and G. Coop, 2013. The geography of recent genetic ancestry across Europe. *PLoS Biology* **11**: e1001555.

Rannala, B. and J. A. Hartigan, 1996. Estimating gene flow in island populations. *Genetical Research* **67**: 147–158.

Rasmussen, M., X. Guo, Y. Wang, and K. E. Lohmueller, 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**: 94–98.

Reich, D., R. E. Green, M. Kircher, J. Krause, and N. Patterson, 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**: 1053–1060.

Reich, D., N. Patterson, D. Campbell, and A. Tandon, 2012. Reconstructing native American population history. *Nature* **488**: 370–374.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009. Reconstructing Indian population history. *Nature* **461**: 489–494.

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

Rosenberg, N. A., S. Mahajan, S. Ramachandran, and C. Zhao, 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* **1**: e70.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, et al., 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913–918.

Salque, M., P. I. Bogucki, J. Pyzel, and I. Sobkowiak-Tabaka, 2013. Earliest evidence for cheese making in the sixth millennium bc in northern Europe. *Nature* **493**: 522–525.

Sankararaman, S., N. Patterson, H. Li, S. Pääbo, and D. Reich, 2012. The date of interbreeding between Neandertals and modern humans. *PLoS Genetics* **8**: e1002947.

Sawcer, S., G. Hellenthal, M. Pirinen, C. C. A. Spencer, N. A. Patsopoulos, et al., 2011 Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**: 214–219.

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

Schaffer, H. E., D. Yardley, and W. W. Anderson, 1977. Drift or selection: a statistical test of gene frequency variation over generations. *Genetics* **87**: 371–379.

Schierup, M. H. and D. Charlesworth, 2000. The effect of hitch-hiking on genes linked to a balanced polymorphism in a subdivided population. *Genetical Research* **76**: 63–73.

Schwarz, G., 1978 Estimating the dimension of a model. *The annals of statistics* **6**: 461–464.

Serre, D. and S. Pääbo, 2004 Evidence for gradients of human genetic diversity within and among continents. *Genome Research* **14**: 1679–1685.

Sheehan, S., K. Harris, and Y. S. Song, 2013. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* **194**: 647–662.

Slatkin, M., 1977. Gene flow and genetic drift in a species subject to frequent local extinctions. *Theoretical Population Biology* **12**: 253–262.

Slatkin, M., 1980. The distribution of mutant alleles in a subdivided population. *Genetics* **95**: 503–523.

Slatkin, M., 1985. Rare Alleles as Indicators of Gene Flow. *Evolution* **39**: 53–65.

Slatkin, M., 1987. The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32**: 42–49.

Slatkin, M., 1991. Inbreeding coefficients and coalescence times. *Genetical Research* **58**: 167–175.

Smith, G. D., D. A. Lawlor, N. J. Timpson, and J. Baban, 2008. Lactase persistence-related genetic variant: population substructure and health outcomes. *European Journal of Human Genetics* **17**: 357–367.

Sokal, R. R., N. L. Oden, and B. A. Thomson, 1999. A problem with synthetic maps. *Human Biology* **71**: 1–13.

Speliotes, E. K., C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, et al., 2010 Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**: 937–948.

Strobeck, C., 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**: 149–153.

Sul, J. H. and E. Eskin, 2013. Mixed models can correct for population structure for genomic regions under selection. *Nature Reviews Genetics* **14**: 300.

Swallow, D. M., 2003. Genetics of lactase persistence and lactose intolerance. *Annual Review of Genetics* **37**: 197–219.

Takahata, N. and M. Nei, 1984 F_{ST} and G_{ST} statistics in the finite island model. *Genetics* **107**: 501–504.

Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* **26**: 119–164.

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

Teshima, K. and H. Innan, 2009. mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics* **10**: 166.

Tian, C., R. Kosoy, A. Lee, M. Ransom, and J. W. Belmont, 2008. Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PLoS One* **3**: e3862.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, and B. F. Voight, 2006. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature* **39**: 31–40.

Tufto, J., S. Engen, and K. Hindar, 1996. Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* **144**: 1911–1921.

Turchin, M. C., C. W. K. Chiang, C. D. Palmer, S. Sankararaman, D. Reich, et al., 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature* **44**: 1015–1019.

Viterbi, A., 1967 Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13**: 260–269.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006. A map of recent positive selection in the human genome. *PLoS Biology* **4**: e72.

Voight, B. F., L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, et al., 2010 Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* **42**: 579–589.

Wakeley, J., 1998 Segregating sites in Wright's island model. *Theoretical Population Biology* **53**: 166–174.

Wakeley, J., 1999 Nonequilibrium migration in human history. *Genetics* **153**: 1863–1871.

Wakeley, J. and N. Aliacar, 2001. Gene genealogies in a metapopulation. *Genetics* **159**: 893–905.

Wall, J. D., M. A. Yang, F. Jay, S. K. Kim, and E. Y. Durand, 2013. Higher levels of neanderthal ancestry in east asians than in europeans. *Genetics* **194**: 199–209.

Wallace, A., 1864. The origin of human races and the antiquity of man deduced from the theory of "natural selection". *Journal of the Anthropological Society of London* **2**: clviii–clxxxvii.

Wang, C., Z. A. Szpiech, and J. H. Degnan, 2010. Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statistical Applications in Genetics and Molecular Biology* **9**: 13.

Wang, J. L., 2001. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research* **78**: 243–257.

Watterson, G. A., 1982. Testing selection at a single locus. *Biometrics* **38**: 323–331.

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier, 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**: 116.

Weir, B. S. and C. C. Cockerham, 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Weiss, G. and A. von Haeseler, 1998. Inference of population history using a likelihood approach. *Genetics* **149**: 1539–1546.

Weiss, G. H. and M. Kimura, 1965. A mathematical analysis of the stepping stone model of genetic correlation. *Journal of Applied Probability* **2**: 129–149.

Wellcome Trust Case Control Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661–678.

West, B., 1993. *Biston betularia* L. (Lep. Geometridae): continued decline in industrial melanism in northwest Kent. *The Entomologist's Record* **115**: 13–16.

Whitlock, M. C. and D. E. McCauley, 1990. Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. *Evolution* **44**: 1717–1724.

Wickham, H. *ggplot2: elegant graphics for data analysis*. Springer.

Wilkinson-Herbots, H. M., 1998. Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology* **37**: 535–585.

Williamson, E. G. and M. Slatkin, 1999. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* **152**: 755–761.

Wright, S., 1921. Systems of mating I-V. *Genetics* **6**: 11–178.

Wright, S., 1922. Coefficients of inbreeding and relationship. *The American Naturalist* **56**: 330–338.

Wright, S., 1931. Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Wright, S., 1938. Size of population and breeding structure in relation to evolution. *Science* **87**: 430–431.

Wright, S., 1943. Isolation by distance. *Genetics* **28**: 114–138.

Wright, S., 1948. On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution* **2**: 279–294.

Wright, S., 1951. The genetical structure of populations. *Annals of Eugenics* **15**: 323–354.

Wright, S. and T. Dobzhansky, 1946. Genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of Drosophila. *Genetics* **31**: 142–160.

Xu, R. and D. Wunsch, 2005. Survey of clustering algorithms. *Neural Networks* **16**: 645–678.

Yang, W. Y., J. Novembre, E. Eskin, and E. Halperin, 2012. A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics* **44**: 725–731.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, et al., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326–3328.