Methods for demographic inference from single-nucleotide polymorphism data

Colette Mair

A Dissertation Submitted to the
University of Glasgow
for the Degree of
Doctor of Philosophy

School of Mathematics & Statistics

November 2012

Abstract

The distribution of the current human population is the result of many complex historical and prehistorical demographic events that have shaped variation in the human genome. Genomic dissimilarities between individuals from different geographical regions can potentially unveil something of these processes. The greatest differences lie between, and within, African populations and most research suggests the origin of modern humans lies within Africa. However, differing models have been proposed to model the evolutionary processes leading to humans inhabiting most of the world.

This thesis develops a hypothesis test shown to be powerful in distinguishing between two such models. The first ("migration") model assumes the population of interest is divided into subpopulations that exchange migrants at a constant rate arbitrarily far back in the past, whilst the second ("isolation") model assumes that an ancestral population iteratively segregates into subpopulations that evolve independently. Although both models are simplistic, they do capture key aspects of the opposing theories of the history of modern humans.

Given single nucleotide polymorphism (SNP) data from two subpopulations, the method described here tests a global null hypothesis that the data are from an isolation model. The test takes a parametric bootstrap approach, iteratively simulating data under the null hypothesis and computing a set of summary statistics shown to be able to distinguish between the two models. Each summary statistic forms the basis of a statistical hypothesis test where the observed value of the statistic is compared to the simulated values. The

global null hypothesis is accepted if each individual test is accepted. A correction for multiple comparisons is used to control the type I error rate of this compound test.

Extensions to this hypothesis test are given which adapt it to deal with SNP ascertainment and to better handle large genomic data sets. The methods are illustrated on data from the HapMap project using two Kenyan populations and the Japanese and Yoruba populations, after the method has been validated by simulation, where the 'true' model is known.

Acknowledgements

I would like to acknowledge the support and encouragement of my supervisor, Dr. Vincent Macaulay, not only for providing me with the skills and confidence in research but also for introducing me to this area of statistics.

I would like to more generally thank the Department of Statistics for their continued support from the beginning of my undergraduate degree until the completion of this thesis. Thanks also to my office mates, I will miss the banter and wish you all success.

Thanks to my family and friends. My brother for his constant reminders that there is more to life than academia and my sister for reading my thesis. Lastly, to my mum. Your determination over the last few years has shown me anything is possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Wakeley (2009) describes population genetics as the study of allele frequencies in populations. Rather than concentrating on particular individuals, interest lies in how evolutionary forces affect the frequencies over time in populations or in samples. Such forces include mutation, selection, drift and population structure. Typically, data are collected, modelled statistically and population parameters are estimated. Theoretically, genetic variability may be modelled in terms of these forces, but, in many cases, even the most simplistic scenarios are difficult to treat analytically, although it may be possible through computational approaches.

This chapter begins by describing a coalescent approach to modelling genetic data and approaches to parameter estimation. Possible problems with inference are described and an overview of this thesis is provided.

1.1 The coalescent

In order to infer relationships between individuals or, more commonly, populations, methods in population genetics rely on theoretical constructions of their history that capture

biological characteristics. The Wright-Fisher model is one such model that was first introduced by Fisher (1930) and Wright (1931) and described well by Wakeley (2009). This model considers a population of N haploid monoecious organisms and assumes non-overlapping generations and a constant population size over time. Furthermore, at each generation, the population is assumed to be a copy of a random sample of the previous generation with replacement. For a biallelic locus with i copies of one allele, the probability that there are j copies in the following generation is

$$p_{ij} = \binom{N}{j}\left(\frac{i}{N}\right)^j\left(1-\frac{i}{N}\right)^{N-j}, \qquad 0 \le j \le N.$$

Given a sample from a panmictic population in the present time, the coalescent process describes the history of the sample backwards in time until the most common recent ancestor of the sample has been found. At each event in the process, two random lineages fission together decreasing the sample size by one. In a sample of size n from a population of size N, a bifurcating tree can be simulated by $n-1$ coalescent times $T_2, \ldots, T_n$ where T_i is the time during which there are i lineages present in the sample. Figure 1.1 shows an example with $n = 5$. In a series of papers, **??** showed that as $N \to \infty$, the coalescent

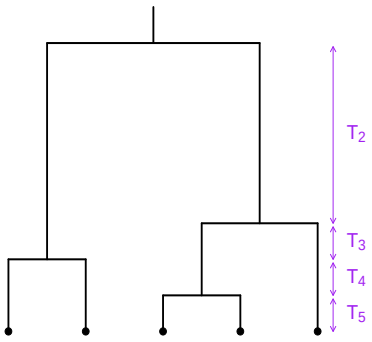


Figure 1.1: Example of coalescent process with $n = 5$.

times are independent and exponentially distributed with rates $\binom{i}{2}$, for $i = 2, \ldots, n$, and scaling time by N, therefore measuring time in N generations, the Wright-Fisher model converges to the coalescent process.

By re-scaling time, the coalescent process also approximates other models. For example, the Moran model, which allows overlapping generations has the coalescent as a limit (scaling time by $N^2/2$ generations). At each time point two individuals are chosen at random with replacement; the first reproduces and the second dies. Many extensions of the coalescent have been studied. For example, Donnelly and Tavaré (1995) allowed for a variable population size over time. Suppose a population evolves according to the Wright-Fisher model but, at each generation, the size of the population changes deterministically. In a generation of size N, the probability that two lineages had their ancestor in the previous generation is $1/N$, and so as N increases this probability decreases. Therefore, the changeable population size affects the rate at which lineages coalesce.

Two properties of a genealogy that are often of interest, since they provide useful summaries, are the time to the most recent common ancestor of the sample $T_{MRCA} = \sum_{i=2}^{n} T_i$ and the total length of the tree $T_{Total} = \sum_{i=2}^{n} iT_i$.

1.1.1 Mutation models

Interest lies in detecting genetic variability between populations and mutation is a source of variability. One common model for mutation is the infinite alleles model, as described by Kimura and Crow (1964). In this model, each mutation generates a new allele, unlike any allele already in existence. On the other hand, the k-alleles model restricts the number of distinct alleles in the population to k. Each allele can mutate to the remaining $k - 1$ alleles with equal probability. The stepwise mutation model was defined by Kimura and Ohta (1978), mutation occurs at rate μ per site per generation and when a mutation occurs, the allele changes state by moving either one step forward or one step backwards with probability 1/2. In addition, Kimura (1969) describes the infinite sites mutation

model with each new mutation occurring at a homozygous site. This model is appropriate when there is a large number of sites and a small mutation rate and so fits well modelling mutations in DNA. In the human genome, SNPs have an estimated mutation rate of 2.7×10^{-8} per site per generation as estimated by Nachman and Crowell (2000). This is low enough to reasonably assume only a single mutation occurred in the ancestry of the SNP.

The number of mutations to occur at a single site is a Poisson random variable with rate per N generations equal to $\frac{1}{2}\theta$ where θ is historically defined as $2N\mu$ with μ the mutation rate per generation and N the diploid population size as described by Wakeley (2009).

1.1.2 Test for neutrality via summary statistics

The Wright-Fisher model describes a sample from a random mating population that is unaffected by natural selection or population structure and it assumes that the population size remains constant through time. That is, it makes many unrealistic assumptions. Any deviations from these assumptions will impact analytical results assuming this so-called standard neutral model. In order to test the standard neutral model, many statistics have been invented to compare theoretical results under a neutral model to observed data. Tajima (1989) used sequence data, an example of which is given below:

$$\begin{aligned} &\text{sequence 1} && \ldots ATGGGCA \ldots \\ &\text{sequence 2} && \ldots ACGGACA \ldots \\ &\text{sequence 3} && \ldots GCGGCTA \ldots \\ &\text{sequence 4} && \ldots GTAGACA \ldots \end{aligned}$$

to compare the average number of pairwise differences between the sequences in the sample, π, and the number of segregating sites, S, under the infinite sites model. In the example, $S = 5$. The number of differences between sequence 1 and sequence 2 is 2, the number of

differences between sequence 1 and sequence 3 is 4. Continuing in the way,

$$\begin{aligned}\pi &= \frac{2+4+3+3+3+4}{\binom{4}{2}} \\ &= 3.17.\end{aligned}$$

Under the neutral model, Watterson (1975) found the distribution of S but in particular showed $E\left(S/\sum_{i=1}^{n-1} i\right) = \theta$. Also, Tajima (1983) showed $E(\pi) = \theta$ and hence a large deviation of $S/\sum_{i=1}^{n-1} i - \pi$ from 0 would indicate a violation of at least one of these assumptions. Tajima approximated the distribution of the statistic

$$D = \frac{\pi - \frac{S}{\sum_{i=1}^{n-1} i}}{\sqrt{\widehat{Var}\left(\pi - \frac{S}{\sum_{i=1}^{n-1} i}\right)}},$$

under the standard neutral model as a Beta distribution and used this result to find critical regions for this test. Interpretations of rejections of the model can be ambiguous, the observed statistic will either be too large in a positive or negative direction. It is clear that in either case, the neutral model can be rejected but it is not clear which assumption has been violated as demonstrated by Simonsen et al. (1995). These authors simulated data under several demographic scenarios and computed Tajima's statistic. The aim of the paper was to test the power of Tajima's statistics, amongst others, to reject the standard neutral model. Notably, the value of the statistics was not informative about the type of model the data were derived from, only that it was inconsistent with the neutral model.

1.1.3 Population structure

In this thesis, two demographic models, described by Wright (1969) and Cavalli-Sforza and Bodmer (1971), are considered. The first model considered assumes the population is divided into subpopulations that exchange migrants at a constant rate arbitrarily far back

in the past. The second model assumes that an ancestral population iteratively segregates into two subpopulations that evolve independently. Both models are depicted in figure 1.2 with 5 subpopulations labelled 1 through to 5.

The two models are somewhat simplistic: the human population is most likely the result of many complex demographic events that have led to complicated patterns of genetic variation. However, one of the models may be better than the other at capturing patterns of variability. In any demographic model, there are unknown parameters that are of importance, for instance, migration rates in the first model and population divergence times in the second; populations sizes would be common to both classes of model. Given observed

(a) (b)

Figure 1.2: Examples of (a) migration and (b) isolation models with 5 subpopulations.

SNP data from a set of subpopulations, Wright (1969) introduced F_{st} which measures the amount of variation within subpopulation compared to the amount of variation between subpopulations and this can be indicative of populations structure. Also, Patterson et al. (2006) used principal components analysis to firstly reduce the dimension of the data and secondly showed that, given the first few components are significant in capturing the structure, they can cluster the data such that those within the same subpopulation are clustered together. Methods of detecting population structure are presented in chapter 2.

1.2 Estimating ancestral parameters

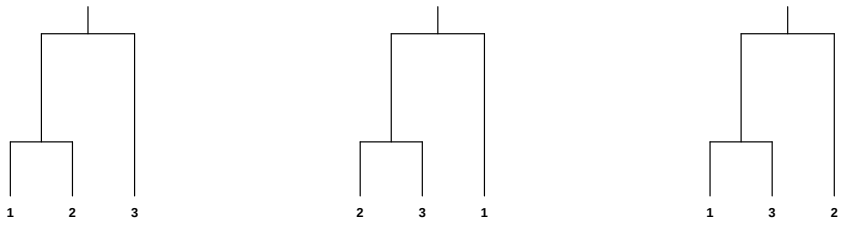
Inference about demographic history relies on past events impacting present-day populations in a distinct way so that genetic data may be informative about the evolutionary history. It is often of interest to estimate a population divergence time or migration rates for a given data set and many methods adopt either a maximum likelihood or Bayesian framework. However, Wakeley and Hey (1997) use a method of moments approach to estimate the population divergence time in an isolation model with two subpopulations by considering different classes of segregating sites using sequence data. They partitioned the segregating sites into four exclusive groups. Two of the groups comprise the sites that are only variable in one of the two subpopulations. Another class consist of sites that are segregating the same two bases pairs in both subpopulations and the last class consists of sites that have a fixed difference between the two subpopulations, i.e., those that partition the sample into n_1 and n_2 sequences, where n_1 and n_2 denote the sample sizes of subpopulation one and two respectively. Wakeley and Hey derived expected values of the number of sites in each class under a neutral model as a function of the model parameters and estimated population parameters by equating the observed number to the expected number.

Model-based methods in inference are based around calculating the likelihood function, $p(D|\phi)$, where D denotes the data and ϕ the parameters in a particular model and then either maximizing the likelihood function with respect to the components of ϕ or placing a prior distribution $\pi(\phi)$ on ϕ and simulating from the posterior distribution $p(\phi|D) \propto \pi(\phi)p(D|\phi)$.

However, the genealogy (the ancestral tree) of the sample is unobserved and is a high dimensional nuisance parameter, g. Therefore, integration over all possible genealogies is required in order to calculate the likelihood function. More precisely,

$$p(D|\phi) = \int_{g \in G} p(D|g, \phi)p(g|\phi)dg, \tag{1.1}$$

where G is the set of all genealogies. A genealogy can be defined in terms of a topology and a set of coalescent times. Disregarding the coalescent times, Wakeley (2009) discusses the total number of possible topologies in a sample of size n. Consider the case $n = 3$, then there are 3 possible rooted bifurcating branching structures:



Generally, beginning at the present time, there are $\binom{n}{2}$ possible coalescent events. Once two lineages have coalesced, the number of possible coalescent events is $\binom{n-1}{2}$ and so forth, until only two lineages are present and $\binom{2}{2} = 1$. The total number of possible genealogies is

$$\prod_{i=2}^{n} \binom{i}{2}.$$

Figure 1.3 illustrates how quickly this number increases. For example, for $n = 6$ there are already 2700 different possible genealogies. Including coalescent times, the set G is uncountably infinite. For this reason, many attempts have been made to either estimate the likelihood function or make inference without the use of the likelihood function.

As in equation (1.1), several statistical problems involve evaluating integrals of the form

$$E\Big(f(X)\Big) \;\; = \;\; \int_{x\in\mathcal{X}} f(x)g(x)dx$$

for a random variable X with distribution function $g(x)$ and some function $f(x)$. By

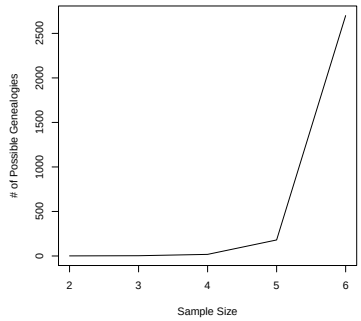


Figure 1.3: Number of branching structures in an samples of size $n = 2, 3, 4, 5$ and 6.

sampling $\{x_1, \ldots, x_m\}$ from $g(x)$ then

$$E\Big(f(X)\Big) \approx \frac{1}{m}\sum_{i=1}^{m} f(x_i) \tag{1.2}$$

is a Monte-Carlo approximation of the integral as described by Gelman et al. (2004). Stephens (2007) demonstrates how Monte-Carlo methods can be used for ancestral inference, for example to estimate T_{MRCA}. By simulating draws, $T_1, \ldots, T_m$, from $p(G|D)$ and considering the time of the most recent common ancestor of the ith genealogy $T_{MRCA}(T_i)$, Stephens makes the approximation

$$E(T_{MRCA}|D) \approx \frac{1}{m}\sum_{i=1}^{m} T_{MRCA}(T_i).$$

Therefore, instead of considering the full set of genealogies, a random sample of genealogies from $p(G|D)$ is taken. Due to the complexities of $g(x)$ it may be infeasible to simulate from it. Importance sampling is a technique that employs another probability function $q(x)$ similar to $g(x)$ but easier to sample from. Re-write $E\Big(f(X)\Big)$ as

$$E\Big(f(X)\Big) = \int f(x)\frac{g(x)}{q(x)}q(x)dx$$

which is approximated using Monte-Carlo:

$$E\Big(f(X)\Big) \approx \frac{1}{m}\sum_{i=1}^{m}\frac{g(\tilde{x}_i)}{q(\tilde{x}_i)}f(\tilde{x}_i),$$

where $\{\tilde{x}_1,\ldots,\tilde{x}_m\}$ are samples from $q(x)$.

Further approximate methods (Approximate Bayesian Computation) in population genetics have been developed that use summary statistics for parameter estimation as described by Tavaré et al. (1997) and Beaumont et al. (2002). Such methods sidestep computing the full likelihood function and rely on suitable statistics η that may be used to replace $p(D|\phi)$ by $p(\eta(D)|\phi)$ for parameters ϕ. These methods are often based on a Monte-Carlo scheme, where parameter values are accepted or rejected depending on how closely summaries derived from them match the observed values of the summaries.

1.2.1 Model selection

Given a data set x and two contending models M_1 and M_2, there are various ways to test which model best fits using a frequentist or Bayesian approach with both methods equally problematic.

A Bayesian hypothesis test computes the marginal likelihood of the data under the models, a challenge in itself when dealing with high dimensional genetic data and their unknown ancestry. The model comparison procedure considers the ratio

$$\begin{aligned}\frac{p(M_1|x)}{p(M_2|x)} &= \frac{\pi(M_1)p(x|M_1)}{\pi(M_2)p(x|M_2)}\\ &= \frac{\pi(M_1)}{\pi(M_2)}\times B_{12},\end{aligned}$$

where B_{12} is the Bayes factor comparing M_1 and M_2. $B_{12} > 1$ shows that the data provide evidence to support M_1 and $B_{12} < 1$ increases the supports for M_2.

A frequentist hypothesis test takes the form

$$\begin{aligned} H_0 &: \text{ data from } M_1 \\ H_1 &: \text{ data from } M_2 \end{aligned}$$

with H_0 the null hypothesis and H_1 the alternative hypothesis. A test statistic is used to assess how compatible the data are with H_0, which is rejected if the measured value of the statistic is unlikely under H_0. If it is possible to estimate adequately the likelihood, then the likelihood ratio often has good properties for model selection as used, for example, by Nielsen and Wakeley (2001). In addition, for a specific model with maximum likelihood value L and p the number of parameters in the model, the Akaike information criterion (AIC) is defined to be

$$\begin{aligned} \text{AIC} &= -2\ln L + 2p \\ &= D + 2p, \end{aligned}$$

and can be computed for several potential models. The model that produces the lowest AIC is deemed the most suitable (from the set of proposed models) since the value D decreases as the likelihood L increases and more complex models, that is models with more parameters, are penalised. Similarly, the Bayesian information criterion (BIC), in which more complex models have a heavier penalty compared with AIC, is defined by Congdon (2003) to be

$$\text{BIC} = D + p\log(n),$$

in a sample of size n.

1.3 Difficulties in inference

There are many difficulties in modelling population genetic data, some of which are highlighted in this section.

1.3.1 Ascertainment

Throughout this thesis, only single nucleotide polymorphism (SNP) data are considered. Many large-scale SNP data sets have been invaluable to human population genetic studies. However, theoretical results may not be directly applicable to SNP datasets because of the non-random way SNPs have been identified or ascertained. SNP discovery tends to begin by genotyping a small sample of individuals at a particular locus and, if there is variability in this small sample, a larger sample is then genotyped. This type of procedure introduces a bias towards SNPs with intermediate allele frequencies which can lead to unreliable results when inferring demographic history using theoretical results. Different ascertainment schemes have been analysed, for example, by Albrechtsen et al. (2010), Nielsen (2004), Nielsen and Signorovitch (2003) and Nielsen et al. (2004).

Wang et al. (1998) discovered the location of 2227 SNPs in the human genome. In order to identify SNPs, the authors began by selecting over 1000 regions across the genome and, for each region, they genotyped an initial set of three individuals and then a pool of ten individuals. If a position was variable in the initial set, they then classified this position as a candidate SNP, and, if the allele frequency in the pool of individuals at the candidate SNP was above a pre-determined threshold, then the SNP was included in the sample, i.e. was considered ascertained.

Due to the bias towards positions with intermediate allele frequencies, ascertainment has been shown to have a dramatic effect on the allele frequency spectrum, for example by Nielsen et al. (2004). Allele frequency spectra are helpful in expressing the patterns of variation in SNP data as many genetic factors can affect the shape of the spectrum.

They are constructed by counting the number of SNPs with each possible number of mutant alleles in the sample. For example, figure 1.4 shows an example of haploid size five genotyped at six SNPs with the data displayed in the matrix on the left hand side.

$$\begin{array}{c} \\ \text{SNP} \end{array} \begin{array}{c} \text{individual} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

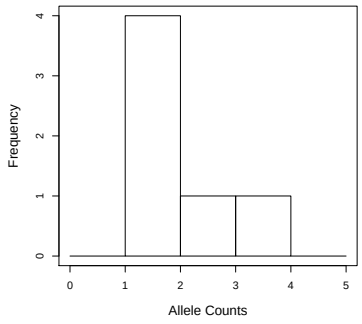


Figure 1.4: Example of allele frequency spectrum (right) for SNP data (left) for five haploids at six SNPs.

Assuming that only a single mutation has occurred, each entry of the matrix is in $\{0, 1\}$ with zero corresponding to those who carry the first allele and a one corresponding to those who carry a copy of the other allele. At each SNP, it may not be known which allele is ancestral and which was caused by the mutation. If the ancestral allele is known, the mutation divides the sample, of haploid size n, into i mutant alleles and $n - i$ ancestral alleles. Across SNPs, Wakeley (2009) lets ξ_i denote the number of SNPs that divide the sample in this way for $1 \leq i \leq n - 1$. If it is unknown which allele is mutant and which is ancestral, then it is not possible to distinguish SNPs that divide the sample into i mutant ($n - i$ ancestral) and $n - i$ mutant (i ancestral) alleles. Therefore, it is only possible to count the number of SNPs that divide the sample into i copies of one allele and $n - i$ copies of the other for $1 \leq i \leq [n/2]$. The distribution of the row sums of the matrix is the bases of the histogram on the right hand side of figure 1.4 (the allele frequency spectrum).

Under a neutral model, Nielsen and Signorovitch (2003) derived expressions for the sampling distribution of allele frequencies under different ascertainment schemes. Namely when the ascertainment panel is included in the final data set, when only some of the panel is included in the final data set and when the final data set is exclusive of the panel. When

the final sample includes the ascertainment panel, figure 1.5 illustrates how the allele frequency spectrum is affected. As described, there is a bias towards intermediate valued allele counts.

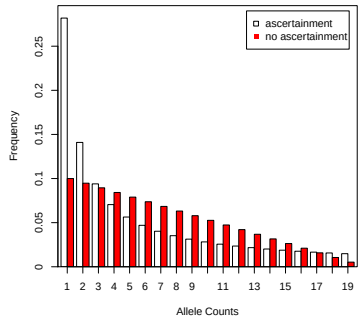


Figure 1.5: Example of allele frequency spectrum under a neutral model with sample size $n = 20$ without ascertainment and with ascertainment of panel size 2.

1.3.2 Ghost populations

In estimating population parameters such as migration rates, it is assumed that samples are available from all of the populations in existence rather than from a sample of the populations. Beerli (2004) described ghost populations as the set of populations in existence but unsampled from and examined their effects on parameter estimation. In particular, he simulated data under several migration scenarios with three subpopulations and took samples from only two. The study showed that if the ghost population had a high migration rate with the two sampled populations, this leads to poorer estimates of the migration rates between the two sampled populations, whereas little or no migration between the ghost and sampled populations leads to little bias in migration rate estimates. Therefore, it will be assumed throughout that samples are available from all the populations in existence. This is not an issue in simulated data but may impact inference using real data.

1.3.3 Sufficient statistics

Estimating a parameter ϕ given observed data x using a function $T(x)$, ideally T would contain all the information, from the data, about ϕ. A statistic T is a sufficient statistic for a parameter ϕ if the conditional distribution of the data given $T(x) = t$ is independent of ϕ.

Given observed data x, The Neyman–Fisher theorem provides an equivalent condition for sufficiency. It states that a statistic T is sufficient for ϕ if and only if there exist functions $a(x)$ independent of ϕ and $b(t|\phi)$ such that

$$p(x|\phi) \quad = \quad b(t|\phi)a(x).$$

For example, given data $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$ then an estimator of μ is $T(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ which is sufficient for μ since

$$\begin{aligned} p(x_1, \ldots, x_n|\mu) \quad &\propto \quad \prod_{i=1}^{n} \exp\left(- \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ &= \quad \exp\left(- \frac{(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2)}{2\sigma^2} \right) \\ &= \quad \exp\left(- \frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2} \right) \exp\left(- \frac{-2\mu \sum_{i=1}^{n} x_i + n\mu^2)}{2\sigma^2} \right). \end{aligned}$$

One recurring problem in population genetics is that it is often the case that estimators of population parameters are not sufficient. A few examples are presented below.

Joyce (1998) described Ewens' sampling formula. Given n objects divided into k distinct objects or classes with a_i the number of classes with i representatives, then the distribution of $a_1, \ldots, a_n$ is

$$P\{a_1, \ldots, a_n\} \quad = \quad \frac{n!\theta^k}{\theta_{(n)}} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!},$$

where $\theta_{(n)} = \theta(\theta+1)\dots(\theta+n-1)$. This scenario is analogous to the infinite alleles mutation model, where each mutation introduces a unique allele with n the total number of genes, k the total number of alleles in the sample, a_i is the number of alleles which have i copies in the sample and, in population genetics, θ is the scaled mutation rate. Wakeley (2009), amongst others, shows k is sufficient for θ by showing the conditional distribution of $a_1, \dots, a_n$ given k is independent of θ.

A possible estimator of a population divergence time is

$$\hat{T} = -\log(1-\hat{F}_{st}),$$

derived by Cavalli-Sforza and Bodmer (1971). Nielsen et al. (1998) derived an expression for the joint likelihood of T and F, where F is the ancestral frequencies which may be estimated by the observed allele frequencies. Using this derived expression, values of the log likelihood of T may be computed. The authors investigated whether this estimator of T was a sufficient estimator and provided an example of two data set with the same estimated F_{st} and F but had a different log likelihood value for T. Therefore, suggesting that estimators that are functions of F_{st} are not a sufficient.

One parameter of interest is the scaled mutation rate θ. Tajima (1983) showed that

$$\begin{aligned} E(\pi) &= \theta, \\ var(\pi) &= \frac{(n+1)\theta}{3(n-1)} + \frac{2(n^2+n+3)\theta^2}{9n(n-1)}. \end{aligned}$$

Therefore, Tajima showed that π is an unbiased estimator of θ but it is not consistent since $var(\pi)$ does not approach zero and $n \to \infty$. In particular, Tajima examined the relationship between the sample size and the variance of $\hat{\pi}$ and showed when $\theta = 0.1$, the $var(\pi)$ using 10 sequences was approximately equal to the variance using 200 sequences. This estimator is not a consistent estimator of θ, Joyce and Marjoram (2008) examined estimating the scaled mutation rate for a range of different estimators and suggest there are no (known) sufficient statistics for θ.

1.4 SNP data sets

Population structure patterns can emerge by analysing patterns of variation in the human genome within and between populations. The International HapMap Consortium (2003) describe the amount of variation in the human genome, in particular, it estimated that two human genomes differ at approximately 0.1% of sites with most sites biallelic. The consortium convey that in around 90% of variable sites, equating to around ten million SNPs in the human population, both alleles have a frequency greater that 0.01 whereas in the remaining 10%, one allele exhibits a frequency less that 0.01.

There are several different types of genetic data set. Section 1.1.2 shows an example of DNA sequences data where positions in the sequences may be mutated producing SNPs. This thesis will focus on this type of data. Some regions of the genome are highly polymorphic and short segments of DNA are found at high frequency. Microsatellites are short repeats of DNA sequences.

Over the last decade, many genetic data sets have been compiled for example the 1000 genome project by the 1000 Genomes Project Consortium (2010), the HapMap project by the International HapMap Consortium (2003) and the human genome diversity panel by Cann et al. (2002).

The human genome diversity panel stored blood samples from around 1050 people from 52 countries in the Foundation Jean Dausset-CEPH in Paris with the intention of building a genetic data base by distributing the samples to be genotyped by different investigators and the results stored in a central database (http://www.cephb.fr/en/index.php). Currently, the database contains around 660918 SNPs.

The international HapMap project was originally designed to discover patterns in DNA sequences in the human genome to identify genes associated with diseases as outlined by the International HapMap Consortium (2003). Over a seven year period, this project was delivered in three phases. The first phase consisted of samples with ancestry from parts

of Africa, Asia and Europe genotyping a total of 269 individuals. In the second phase, the International HapMap Consortium (2007) improved the coverage of the genome by genotyping the 269 samples at 3.1 million SNPs. In the third phase, the International HapMap 3 Consortium (2010) genotyped samples from an addition 7 populations bringing the total sample size to 1184.

The 1000 Genomes Project Consortium (2010) used samples from the third phase of the HapMap project to compare sequencing strategies. The pilot projected compared three strategies and their rate of discovery of variant positions in the genome. The consortium used whole genome shotgun sequencing on unrelated individuals and on two trios, one from the Yoruba population and the other from the CEU sample, consisting of individuals with European ancestry. The last method targeted exons, short and functionally important sequences that are thought to contain diseases causing mutations in the human genome. For example Li et al. (2010) created a database of genetic data by targeting exons in 200 Danish individuals.

1.5 Overview of thesis

The remainder of this thesis is organised into eight chapters. Population structure and methods of detecting any structure are explored in chapter 2. Chapter 3 presents the details of simulating SNP data under the isolation and migration models introduced in section 1.1.3 and also possible parameter estimation of the model parameters. Data are simulated and examined under both models. Chapter 4 considers established ways of distinguishing the two models and examines the effects of ascertainment on SNP data. Further methods of parameter estimation are examined in chapter 5. Chapter 6 introduces a summary statistic–based hypothesis test that may be used to distinguish between the migration and isolation models. Details are given of the choice of summary statistics and methods of treating multiple comparisons. The hypothesis test examines how distinguishable the two models are under certain SNP ascertainment schemes and also looks at

the usefulness of the significant components from principal components analysis in chapter 7. This methods is applied to real data in chapter 8. Lastly, a discussion of the methods employed in the thesis is presented with possible extensions and criticisms.

Chapter 2

Population structure

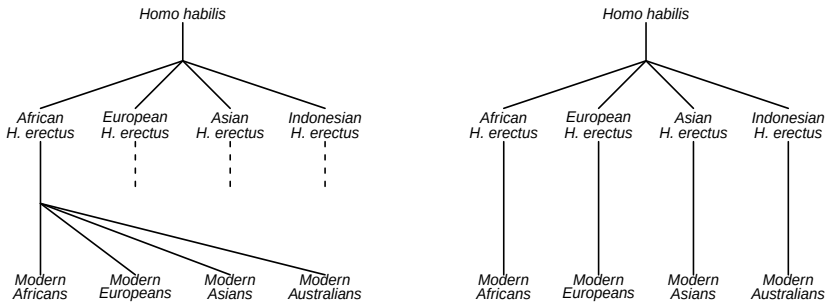
Cann et al. (1987) analysed mitochondrial DNA from 147 people from five geographical regions. They reconstructed a genealogy using the parsimony method.[1] The authors found that, from the reconstructed genealogy, it is likely that human mitochondrial expanded quite recently from Africa. Since then, the out-of-Africa theory has received support from many different types of data, as summarised by Nei (1995). However, reconstructing evolutionary history has led to much controversy. Two main theories that are at the opposite ends of a spectrum of hybrid models, as discussed by Relethford (2008), are

1. the replacement theory and,
2. the theory of multiregional evolution.

The first posits that modern humans appeared in Africa around 200,000 years ago and spread through Asia, Australasia and Europe after 100,000 years ago, replacing archaic humans who already existed there. The second assumes that humans diverged from Africa around two million years ago through Asia, Australasia and Europe and each settlement

[1] For each pair of sequences (x, y) out of a total of n sequences, the distance between the pair is the number of difference between x and y, denoted by d_{xy}. The parsimony score of a genealogy is $\sum d_{xy}$, summed over all possible pairs. The maximum parsimony tree minimise the parsimony score.

evolved into modern humans. Migration events occurred over time such that the separate settlements remained genetically similar instead of evolving into separate species. Figure 2.1 shows both theories, as displayed by Cavalli-Sforza and Piazza (1993).



(a) (b)

Figure 2.1: Reconstruction of the origins of humans as shown by Cavalli-Sforza and Piazza (1993).

Although the two theories present different explanations of demographic history, it is clear that the human population is structured in some way. This chapter aims to review some methods that have been employed to detect population structure, in particular using SNP data.

2.1 Inferring population structure

The most basic coalescent process assumes a sample from one random–mating population. One deviation from this assumption is to assume that the larger population can be divided into smaller subpopulations that are more or less genetically distinct. Many current methods in inferring population structure aim to quantify the amount of gene flow between two or more populations. A limitation of these methods is the confounding of long–term and historical events, namely recurrent gene flow with a small number of historical population splits. The two different demographic scenarios can present similar characteristics in allele

frequencies and other summary statistics, which can make differentiating between the two difficult.

2.1.1 F-statistic

A commonly quoted statistic in population genetics is Wright's F_{st} statistics, defined vaguely to be "the correlation between random gametes within subdivisions, relative to gametes of the total population" by Wright (1969). This statistic is a measure of population diversity since it compares the amount of genetic variability between and within populations and is closely related to the so called inbreeding coefficient. It has since been redefined in several ways and many different estimators of it have been proposed. Wright (1969) showed that

$$F_{st} = \frac{F_{it} - F_{is}}{1 - F_{is}},$$

where F_{is} and F_{it} were defined to be the "correlation between uniting gametes relative to those of their own subdivision" and "correlation between gametes that unite to produce the individuals relative to the gametes of the total population", respectively. Cavalli-Sforza and Bodmer (1971) discussed the case of k subpopulations in Hardy–Weinberg equilibrium so that the amount of genetic variation in each subpopulation remains constant over time. The authors showed, under population structure, there is a deficiency in heterozygotes in the entire population by deriving an expression for the average frequency of heterozygotes. For a biallelic locus, let $\bar{p}$ be the average sample allele frequency and σ^2 be the variance of the sample allele frequencies across subpopulations. Then,

$$\hat{F}_{st} = \frac{\sigma^2}{\bar{p}(1-\bar{p})}. \tag{2.1}$$

Furthermore, Cavalli-Sforza and Piazza (1993) also provided an estimator of F_{st} by considering the heterozygosity in the whole population compared to that within populations. At a single loci, they define p_{ij} to be the gene frequency, in the sample, of allele i in

subpopulation j for $i = 1, \ldots, L$ and $j = 1, \ldots, S$.

$$\hat{F}_{st} = \frac{h - h_s}{h},$$

where h is the heterozygosity of the sample,

$$h = 1 - \sum_{i=1}^{L} \left[\frac{1}{s} \sum_{j=1}^{S} p_{ij} \right]^2,$$

and h_s is the average heterozygosity in subpopulation s,

$$h_s = \frac{1}{s} \sum_{j=1}^{s} \left(1 - \sum_{i=1}^{L} p_{ij}^2 \right).$$

Calculating F_{st} under the migration model, Wright (1969) assumed that a population consists of D random–mating islands or subpopulations. Each subpopulation is of the same constant haploid population size N and a proportion of each subpopulation is accounted for by migrants from other subpopulations. He defined the set of parameters $\{m_{ij} : i, j = 1, \ldots, D\}$ where m_{ij} is the probability that a lineage now in subpopulation i had its parent in subpopulation j. Therefore, the number of individuals that migration from island i to j in a single generation (backwards in time) is $N_i m_{ij}$. If the migration rate is constant, i.e. $m_{ij} = m \; \forall i \neq j \in \{1, \ldots, D\}$, and reproduction occurs according to the Wright-Fisher model, Wright derived an estimate F of F_{st} based on the migration rate:

$$F = \frac{1}{1 + 2Nm}. \tag{2.2}$$

This formula was first introduced by Wright (1969) and produced a way of estimating the amount of population diversity and also provided an estimate of the level of gene flow between subpopulations

$$2Nm \approx \frac{1}{F} - 1.$$

Although since reality often violates the assumptions in this model, Whitlock and Mc-

Cauley (1998) demonstrated that it is often the case that F_{st} cannot be used to estimate $2Nm$. This formula assumes that each subpopulations is of the same population size and each pair of subpopulations exchange migrants at the same constant rate. Whitlock and McCauley systematically illustrate the issues with these assumptions, for example, geographically closer subpopulations might exchange migrants at a higher rate than those further apart. Even if these assumptions are valid, other factors, such as selection, may affect allele frequencies within each subpopulation and hence F_{st}. As a result, if N and m remain fixed, it is still possible to estimate different F_{st} values without accounting for other possible factors.

Cavalli-Sforza (1969) considered expected values of F_{st} under different models. In particular, in the isolation model with k subpopulations all diverging from a common ancestral population t generations ago, a relationship between $E[F_{st}]$ and the population divergence time t was derived;

$$\begin{aligned} E[F_{st}] &= 1 - \left(1 - \frac{1}{N}\right)^t \\ E[F_{st}] &\approx 1 - e^{-\frac{t}{N}}, \end{aligned} \tag{2.3}$$

with N is the haploid population size of each subpopulation. This formula is examined more closely in chapter 5.

2.1.2 Genetic distance

Genetic distance can be used to convey the difference between two populations as discussed by Weir (1996). Any distance measure d should satisfy four axioms:

For three points A, B and C,

1. $d(A, B) \geq 0$,
2. $d(A, A) = 0$,

3. $d(A, B) = d(B, A)$ and,

4. the triangle inequality $d(A, B) + d(B, C) \geq d(A, C)$.

In this context, points A, B and C can correspond to individuals or populations.

Weir (1996) provides some examples of possible distance measures between populations. The Euclidean distance between sample allele frequencies may be used. For example, at a particular biallelic locus, if p_1 and p_2 are the frequencies of one of the alleles in the sample from subpopulation 1 and subpopulation 2 respectively, then

$$d(\text{subpopulation 1, subpopulation 2}) \quad = \quad \sqrt{2(p_1 - p_2)^2}.$$

F_{st} between two populations can be considered as a distance measure. Genetically similar populations have smaller pairwise F_{st} values and genetically diverse subpopulations have larger pairwise F_{st}. Since F_{st} is in the range $[0, 1]$, whereas a distance measure should be in the range $[0, \infty)$, a F_{st} based distance measure can be defined by

$$D \quad = \quad -\log(1 - F_{st}). \tag{2.4}$$

Consider data from the HGDP-CEPH diversity panel and exploiting equation 2.4 as a distance measure, a neighbour–joining tree can be constructed as illustrated in figure 2.2. The subset of populations (from the 52 in the panel) are labelled from 1 to 27 and the colour of each label is determined by the continent of the corresponding population. The neighbour–joining algorithm, introduced by Saitou and Nei (1987) and modified by Studier and Keppler (1988), attempts to reconstruct a tree which matches the observed pairwise distances as closely as possible by iteratively joining the nodes that minimise

$$S_{ij} \quad = \quad (N - 2)d_{ij} - R_i - R_j,$$

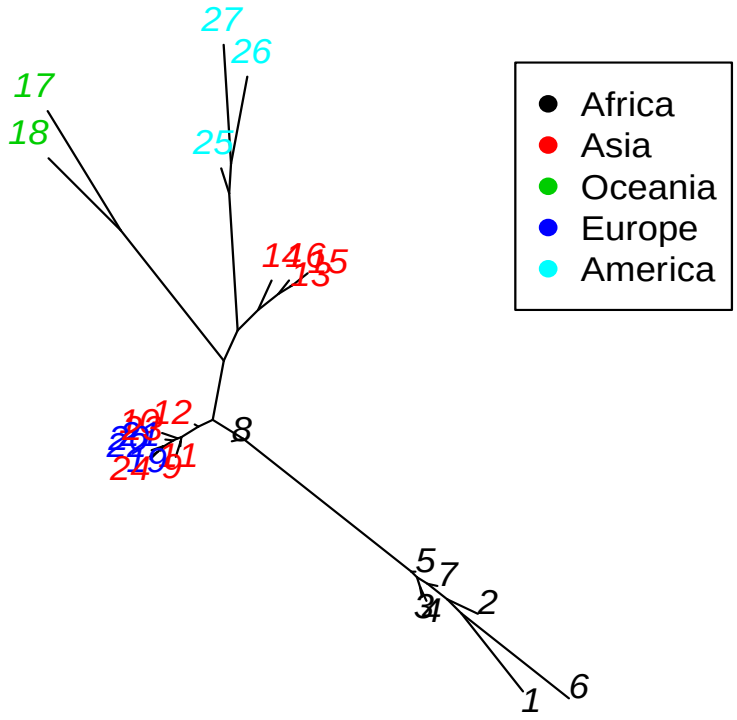


Figure 2.2: Neighbor joining tree of populations from HGDP-CEPH data

where N is the total number of nodes (or populations), $R_i = \sum_{j=1}^{N} d_{ij}$ and d_{ij} is the distance between node i and node j.

2.1.3 Principal components analysis

More that 30 years ago, Cavalli-Sforza and Piazza (1978) studied the relationship between genetic differences and evolutionary history. To represent the evolutionary history of a large area, they applied principal components analysis to gene frequency data and constructed "synthetic maps" based on the components' values. Higher component values were expected to correspond to the place of origin of the sample and the authors hypothesised that patterns of migration would emerge as gradients in the maps. One famous application was to migration events of farmers from the Middle East to Europe by Cavalli-Sforza et al. (1993). However, by simulating data from a set of populations arranged in a square lattice and allowing for migration between neighbouring populations, Novembre and Stephens (2008) illustrated that principal components under this model demonstrated a sinusoidal patterns and hence, peaks and troughs in a principal component map do not necessarily show underlying migration patterns. Although principal components may not show patterns of migration, the use of such dimension reduction techniques has become widespread in the analysis of population structure. In particular, it provides a way of graphically displaying vast amounts of SNP data. In a structured population, such graphical summaries are correlated to geographical distance. For example, using SNP data from Europeans, Novembre (2008) produced a plot of individuals' SNP profiles' projected onto the first two principal components and demonstrated how it resembled a map of the individuals' location in Europe.

Principal components analysis is a common statistical tool that reduces the dimension of high dimensional data through linear combinations of the original variables. Consider a data set that consists of relatively few individuals compared to the number of independent variables p. Let $Y = (y_1, \ldots, y_p)^T$ be the variable measurements for a single individual. The objective is to find uncorrelated 'synthetic' variables, denoted by $PC_1, \ldots, PC_p$, which are linear combinations of the original p variables ordered such that PC_1 accounts for most of the variance in the data and PC_p accounts for the least amount of the variance. Each

component, PC_i $(i = 1, \ldots, p)$, is of the form

$$PC_i = \sum_{j=1}^{p} a_{ij} y_j.$$

The vector $a_i = (a_{i1}, \ldots, a_{ip})$ is equal to the eigenvector of the corresponding ith largest eigenvalue of the covariance matrix $X = YY^T$ as shown by, for example Dunteman (1989).

A test for the presence of population structure was present by Patterson et al. (2006). This method supposes the data consist of L SNPs and n individuals where $L \gg n$. The $n \times L$ matrices C and $\tilde{M}$ were defined with elements

$$C_{i,j} = \text{the number of variant alleles for SNP } j \text{ in individual } i,$$

$$\tilde{M}_{i,j} = \frac{C(i,j) - \mu_j}{\sqrt{p_j(1-p_j)}},$$

where μ_j is the average allele count and $p_j = \frac{\mu_j}{2}$ the allele frequency of the jth SNP. Principal components analysis is performed on the $n \times n$ covariance matrix $\tilde{M}\tilde{M}^T$ and any evidence of population structure is tested based on results of the distribution of the largest eigenvalue of a covariance matrix given by Johnstone (2001). Given the ordered eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_{n^T-1}$ of covariance matrix $\tilde{M}\tilde{M}^T$, the Patterson et al. (2006) set parameters

$$\begin{aligned} \mu(n,L) &= \frac{(\sqrt{L-1}+\sqrt{n})^2}{L}, \\ \sigma(n,L) &= \frac{(\sqrt{L-1}+\sqrt{n})}{L}\left(\frac{1}{\sqrt{L-1}} + \frac{1}{\sqrt{n}}\right)^{\frac{1}{3}}, \end{aligned}$$

and scale the eigenvalues by

$$x = \frac{\lambda_1 - \mu(n-L)}{\sigma(n,L)}.$$

Under the null model of no population structure, the statistic x approximately follows

a Tracy-Widom distribution. The density function is given in figure 2.3. Testing at significance level 0.05, eigenvalues are rejected if the corresponding value of x falls in the red tail of the distribution . If λ_1 is found to be significant, the next eigenvalue is tested by

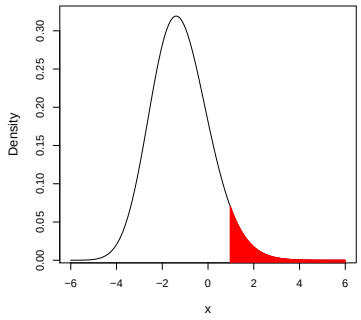


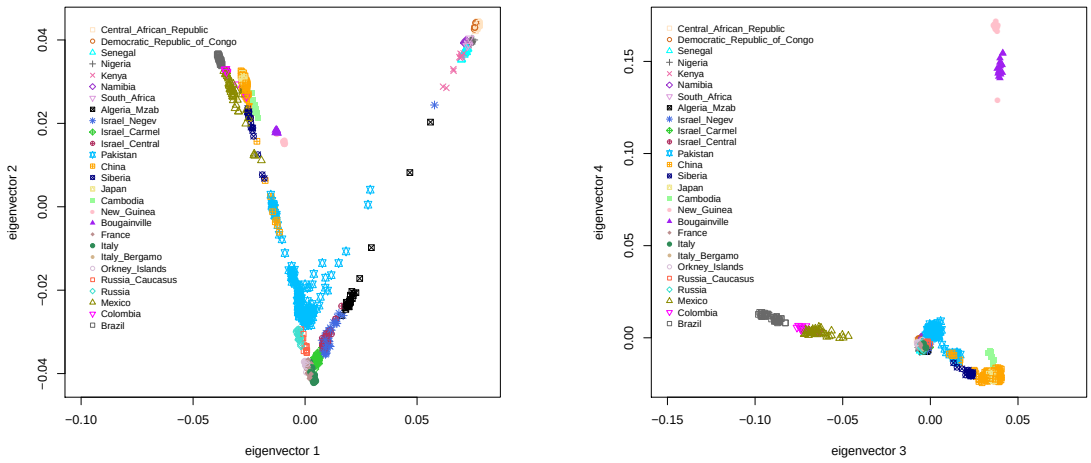
Figure 2.3: Tracy-Widom density.

recalculating x with $n = n-1$ continuing until an eigenvalue is found to not be significant.

Astle and Balding (2009) provide an interpretation of components from principal components analysis. Concentrating on the first component, they note that it will be "correlated with many SNPs". In an example with two populations that exchange migrants, they suggests that the first component will predict population origins for the individuals in the sample since it will be correlated with SNPs that show the most discrepancy between the populations. Generally, in such a model with S populations, the first $S-1$ components may predict population memberships.

2.1.3.1 Human Genetic Diversity Panel

Using SNPs from the HGDP-CEPH panel from only chromosomes 1 to 22 and 1043 individuals from 27 countries out of the 52 in the panel, the first two components from principal components analysis are plotted in figure 2.4(a). In this figure, individuals from the same countries are clustered together. More generally, clusters geographically close appear closer in this plot, corresponding to the results by Novembre (2008) in Europe. A plot of the 3rd and 4th eigenvectors is given in figure 2.4(b) and shows three main cluster.



(a) (b)

Figure 2.4: (a) Biplot of first two components using data from the HGDP-CEPH diversity panel. Each point on the plot is representative an individual and is shaped and coloured associatively to one of 27 countries. (b) Biplot of the 3rd and 4th components.

The first consists of individuals from the New Guinea and Bougainville populations, the second of individuals from South America, with the remaining individuals forming the last cluster. In total, 80 components were found to be significant.

In order to gauge a range of F_{st} values applicable to the human population, pairwise F_{st} values are also computed using the program SMARTPCA introduced by Patterson et al. (2006), and are summarized in table 2.1. This program uses a jacknife estimator of F_{st}. The smallest F_{st} was found to be between the two Italian populations (0.005) followed closely by the F_{st} between France and Italy, 0.007, highlighted by red on the table. The largest F_{st} values were found to be between African and South American populations, for example $F_{st} = 0.343$ between Brazil and Namibia, followed closely by those between African and Oceania populations, for example $F_{st} = 0.312$ between New Guinea and Namibia.

2.1.4 Bayesian clustering approach

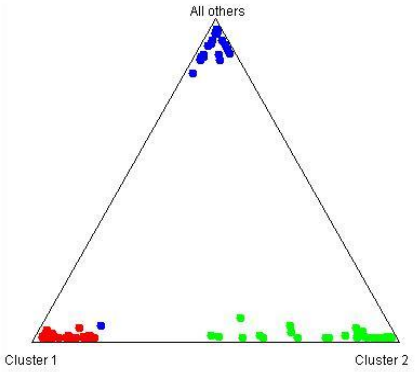
Another method of detecting population structure is STRUCTURE, originally introduced by Pritchard et al. (2000). This is a model-based clustering technique used to cluster individuals into populations. This program assumes that the number of clusters (or populations), k, is unknown and uses a Dirichlet distribution to model the allele frequencies of the alleles at a particular locus in each of the k populations. Draws from the joint distribution of the allele frequencies for each population, P and individual population allocations, Z given the data, X are made using an MCMC algorithm that iteratively samples P from $Pr(P|X, Z)$ and then Z from $Pr(Z|X, P)$. Initial values of Z are chosen by assuming that the probability an individual belongs to the kth population is equal to $\frac{1}{k}$. This method was extended to allow for admixture. Each individual is associated with a vector $q = \{q_1, \ldots, q_k\}$, where q_i is the proportion of that individual's genome belonging to population i.

In order to infer k, the number of populations, the authors approximate the posterior distribution of k given data X although they also suggest their estimation may not be accurate. Given a prior $p(k)$ on k and the likelihood of data X given k $p(X|k)$ then

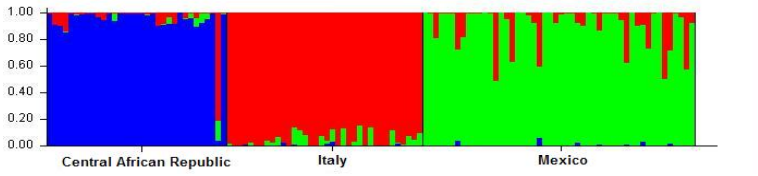
$$p(k|X) \propto p(X|k)p(k).$$

The troublesome aspect of this approach is finding $p(X|k)$, however, the authors provide an estimation of this probability.

An example is given in figure 2.5(a) from individuals from the Mexican (coloured green), Italian (red) and Central African Republic (blue) populations in the HGDP-CEPH diversity panel. The program labels the clusters cluster 1, cluster 2 and all others. This simple example shows the results with $k = 3$. Each point in the equilateral triangle represents an individual in the sample and is coloured depending of its origin. The position of each point is determined by $q_i = \{q_{i1}, q_{i2}, q_{i3}\}$, such that $\sum_{j=1}^{3} q_{ij} = 1$, and q_{ij} is the perpendicular distance from side j of the triangle, and represents the estimated ancestry



(a)



(b)

Figure 2.5: (a) STRUCTURE clustering results with $k = 3$ using data from 3 populations from the HGDP-CEPH diversity panel (b) Another graphical representation of the data.

of individual i in component j. Individuals are clustered depending on their originating population, with only one person from the Central African Republic, in blue, clustered with people from Italy, shown as cluster one. Figure 2.5(b) gives another graphical representation of the results for each individual as described by Rosenberg (2004). In this diagram, each individual is represented through a bar and each bar is coloured depending on the individual's estimated ancestry.

Table 2.1: Pairwise F_{st} values of 27 populations from the HGDP-CEPH data.

	CAR	Congo	Sen	Nig	Ken	Nam	S Afr	Alg	Isr Neg	Isr Car	Isr Cen	Pak	China	Sib	Jap	Camb	New Gui	Boug	Fra	Ita	Ita Ber	Ork	Rus Cau	Russ	Mex	Col	Brazil
CAR	0	0.061	0.046	0.042	0.039	0.085	0.031	0.121	0.145	0.164	0.148	0.152	0.196	0.198	0.205	0.193	0.262	0.250	0.173	0.174	0.171	0.176	0.166	0.171	0.241	0.277	0.292
Congo		0	0.085	0.081	0.075	0.101	0.066	0.159	0.183	0.202	0.186	0.189	0.232	0.235	0.241	0.230	0.297	0.285	0.211	0.212	0.209	0.214	0.204	0.208	0.277	0.313	0.328
Sen			0	0.011	0.017	0.107	0.015	0.095	0.123	0.143	0.127	0.131	0.176	0.178	0.184	0.173	0.242	0.230	0.153	0.154	0.151	0.156	0.146	0.151	0.221	0.258	0.272
Nig				0	0.012	0.105	0.010	0.099	0.127	0.147	0.130	0.134	0.178	0.181	0.186	0.175	0.244	0.233	0.157	0.157	0.154	0.160	0.149	0.154	0.224	0.260	0.274
Ken					0	0.098	0.010	0.088	0.113	0.132	0.116	0.121	0.166	0.168	0.174	0.163	0.233	0.221	0.142	0.143	0.140	0.146	0.135	0.140	0.211	0.248	0.262
Nam						0	0.077	0.174	0.196	0.215	0.199	0.203	0.247	0.249	0.255	0.244	0.312	0.300	0.224	0.225	0.222	0.227	0.217	0.221	0.292	0.329	0.343
S Afr							0	0.099	0.126	0.146	0.129	0.133	0.177	0.179	0.185	0.174	0.243	0.231	0.155	0.156	0.153	0.159	0.148	0.153	0.223	0.259	0.274
Alg								0	0.023	0.030	0.023	0.036	0.105	0.102	0.115	0.104	0.185	0.170	0.033	0.031	0.030	0.039	0.034	0.038	0.146	0.189	0.202
Isr Neg									0	0.014	0.009	0.023	0.099	0.095	0.110	0.098	0.182	0.167	0.021	0.019	0.017	0.027	0.018	0.027	0.140	0.182	0.196
Isr Car										0	0.011	0.021	0.102	0.097	0.113	0.101	0.187	0.171	0.017	0.015	0.013	0.022	0.013	0.022	0.141	0.184	0.198
Isr Cen											0	0.018	0.096	0.092	0.107	0.095	0.180	0.164	0.016	0.014	0.011	0.021	0.012	0.021	0.136	0.180	0.193
Pak												0	0.064	0.062	0.074	0.064	0.157	0.141	0.022	0.028	0.020	0.023	0.011	0.017	0.108	0.150	0.163
China													0	0.024	0.007	0.013	0.166	0.138	0.102	0.108	0.101	0.103	0.088	0.086	0.099	0.139	0.151
Sib														0	0.028	0.042	0.176	0.151	0.095	0.102	0.094	0.095	0.082	0.077	0.098	0.138	0.151
Jap															0	0.023	0.174	0.146	0.122	0.119	0.112	0.113	0.098	0.096	0.100	0.139	0.152
Camb																0	0.163	0.136	0.102	0.107	0.100	0.103	0.088	0.088	0.110	0.150	0.163
New Gui																	0	0.104	0.189	0.194	0.188	0.190	0.179	0.180	0.230	0.269	0.282
Boug																		0	0.173	0.178	0.171	0.174	0.162	0.163	0.208	0.248	0.260
Fra																			0	0.007	0.003	0.007	0.012	0.008	0.137	0.180	0.193
Ita																				0	0.005	0.014	0.015	0.016	0.145	0.189	0.203
Ita Ber																					0	0.008	0.009	0.009	0.136	0.180	0.194
Ork																						0	0.014	0.008	0.135	0.179	0.192
Rus Cau																							0	0.012	0.126	0.169	0.182
Russ																								0	0.120	0.163	0.177
Mex																									0	0.064	0.076
Col																										0	0.091
Brazil																											0

Chapter 3

Data simulation

This thesis will look in particular at SNP data. Therefore, this chapter begins by introducing some software for simulating SNP data. Data simulation is an invaluable tool in many fields to examine characteristics of statistical models.

Two demographic models are to be considered in this thesis: the island model (or migration model) and an isolation model. Particulars on each model are provided including assumptions and methods of simulating SNP data.

In both models, let D denote the number of subpopulations with N_i and n_i the population sample sizes of subpopulation i, respectively, and $\sum_{i=1}^{D} N_i = N^T$, $\sum_{i=1}^{D} n_i = n^T$.

3.1 Software for simulating data

Many computer programs are available to simulate genetic data, many of which take a coalescent approach, simulating backwards in time, although others adopt a forward-in-time approach. An in-depth review of many of these programs is given by Excoffier and Heckel (2006). Data simulators include 'ms' by Hudson (2002), 'SimCoal' by Excoffier et al. (2000) and 'Fregene' by Chadeau-Hyam et al. (2008).

Hudson's ms program uses a coalescent approach to simulate sequence data with an infinite sites mutation model for a wide range of population structure scenarios, for example migration, divergence and variable population size. There are more such backwards-in-time simulators such as 'SimCoal', a coalescent simulation of different types of genetic data, originally DNA sequences and microsatellite repeat counts under complex demographic histories such as population expansion, bottlenecks, divergence and migration. 'SimCoal2' by Laval and Excoffier (2004) was later introduced to simulate SNP data and incorporate ascertained data. 'Serial SimCoal' by Anderson et al. (2005) is an extension of SimCoal, which allows multiple sampling time points, where present-day and ancestral data may be generated in order to compare the amount of diversity of a set of populations through time.

Similarly, Fregene by Chadeau-Hyam et al. (2008) simulates DNA sequence data from a possibly subdivided population. However, Fregene adopts a forward-in-time approach. The authors argue that simulating forward in time allows for more flexibility in modelling recombination and selection, neither of which will be considered here. Naturally, it incorporates the other factors such as changing population size, mutation and migration.

This list is in no way exhaustive. In this thesis, data simulation was approached using the backwards-in-time coalescent model and were implemented in the statistical software package, R, R Development Core Team (2008).

3.2 Strategy for simulating data

Hudson (1991) describes simulating a genealogy and then adding mutations. In a sample of haploid size n from a population of haploid size N with $N \gg n$, beginning at time 0, the time until the next coalescent event measured in units of N generations, backwards in time, is exponentially distributed with rate $\binom{n}{2}$ and each of the $\binom{n}{2}$ pairs have equal probability of being chosen to coalesce. That is, a pair of lineages are chosen to coalesce at time $T_n \sim Exp\left(\binom{n}{2}\right)$. After this time, the sample size reduces by one and, from the remaining

$n-1$, a pair are randomly chosen to coalesce after a further time $T_{n-1} \sim Exp\left(\binom{n-1}{2}\right)$. This process continues until the final pair coalesce at time $T_{MRCA} = \sum_{i=2}^{n} T_i$. Figure 1.1 illustrates the coalescent process with $n = 5$.

Once the genealogy has been simulated, a Poisson number of mutations are added to the tree. Let S denote the number of mutations. On a branch of length t generations, the number of mutations is a $Poi(\mu t)$ random variable where μ is the total mutation rate per generation. On a genealogy of length T_{total}, assuming mutations occur independently on each branch and measuring time in N generations,

$$\begin{aligned} E(S) &= \mu N \sum_{i=2}^{n} i T_i \\ &= \frac{\theta}{2} T_{total}, \end{aligned}$$

where $\theta = 2N\mu$. In the simulation, $S \sim Poi(\frac{1}{2}\theta T_{total})$ mutations are randomly places on the genealogy.

Depaulis and Veuille (1998) simulated data in a slightly different manner. They were interested in the distributions of some statistics under a neutral model, hoping to show the statistics are powerful in testing whether observed data are consistent with a neutral model, for a given number of mutations so they simulated genealogies and added the fixed number of mutations. However, the number of mutations depends on θ and the total length of the tree. For example, given θ, longer trees are expected to have more mutations than shorter trees. Markovtsova et al. (2001) show that the distribution of a statistic given S is not independent of θ, suggesting the power of the test depends on θ. By simulating data under the fixed S method employed by Depaulis and Veuille (1998) to find rejection regions (at the 5% level) for some statistics and testing data simulated under the method used by Hudson (1991), Wall and Hudson (2001) found that the type I error rate of the various statistics was around 5% for a range of values of θ. They suggest that the fixed number of mutations method was acceptable assuming the true value of θ was not too large or too small. Specifically, Depaulis and Veuille (1998) illustrated their test using

data with $S = 44$ and $n = 20$ and found the data to be inconsistent with the neutral model. Using Hudson's method of simulation, Markovtsova et al. (2001) found similar results when

$$\theta \leq \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}},$$

where the right hand side is just the estimator of θ given by Watterson (1975). Therefore, Markovtsova et al. (2001) suggest that the fixed number of mutations method is adequate assuming θ is less than Watterson's estimate. The method, although approximate, is algorithmically convenient.

Attention in this thesis is given to biallelic SNP data. The method of Depaulis and Veuille (1998) is employed with $S = 1$. A genealogy is simulated under a specific model and a single mutation randomly added to the tree. Given Hudson's method for simulating data under the standard neutral model, details are now given for simulating a genealogy under the migration and isolation models respectively.

3.3 Migration model

This model was first described by Wright (1969) and an example with 4 subpopulations is illustrated in Figure 3.1. Consider D subpopulations, with the ith subpopulation of haploid population size N_i. In each generation, migration events occur between subpopulations at a constant rate q_{ij} for $i \neq j \in \{1, \ldots, D\}$. More specifically,

q_{ij} = the probability that a parent in population i has its child in population j.

In each generation, each individual produces a finite number of offspring. In subpopulation i, the expected number of offspring to migrate to subpopulation j in a generation is equal to $N_i q_{ij}$. Simulating under the coalescent begins at the present and works backwards in time, therefore migration events need to be defined backwards in time. Given an individual

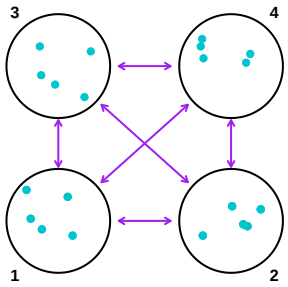


Figure 3.1: Example of migration model with 4 subpopulations. Arrows show possible migrations between each subpopulation.

in subpopulation i, the probability that their parent belonged to subpopulation j in the previous generation is denoted by m_{ij}. The probability that the parent of an individual in subpopulation i did not belong to subpopulation i is equal to $\sum_{j \neq i} m_{ij}$. The parameters m_{ij} are related to q_{ji} by applying Bayes theorem:

$$m_{ij} = \frac{N_j q_{ji}}{\sum_{k=1}^{D} N_k q_{ki}}. \tag{3.1}$$

This expression for m_{ij} is the number of migrants from j to i, $N_j q_{ji}$, divided by the total number to migrate to i in a single generation. It is assumed that the population size in each subpopulation remains constant through time. If subpopulation i is of size N_i in the current generation then, in the next generation, after all migration events occur, it is then of size

$$N_i' = \left(N_i - \sum_{j \in J} N_i m_{ij} \right) + \sum_{j \in J} N_j m_{ji},$$

where $J = \{1, \ldots, i-1, i+1, \ldots, D\}$. In the case that $N_1 = \cdots = N_D$ and there is a constant migration rate between all subpopulations, then the population size in the next generation will be N_i. Otherwise, a sample of size N_i is taken, with replacement, from N_i'.

Measuring time in N^T generations and allowing $N^T \to \infty$, this structured model converges to the structured coalescent. The structured coalescent is a continuous-time Markov process with state space equal to the set of vectors $\alpha = \{\alpha_1, \ldots, \alpha_D\}$, where α_i is the number of lineages in subpopulation i in a particular generation. In the structured coalescent, only two possible events can occur: either two individuals from the same subpopulation coalesce or an individual migrates from one subpopulation to another. Coalescent and migration events occur as independent Poisson processes. Coalescent events occur at rate

$$\sum_{i=1}^{D} \frac{1}{c_i} \binom{\alpha_i}{2},$$

where $c_i = \frac{N_i}{N^T}$ and α_i is the sample size in subpopulation i at the given time. Migration events occur at rate

$$N^T \sum_{i=1}^{D} \alpha_i m_i.$$

Event times are exponential draws with rate given by the sum of the rates of all possible events, i.e.,

$$\Lambda \quad = \quad N^T \sum_{i=1}^{D} \alpha_i m_i + \sum_{i=1}^{D} \frac{1}{c_i} \binom{\alpha_i}{2}.$$

3.3.1 Simulation

The details of simulating data under the structure coalescent are given by Nordborg (2007). The sample is considered in the present time, time zero. At each step in the simulation, an exponential event time of rate Λ is drawn and either a migration or coalescent event occurs. The event is a coalescent event with probability

$$P_C = \frac{\sum_{i=1}^{D} \frac{1}{c_i} \binom{\alpha_i}{2}}{\Lambda},$$

and is a migration event with probability

$$P_M = \frac{N^T \sum_{i=1}^{D} \alpha_i m_i}{\Lambda}.$$

If a migration event takes place, then a random pair of subpopulations is selected with probabilities $\{m_{ij} : i \neq j \in 1, \ldots, D\}$, with one of the subpopulations the originating population and a lineage chosen randomly to migrate to the other subpopulation. The sample size of the originating subpopulation decreases by one and the receiving subpopulation sample size increases by one. If the event is a coalescent event, a subpopulation is selected with probability $\frac{1}{D}$ and two distinct lineages within that subpopulation are selected to coalesce. The sample size of that subpopulation decreases by one. This process is repeated until the most common recent ancestor of the sample in reached, when a genealogy of a sample of size n^T has been simulated. At this stage, a single mutation time is drawn from a $Un(0, T_{total})$ distribution and randomly placed on one of the branches in the genealogy at that time. For instance, consider the example shown in figure 1.1. Figure 3.2 illustrates the procedure used to add a mutation to the simulated genealogy. Firstly, the tree is partitioned, as shown on the left hand side, where each element of the partition is differently coloured. The elements are placed side by side creating the bottom line in figure 3.2 of length T_{total} (not drawn to the same scale). A mutation time is drawn uniformily, as shown by the red bar, and added to the corresponding branch on the genealogy shown on the right hand side, as a red square.

3.3.2 Example of simulated data set

In order to illustrate this method of simulation, data were simulated under the migration model with four subpopulations with each of population size 500 and sample size 50. The general case would have $\binom{4}{2} = 6$ migration rates. In this example the rates are chosen symmetrically, hence $m_{ij} = m_{ji}$, which requires 3 migration rates to be defined. By choosing a high migration rate, m_1, between subpopulations 1 and 2, these two subpopulations will be more genetically similar. Subpopulation 3 will be more genetically distinct from subpopu-

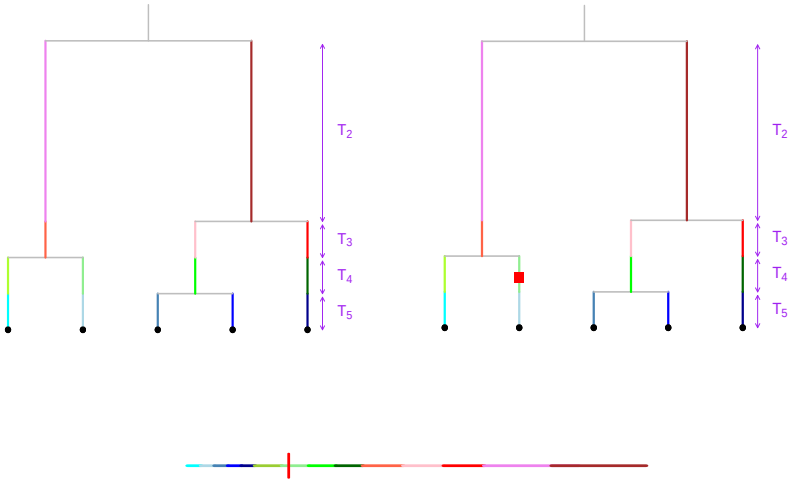
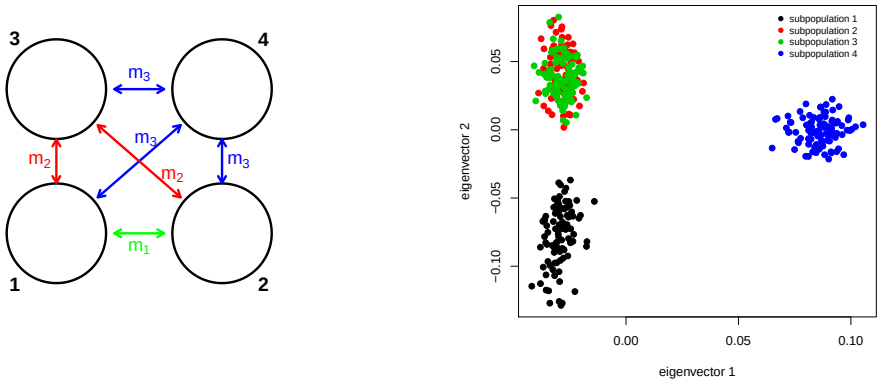


Figure 3.2: Adding mutation to genealogy with $n = 5$.

lations 1 and 2 by choosing an intermediate migration rate m_2 between $\{1, 3\}$ and $\{2, 3\}$. Subpopulation 4 will be the most distinct subpopulation in this sample by choosing a low migration rate, m_3, between subpopulation 4 and the other subpopulations in the sample. 1000 SNPs were simulated under this model with $\{m_1, m_2, m_3\} = \{0.08, 0.008, 0.0005\}$ and principal components analysis was performed with figure 3.3(a) showing the structure of this example and figure 3.3(b) showing a plot of the first two components. The first two components are unable to distinguish subpopulations 1 and 2. The first component isolates subpopulation 4 from the other subpopulations and the second component separates $\{1, 2\}$, 3 and 4. In this example, only the first two components were found to be significant in capturing structure.



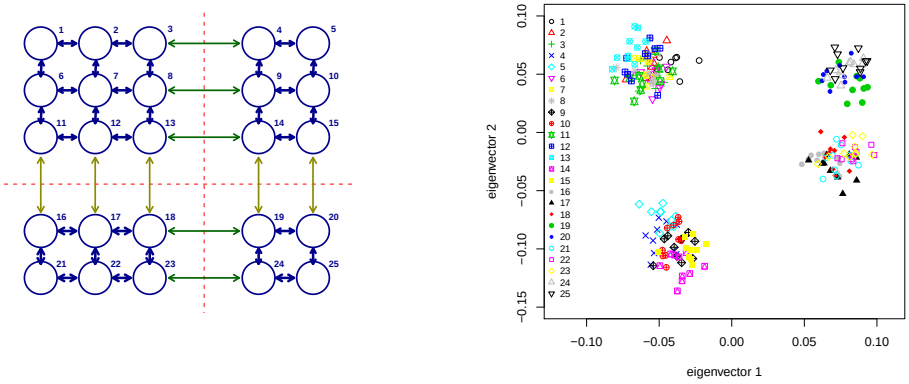
(a) (b)

Figure 3.3: Example of migration model with 4 subpopulations and three migration rates $m_1 < m_2 < m_3$, corresponding to high, intermediate and low migration between subpopulations.

3.3.3 25 population example

To demonstrate how genetic distances mirror geographical distances, as shown by Novembre (2008) in the case of Europeans, a lattice of 25 subpopulations was constructed such that each subpopulation can exchange migrants with neighbouring populations. In addition, migration was restricted in certain places to mimic barriers to gene flow such as a challenging topography. Figure 3.4(a) illustrates the construction of this example.

1000 SNPs were simulated under this model with a sample size of 10 from each subpopulation. Three migration rates were set and the red dotted lines correspond to areas of restricted migration. High levels of migration were assigned between neighbouring pairs of subpopulations from the same set within $\{1, 2, 3, 6, 7, 8, 11, 12, 13\}$, $\{4, 5, 9, 10, 14, 15\}$, $\{16, 17, 18, 21, 22, 23\}$ and $\{19, 20, 24, 25\}$ shown with blue arrows and lower migration between the remaining neighbouring pairs of subpopulations, corresponding to the green and yellow arrows in figure 3.4(a). Figure 3.4(b) shows a plot of the first two components from principal components analysis performed on the simulated data set. The resulting clustering shows a similar pattern to the original construction. The red dotted lines, shown



(a) (b)

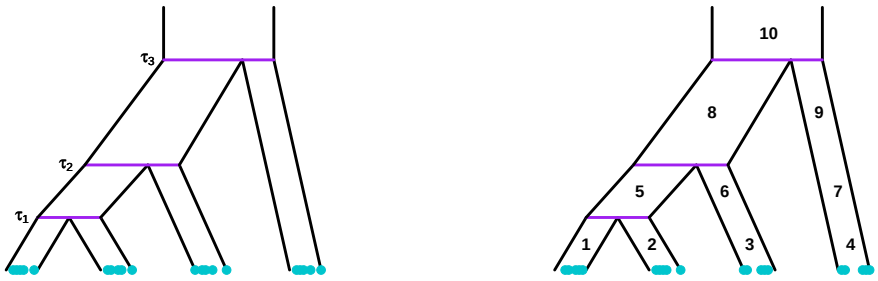
Figure 3.4: (a) 5×5 lattice of 25 subpopulations with migration between neighbouring populations. Blue arrows correspond to higher migration whereas green and yellow arrows correspond to restricted migration. (b) Plot of first two components from principal components analysis.

on the left, separate the 25 subpopulation into four groups. The plot on the right shows four main clusters corresponding to these groupings. The high level of migration within each group results in subpopulations within groups being less distinct than subpopulations taken from different groups.

3.4 Isolation model

The second model to be considered is a model of isolation. In the beginning, only a single population existed. At a given time in the past, τ, the population split into two subpopulations which subsequently evolved independently, i.e. without exchanging migrants. The further back in time the population split occurs, the more genetically diverse the subpopulations are. An example of this model is shown in figure 3.5(a) showing three splitting episodes.

This model assumes that individuals can only coalesce if they belong to the same subpopulation. Once two subpopulations join together, backwards in time, they are consider to



(a) (b)

Figure 3.5: (a) Isolation model with fout subpopulations. (b) Populations labelled from the present time backwards until there is a single ancestral population.

be a single subpopulation. As in the previous case of migration, within subpopulation i, individuals coalesce at the rate

$$\frac{N^T}{N_i}\binom{\alpha_i}{2},$$

where α_i is the number of lineages present in subpopulation i at the given time. In this case, the intensity of this process depends on time through the number of lineages present in the sample as well as the number of subpopulations in existence. If $I(t)$ is the set of populations in existence at time t, the waiting time till a coalescent event in any of the members of $I(t)$ is exponentially distributed with rate

$$R(t) = \sum_{i \in I(t)} \frac{N^T}{N_i}\binom{\alpha_i(t)}{2}. \tag{3.2}$$

Furthermore, the probability that the event occurs in the ith subpopulation at time t is

$$\frac{1}{R(t)}\frac{N^T}{N_i}\binom{\alpha_i}{2}. \tag{3.3}$$

3.4.1 Simulation

In simulating data under this model, the sets $I(t)$ are recorded by labelling the subpopulations from 1 to $\frac{1}{2}D(D+1)$ as illustrated in figure 3.5(b), since D populations in the present are generated from $D-1$ binary splits, and, at each split, a relabelling with one fewer populations is made. In this example, there are three population split times, τ_1, τ_2 and τ_3. Hence

$$I(t) = \begin{cases} \{1,2,3,4\}, & 0 \leq t < \tau_1; \\ \{5,6,7\}, & \tau_1 \leq t < \tau_2; \\ \{8,9\}, & \tau_2 \leq t < \tau_3; \\ \{10\}, & \tau_3 \leq t. \end{cases}$$

To generate the first event time t_{n^T} (the time during which there are n^T lineages present in the sample), a draw, $t^*_{n^T}$, is made from an exponential distribution with rate $R(0)$. The difficulty lies in whether $t^*_{n^T}$ is more or less recent than the first split time, τ_1.

1. If $t^*_{n^T} < \tau_1$ then $t_{n^T} = t^*_{n^T}$ and we may proceed to select a subpopulation for the coalescent event to occur with the probabilities in (3.3).

2. If $t^*_{n^T} \geq \tau_1$ then the additional time, from τ_1, until the coalescent event needs to be considered.

The procedure used in scenario two is described by Ross (1997). Let $F(x)$ and $F_{\tau_1}(x)$ be the distribution functions of the time until the next coalescent event and the additional time from τ_1 until the next coalescent event, respectively. Then,

$$\begin{aligned} F_{\tau_1}(x) &= P\{\text{time from } \tau_1 \text{ until next coalescent event is less than } x \mid \text{event time} \geq \tau_1\} \\ &= P\{\text{event occurred in } (\tau_1, \tau_1 + x)\} \\ &= 1 - P\{\text{no event occurred in } (\tau_1, \tau_1 + x)\} \\ &= 1 - P\{\text{event time} > \tau_1 + x \mid \text{event time } \geq \tau_1\} \\ &= 1 - P\{\text{event time} > x\} \end{aligned} \tag{3.4}$$

$$\begin{aligned} &= P\{\text{event time} < x\} \\ &= F(x). \end{aligned}$$

This proof implements the memoryless property of the exponential distribution at (3.4). That is, any random variable, $X \sim Exp(\theta)$,

$$P(X > x + y | X > x) = P(X > y) \qquad \text{for any } x, y \geq 0.$$

Therefore, if the coalescent time is greater than τ_1 then the additional time after τ_1 until the event time is independent of τ_1. Hence, the first event after τ_1 would occur at $t_{n^T} = \tau_1 + x$, where $x \sim Exp\Big(R(\tau_1)\Big)$.

From here, we wish to simulate the remaining $t_{n^T-1}, \ldots, t_2$ events times. Let n_{split} = the number of population split times and $\tau_0, \ldots, \tau_{n_{split}}$ be the ordered split times with $\tau_0 = 0$ and $0 < \tau_1 \leq \ldots \leq \tau_{n_{split}}$. For waiting time t_i, identify k such that

$$\sum_{j=i}^{n^T} t_j < \tau_k \text{ and } \sum_{j=i}^{n^T} t_j \geq \tau_{k-1}.$$

To find the next waiting time, make draw t^* from $Exp\Big(R(t_i)\Big)$.

1. If
$$\sum_{j=i}^{n^T} t_j + t^* < \tau_k,$$
then $t_{i-1} = t^*$.

2. If
$$\sum_{j=i}^{n^T} t_j + t^* \geq \tau_k,$$
then $t_{i-1} = \tau_k + x$ where x is an exponential draw with rate $R(\tau_k)$.

3. If

$$\sum_{j=i}^{n^T} t_j + \tau_k + x > \tau_{k+1},$$

then another draw $x_1 \sim Exp\Big(R(\tau_{k+1})\Big)$ is made and $t_{i-1} = \tau_{k+1} + x_1$.

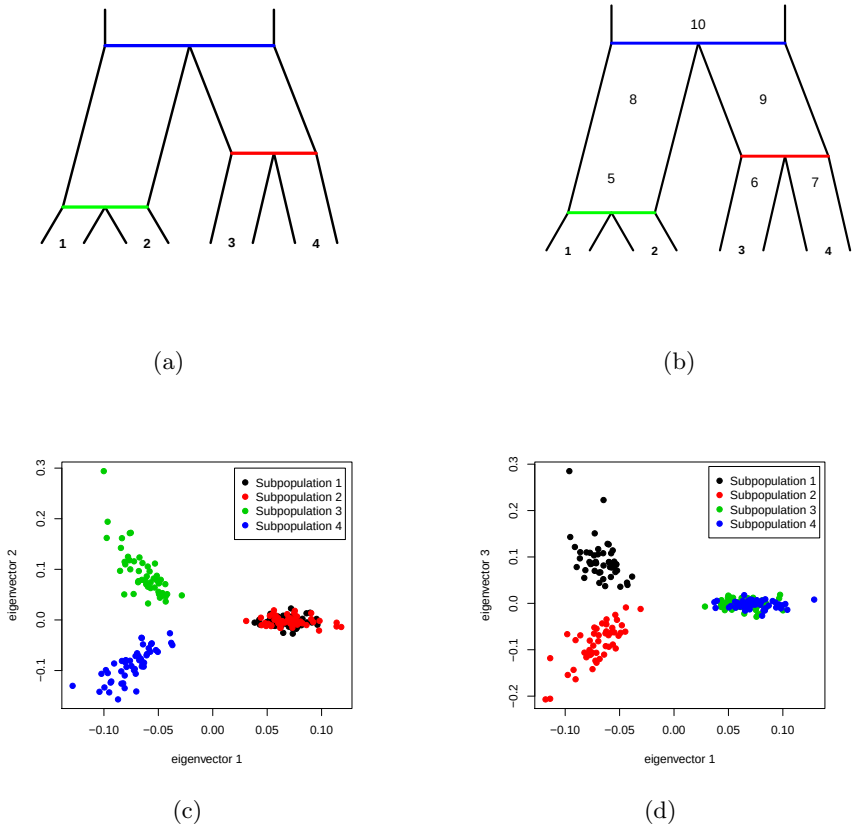
It is possible that $\sum_{j=i}^{n^T} t_j + \tau_{k+1} + x_1 > \tau_{k+2}$, in which case, $t_{i-1} = \tau_{k+2} + x_2$ for $x_2 \sim Exp\Big(R(\tau_{k+2})\Big)$. This step is repeated until $q(\geq k)$ is found such that

$$\sum_{j=i-1}^{n^T} t_j < \tau_q \text{ and } \sum_{j=i-1}^{n^T} t_j \geq \tau_{q-1}.$$

The process is applied for all $i = n^T, \ldots, 2$, until the most recent common ancestor of the sample is reached and then a single mutation is added randomly to the genealogy, as before.

3.4.2 Example of simulated data set

To illustrate, data were simulated from four subpopulations, each with population size 500 and sample size 50 under the isolation model. Figure 3.6(a) shows the model. Population divergence times τ_1 (between subpopulations 1 and 2), τ_2 (between subpopulations 6 and 7) and τ_3 (between 8 and 9) were specified such that τ_1 is the most recent, τ_3 the oldest and τ_2 an intermediate time. Simulating 1000 SNPs under this model with $\{\tau_1, \tau_2, \tau_3\} = \{0.003, 0.03, 0.4\}$, figure 3.6(c) displays the first two components from principal components analysis and figure 3.6(d) shows the first and third components. The first component separates subpopulations 1 and 2 from subpopulations 3 and 4. The second component separates $\{1, 2\}$, 3 and 4 and the third component separates 1,2 and $\{3, 4\}$.



(a) (b)

(c) (d)

Figure 3.6: (a) Isolation model with 4 subpopulations and 3 population divergence times. (b) Labelling of subpopulations from 1 to 10. (c) Biplot of first two components from data simulated under the isolation model. (d) Biplot of first and third components.

3.5 Computing F_{st}

It is possible to simulate data from both the migration and isolation models to produce similar F_{st} values between subpopulations. Section 2.1.1 showed ways of estimating a migration rate, through (2.2), and population divergence times, via (2.3). Slatkin (1991, 1993) approximated F_{st} through coalescent times by considering a definition of F_{st} in terms of identity by descent, the probability that two genes reach their most recent common

ancestor unaffected by any mutation. That is,

$$F_{st} = \frac{f_0 - \bar{f}}{1 - \bar{f}}, \tag{3.5}$$

where f_0 is the probability of identity by descent of two genes sampled from the same subpopulation and $\bar{f}$ is the probability of identity by descent of two genes sampled from the whole population, regardless of subpopulation. Wakeley (2009) writes the probability of identity by descent as an infinite sum:

$$\begin{aligned} P\{IBD\} &= \sum_{t=1}^{\infty} (1-\mu)^{2t} \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N} \\ &= \sum_{i=1}^{\infty} (1-\mu)^{2t} P(t), \end{aligned} \tag{3.6}$$

where μ is the mutation rate per generation. $P(t)$ is the probability that the two lineages reach their most common recent ancestor in generation t and $(1-\mu)^{2t}$ is the probability that neither lineage is affected by a mutation up to and including generation t. For a small mutation rate, Slatkin (1993) approximated (3.6) by

$$\begin{aligned} \bar{f} &\approx \sum_{t=1}^{\infty} (1 - 2t\mu) P(t) \\ &= 1 - 2\mu\bar{t}. \end{aligned}$$

and showed from (3.5) that

$$\begin{aligned} F_{st} &= \frac{f_0 - \bar{f}}{1 - \bar{f}} \\ &\approx \frac{(1 - 2\mu t_0) - (1 - 2\mu\bar{t})}{1 - (1 - 2\mu\bar{t})} \\ &= \frac{\bar{t} - \bar{t}_0}{\bar{t}}, \end{aligned} \tag{3.7}$$

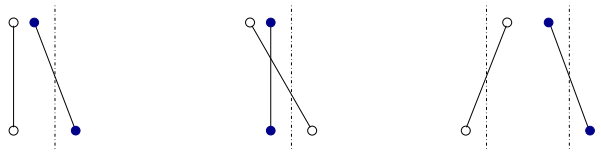
where $\bar{t}_0$ is the average coalescent time between two genes from the same subpopulation and $\bar{t}$ is the average coalescent time between two genes samples from the whole population.

3.5.1 Migration model

An in-depth analysis of the migration model with D subpopulations was given by Slatkin (1991). Suppose each subpopulation is of haploid size N. In this model the migration and coalescent processes occur independently. Slatkin (1991) presents the result that if subpopulations exchange migrants at constant rate m, then $\bar{t}_0 = N^T = DN$, that is the total population size. Using this results, the probability that two genes from the same subpopulations coalesce in the next generation is $\frac{1}{DN}$ and so, measuring time in DN generations,

$$\bar{t}_0 \quad = \quad 1.$$

In the case of two genes that are from different subpopulations, Slatkin considered the probability that the two genes were in the same subpopulation in the previous generation. Let m be the probability that a gene migrates to any of the other $D-1$ subpopulations, so the probability that a gene migrates to a specific subpopulations is $\frac{1}{D-1}$. There are three possible ways that two genes, now in different subpopulations, were in the same subpopulation in the previous generation as demonstrated below. The first two scenarios



are the cases of one gene migrating, backwards in time, into the occupying subpopulation of the other gene. The last is the case that both genes migrate from different subpopulations in the current generation to the same subpopulation in the previous generation. When

$m \to 0$, the last scenario has negligible probability. Therefore,

$$\begin{aligned} Pr\{\text{two genes in same subpopulation in the previous generation}\} &= 2m(1-m)\frac{1}{D-1} \\ &\approx \frac{2m}{D-1}, \end{aligned}$$

for small m. Hence,

$$\begin{aligned} \bar{t}_1 &= \text{the time until two genes are in the same subpopulation} \\ &\quad + \text{the average time until two genes in the same subpopulation coalesce} \\ &= \frac{D-1}{2m} + DN. \end{aligned}$$

Measuring time in DN generations leads to

$$\bar{t}_1 = 1 + \frac{D-1}{2DNm}.$$

Lastly, to find $\bar{t}$, Slatkin applied the law of total probability, namely

$$\begin{aligned} \bar{t} &= Pr\{\text{two genes in same subpopulation}\}\bar{t}_0 \\ &\quad + Pr\{\text{two genes from different subpopulation}\}\bar{t}_1 \\ &= \frac{1}{D}\bar{t}_0 + \left(1 - \frac{1}{D}\right)\bar{t}_1 \\ &= 1 + \frac{(D-1)^2}{2D^2Nm}. \end{aligned}$$

As a result, by applying (3.7),

$$F_{st} = \left(1 + \frac{2NmD^2}{(D-1)^2}\right)^{-1}. \tag{3.8}$$

3.5.2 Isolation model

In the isolation model, to find $\bar{t}_0$ consider two genes belonging to the same subpopulation. Within each isolated subpopulation, it is assumed that the lineages coalesce according to a neutral Wright-Fisher model with population size N, as illustrated in figure 3.7. Measuring time in generations, the probability that the two genes coalesce in the previous generation is $\frac{1}{N}$. Since it has been assumed the ancestral population size is also N, the average time till the two genes coalesce is N whether or not they coalesce before or after the split. Measuring time in DN generations,

$$\bar{t}_0 \quad = \quad \frac{1}{D}.$$

Two genes from different subpopulations can only coalesce after the population split time, τ. Therefore, the expected coalescent time of two genes that belong to different subpopulations, $\bar{t}_1$, equals the expected time that two genes coalesce from the same subpopulation plus the population split time τ,

$$\bar{t}_1 \quad = \quad \tau + \frac{1}{D}.$$

To find the average time until two genes coalesce from the entire sample, then either the genes belong to the same or different subpopulations with equal probability. Therefore,

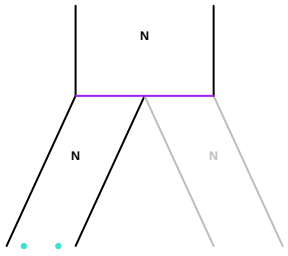


Figure 3.7: Isolation model with two subpopulations.

$\bar{t} = \frac{1}{2}(\bar{t}_1 + \bar{t}_0)$ and so

$$\bar{t} = \frac{1}{D} + \frac{\tau}{2}.$$

From (3.7),

$$F_{st} = \frac{\tau}{2D^{-1} + \tau}. \tag{3.9}$$

As a result, it is possible to choose the parameters τ and m to simulate SNP data from D subpopulations that produce similar pairwise F_{st} values. In the case of two subpopulations, for a given F_{st}, the appropriate estimates are, from (3.9) and (3.8), respectively

$$\hat{\tau} = \frac{F_{st}}{1 - F_{st}}, \tag{3.10}$$

$$\hat{m} = \frac{1 - F_{st}}{8NF_{st}}. \tag{3.11}$$

Simulation results are shown in figure 3.8. In each case, $D = 2$, $N_1 = N_2 = 500$ and $n_1 = n_2 = 50$. For a range of F_{st} values, the migration rate m and

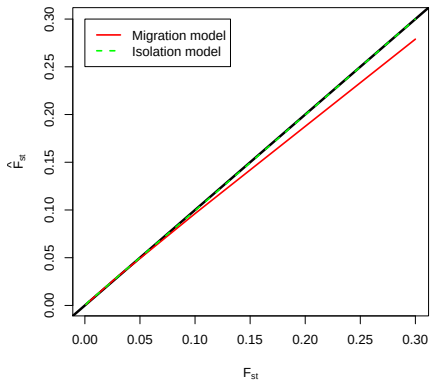


Figure 3.8: Estimates of F_{st} under migration and isolation models with two subpopulations and m and τ estimated using (3.10) and (3.11).

population split time τ were calculated using (3.11) and (3.10) and data were simulated under both models and F_{st} estimated using (3.7). Under the isolation model, the estimated F_{st} values appear almost identical to the predetermined F_{st} values as expected, whereas under the migration model, there is a slight underestimation for larger F_{st} values.

Chapter 4

Distinguishing models

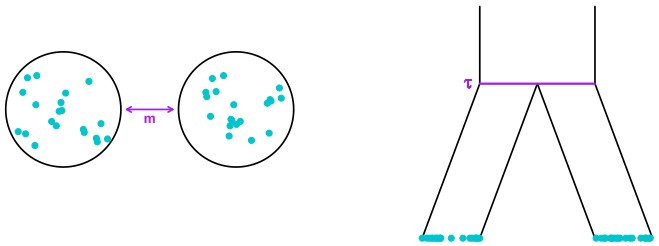
For a given data set, it is often possible to fit both an isolation model and a migration model. In order to infer demographic history, it is important to be able to distinguish between these models. Consider the two models shown in figure 4.1. Figure 4.1(a) shows a two subpopulation migration model. The two subpopulations exchange migrants at rate m per generation. Figure 4.1(b) shows the isolation model with two subpopulations that diverged at time τ in the past. In both cases, assume each of the two subpopulations are of haploid size N and time is measured in units of $2N$ generations.

4.1 Methods of distinguishing migration from isolation

This section introduces and briefly describes established ways of distinguishing these models.

4.1.1 Pairwise differences

Wakeley (1996) showed that the isolation model can be identified by a hypothesis test involving the variance of pairwise differences within and between subpopulations. Suppose



(a) (b)

Figure 4.1: (a) An example of the migration model with two subpopulations that exchange migrants at rate m. (b) An example of the isolation model with 2 subpopulations that diverged at time τ.

DNA sequence data are available from two populations with sample sizes n_1 and n_2 and let $k_{jj'}$ be the number of differences between sequences j and j'. Wakeley defined the average pairwise differences between sequences within subpopulation i as

$$d_i = \frac{1}{\binom{n_i}{2}} \sum_{j=1}^{n_i-1} \sum_{j'=j+1}^{n_i} k_{jj'},$$

where $j, j' \in \{1, \ldots, n_i\}$ and $i = 1, 2$. He also defined the average pairwise difference between the two subpopulations as

$$d_{12} = \frac{1}{n_1 n_2} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} k_{jj'},$$

where $j \in \{1, \ldots, n_1\}$ and $j' \in \{1, \ldots, n_2\}$. He further provided an expression for a variance of the average pairwise differences (within and between subpopulations):

$$s_i^2 = \frac{1}{\binom{n_i}{2}} \sum_{j=1}^{n_i-1} \sum_{j'=j+1}^{n_i} (k_{jj'} - d_i)^2 \quad \text{and,}$$

$$s_{12}^2 = \frac{1}{n_1 n_2} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} (k_{jj'} - d_{12})^2,$$

respectively.

Wakeley considered many function of the intra populations statistics s_1^2/d_1 and s_2^2/d_2 that may be used to distinguish the two model. From the set of functions considered, he found that when the migration rate was high, or a low population divergence time,

$$\Psi = \left[n_1(n_1-1)\frac{s_1}{d_1} + n_2(n_2-1)\frac{s_2}{d_2} + 2n_1n_2\frac{s_{12}}{k} \right],$$

was most successful in distinguishing the two models. He compared the expectation of Ψ over a range of migration rates. As the migration rate increased, the expectation under both models converged to the same value, whereas for smaller migration rates, and so more ancestral population divergence times, the expectation under the migration model is higher and so the isolation model is rejected for 'large' values of Ψ.

4.2 MCMC approach

Nielsen and Wakeley (2001) described an isolation with migration (IM) model consisting of three populations: the ancestral population of population size N_A that branches into two subpopulations (1 and 2) at some time in the past with migration occurring between the two subpopulations. This model assumes constant population size over time, that the populations evolve according to the Wright-Fisher model and that there are no further population subdivisions. Let $\{m_{12}, m_{21}\}$ denote migration rates, with m_{ij} the rate from subpopulation i to j for $i \neq j \in \{1, 2\}$ and let τ denote the population split time, as illustrated in figure 4.2.

Neilsen and Wakeley developed a method for distinguishing this model to one with no migration by firstly fitting the IM model to observed sequence data x and, measuring time in units of N_1 generation, where N_1 is the population size of subpopulation 1, estimating the (scaled) parameters in the model, namely $\phi = \{\theta, M_1, M_2, T, \frac{N_2}{N_1}, \frac{N_A}{N_1}\}$ where θ is the scaled mutation rate, N_i the population size of subpopulation i, M_i is the scaled migration

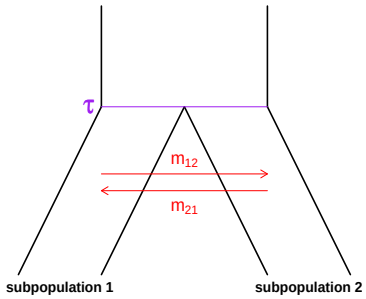


Figure 4.2: Example of isolation with migration model described by Nielsen and Wakeley (2001).

rate of migrants from subpopulation i and T the scaled population divergence time.

In order to estimate the parameters, the authors estimate the likelihood function $L(\phi) = p(x|\phi)$. Taking a Bayesian approach and placing uniform priors on the components of ϕ then, $p(\phi|x) \propto p(x|\phi)$. Therefore, this problem reduces to that described in section 1.2 and is fully described by Beerli and Felsenstein (2001). Treating the likelihood as an expectation, then

$$\begin{aligned} L(\phi) &= \int_G p(x|g)p(g|\phi)dg \\ &= E_p(f), \end{aligned}$$

where f and p are defined as

$$\begin{aligned} p &= p(g|\phi) \quad \text{and} \\ f &= p(x|g). \end{aligned}$$

Importance sampling simulates from another function q similar to p. Neilsen and Wakeley

simulating ϕ_0 from $\pi(\phi)$ and they set

$$q = \frac{p(g|\phi_0)p(x|g)}{\sum_{g\in G} p(g|\phi_0)p(x|g)} = \frac{p(g|\phi_0)p(x|g)}{L(\phi_0)}.$$

Therefore, from (1.2),

$$\begin{aligned} L(\phi) &= E(f) \\ &\approx \frac{1}{m}\sum_{i=1}^{m}\frac{p}{q}f \\ &= L(\phi_0)\frac{1}{m}\sum_{i=1}^{m}\frac{p(g_i|\phi)}{p(g_i|\phi_0)}, \end{aligned}$$

where $\{g_1, \ldots, g_m\}$ are draws from a Markov chain with stationary distribution proportional to $p(x|\phi, G)p(G|\phi)$.

By use of the likelihood ratio test, they compared a model of isolation with migration against a model with no migration, that is the scaled mutation rate M between the two subpopulations equals zero. Given data x and parameters $\phi_0 = \{\theta, M_1 = 0, M_2 = 0, T, \frac{N_2}{N_1}, \frac{N_A}{N_1}\}$, they used the log likelihood ratio

$$T = \log\left[\frac{p(x|\phi_0)}{p(x|\phi)}\right].$$

The test statistic, $-2T$, has an approximate χ^2 distribution with degree of freedom equal to the difference in the number of parameters in the two models under the usual asymptotic theory described by Cox (2006) which requires independent data. However, because of the shared ancestry, genetic data are not independent. Therefore, this standard result is not applicable in this context. Beerli and Felsenstein suggested that the distribution under the null hypothesis may be approximated through simulation.

4.3 Allele frequency spectrum

The allele frequencies under the isolation and migration model are expected to act in similar ways for a range of migration rates and population split times. High migration rates correspond to low F_{st} values and hence subpopulations in the sample will tend to resemble a single panmictic population. Similarly, subpopulations arising from a recent population split again corresponding to a low F_{st} value, will mimic a single panmictic population. On the other hand, for low migration rates (and a more ancestral population split time), it becomes more likely that the subpopulations are more distinct. Figure 4.3 illustrates possible genealogies under the migration (LHS) and isolation (RHS) models. Migration events are represented by red dashed arrows. The top figures correspond to a low migration rate and a large ancestral population split time. Both genealogies have similar characteristics in that the subpopulations find their most common recent ancestor, with an extended T_2. On the left, T_2 is extended due to the time required for the last two lineages to meet in the same populations, whilst on the right, it is extended until the isolated daughter populations have merged.

Data were simulated under both the isolation and migration models, from two subpopulations each of sample size $n = 50$. Figure 4.4 shows allele frequency spectra for a range of migration rates and population split times chosen to produce similar F_{st} values. For small F_{st} values between 0.001 and 0.005, given in the first row of figure 4.4, the distributions of allele frequencies are more or less indistinguishable between the two models. In the second row, with F_{st} ranging between 0.01 and 0.05, slight differences in distributions between the two models develop. In particular, the isolation model's sample has a larger number of SNPs with an allele count less than 5. Also, the variance is greater in the migration model, with a more dramatically right skewed distribution. Lastly, for higher F_{st} values, given in the last row, both models show an increase in the number of SNPs with allele count around 50. This can be explained by considering figure 4.3. The first row gives examples of genealogies from the migration model with a low migration rate, and the isolation model with an ancient population split time τ.

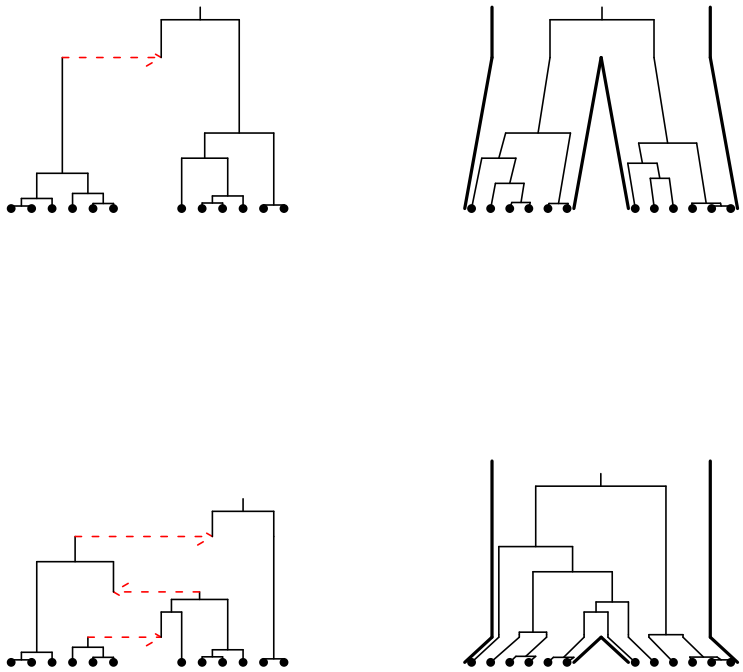


Figure 4.3: Example of genealogies under migration and isolation. The left hand side shows the migration model with low migration (top) and high migration (bottom). The right hand side shows the isolation model with a long population split time (top) and recent population split time (bottom).

In the isolation model, since the population divergence time is large then

$$E(T_2) \quad = \quad \tau + t, \qquad \text{where} \quad t \sim Exp(1).$$

Therefore, as τ increase, the time during which there are two lineages increase hence as $\tau \to \infty$, $E(T_2) \to \infty$. If it is assumed that a mutation occurred randomly at some point on this genealogy, as shown in figure 3.2, it becomes more probable that is occurs during T_2. Therefore separating the sample into n copies of the mutant allele and n copies of the ancestral allele (with n the sample size in each subpopulation). Likewise, in the

migration model, as the migration rate decrease, the time during which there are two lineages increases since the waiting time until a migration event increases. The second row of figure 4.3 shows examples of genealogies with a higher migration rate (LHS) and a lower divergence time (RHS). As the migration rate increase, there is more gene flow between the two subpopulations. As the divergence time approaches zero, the time during which the two subpopulations evolve independently decreases. In both models, the sample increasingly resembles a single panmictic population.

4.3.1 Ambiguities in allele frequency spectra

Data simulated under the isolation model produced an allele frequency spectrum with an excess of rare SNPs but this may be the result of several different demographic characteristics, for example, population growth or natural selection. Williamson et al. (2005) described patterns of allele frequencies under positive, negative and balancing selection. In particular, negative selection leads to an access of rare alleles. Similar results are found under population growth.

4.4 Effects of ascertainment on allele frequency spectrum

Locus ascertainment has a noticeable effect on the allele frequency spectrum, as shown for the neutral model in section 1.3.1, with the degree of the effect depending on the size of the ascertainment sample and, in a structured population, on the subpopulation involved in the SNP discovery process. Nielsen (2004) considered a migration model with two subpopulations. He assumed the ascertainment sample was taken equally from both subpopulations and showed there are no difference in the expected distributions of allele frequencies between the two subpopulations. However, when the ascertainment sample is taken from only one of the subpopulations, the differences between the two subpopulations are more dramatic. In order to see how ascertainment affects the distribution of allele frequencies, data will be simulated under both models with ascertainment.

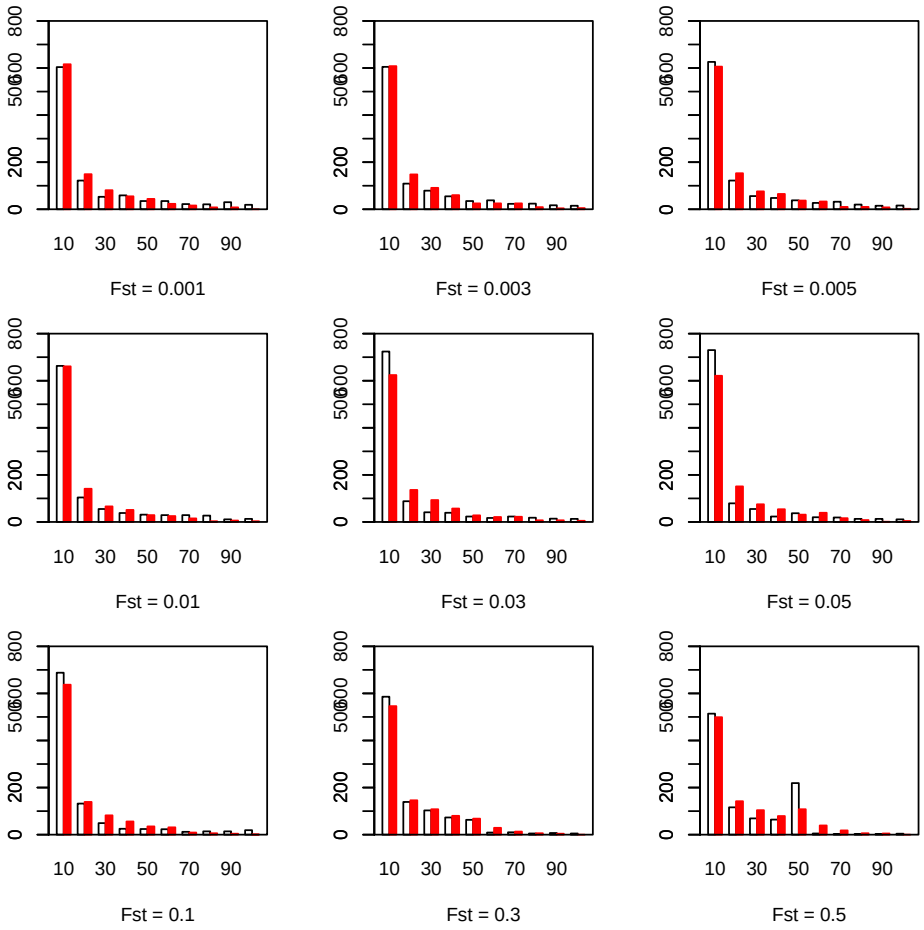


Figure 4.4: Allele frequency spectra from 1000 SNPs simulated from isolation model (white bars) and migration model (red bar) with two subpopulations each of sample size 50 for a range of corresponding F_{st} values

4.4.1 Simulating samples under ascertainment

Consider a sample from a single population of haploid size N, Figure 4.5(a) presents a possible genealogy of a sample of size 10. Once the genealogy has been simulated, a Poisson number of mutations are added to the tree with rate $\theta T_{total}/2$ where $\theta = 2N\mu$, μ is the mutation rate, per generation and T_{total} is the total length of the tree. This simulation incorporates an infinite-alleles model and so every mutation gives rise to a distinct allele.

The red square in figure 4.5(b) represents a mutation on this tree. This particular locus is ascertained by genotyping only a small sample, of size b, of individuals. If there is variability in the ascertainment sample, then a larger sample is genotyped at this locus. In the simulation, the b ascertainment samples corresponding to the blue dots in figure 4.5(c) with $b = 3$. Considering only the branches connecting the ascertainment sample, shown in figure 4.5(d), if there is variability in this small sample, then the remaining $n - b$ nodes are genotyped and only biallelic sites are included in the final sample. In this example, the ascertainment sample shows variability and hence, the remaining nodes are genotyped. Figure 4.6 shows the final data at this locus excluding the ascertainment data.

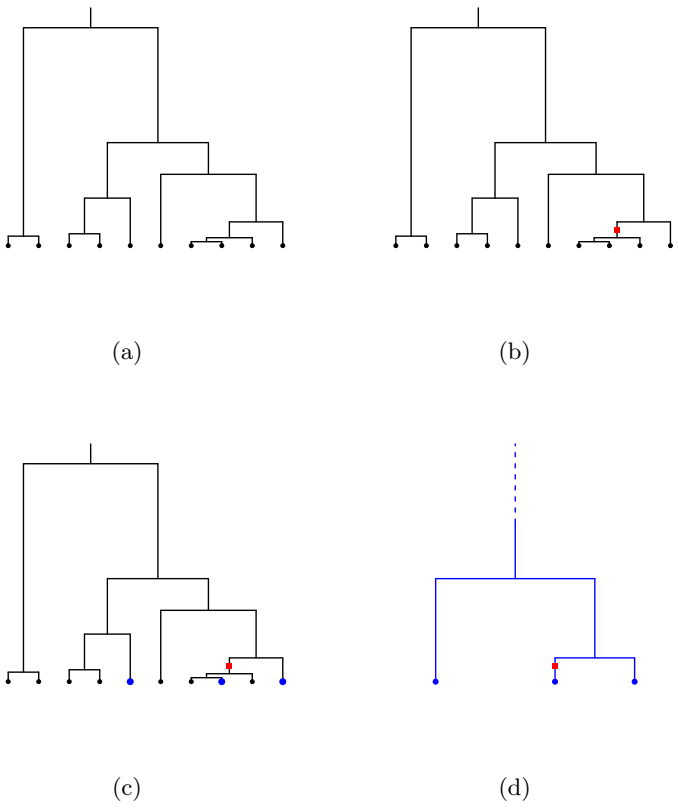


Figure 4.5: Example of simulating data with ascertainment. Blue dots and the red square represent ascertainment sample and a mutation respectively.

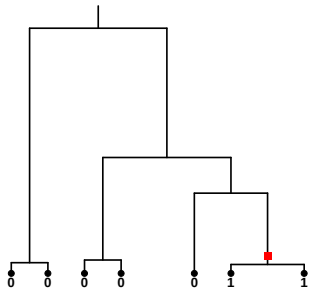


Figure 4.6: Final SNP data from a particular locus under the ascertainment process described in figure 4.5.

4.4.2 Simulating ascertained samples under migration and isolation

1000 SNPs were simulated from the migration and isolation models with a sample size of 50 from each subpopulation for the range of migration rates and population split times used in figure 4.4. As the ascertainment sample size decreases, the bias towards SNPs with intermediate allele counts increases. To see the full extent of this bias under the isolation and migration models, ascertainment samples of size 2 were taken from each subpopulation in both models.

Figure 4.7 shows allele frequency spectra from data simulated under the migration model with and without ascertainment. As the migration rate decreases (or equivalently, as F_{st} increases), the number of SNPs with an allele count around 50 increases, as in the case of no ascertainment. Similar patterns emerge in the case of the isolation model (figure 4.8).

Figure 4.9 compares data simulated under the isolation (white bars) and migration (red bars) models with ascertainment. In terms of distinguishing the two models, similar comparisons can be made to the case of no ascertainment (figure 4.4). As F_{st} increases (corresponding to a decreasing migration rate or an increasing population divergence time), slight differences emerge between the two models. In the second row, the isolation model

shows more alleles with a count less that 5 and shows a bigger increase in alleles with a count equal to 50 in the last row.

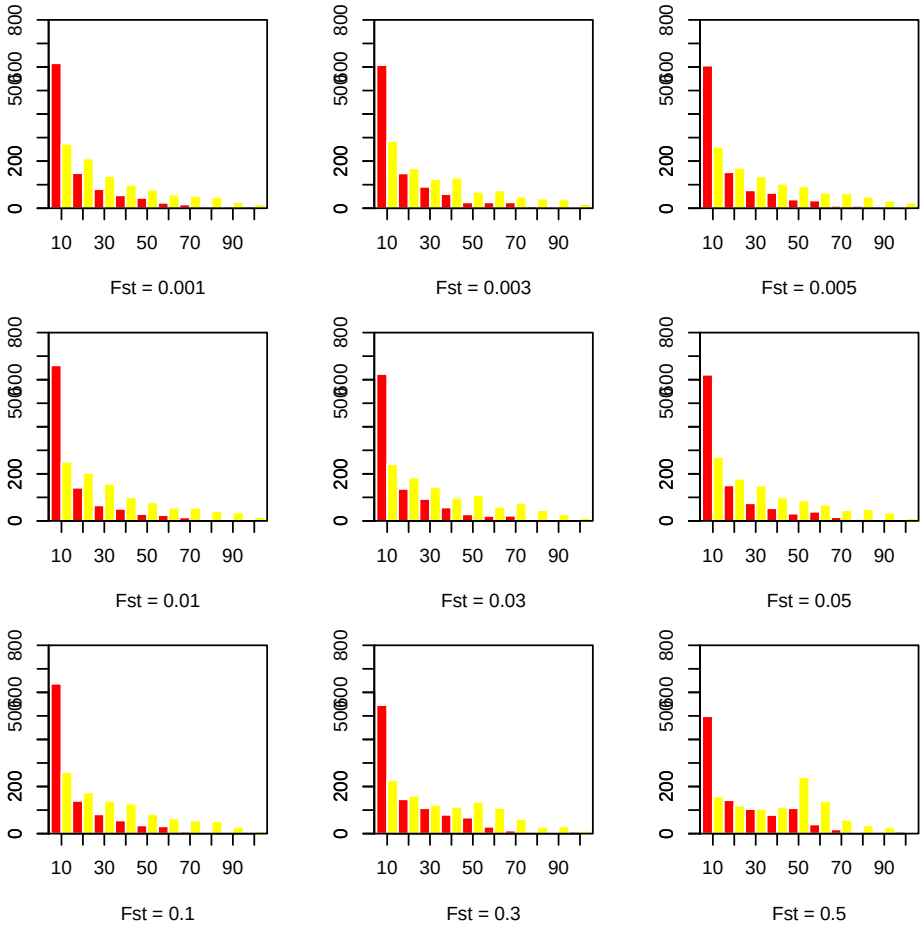


Figure 4.7: Allele frequency spectra from 1000 SNPs simulated from the migration model without any ascertainment (red bars) and with an ascertainment sample of size 2 from each of the two subpopulations (yellow bars) for the range of F_{st} values selected in figure 4.4.

4.5 Example using four subpopulations

In order to explore the similarities and difference between data from the migration and isolation models, an example with four subpopulations was conctructed and data simulated

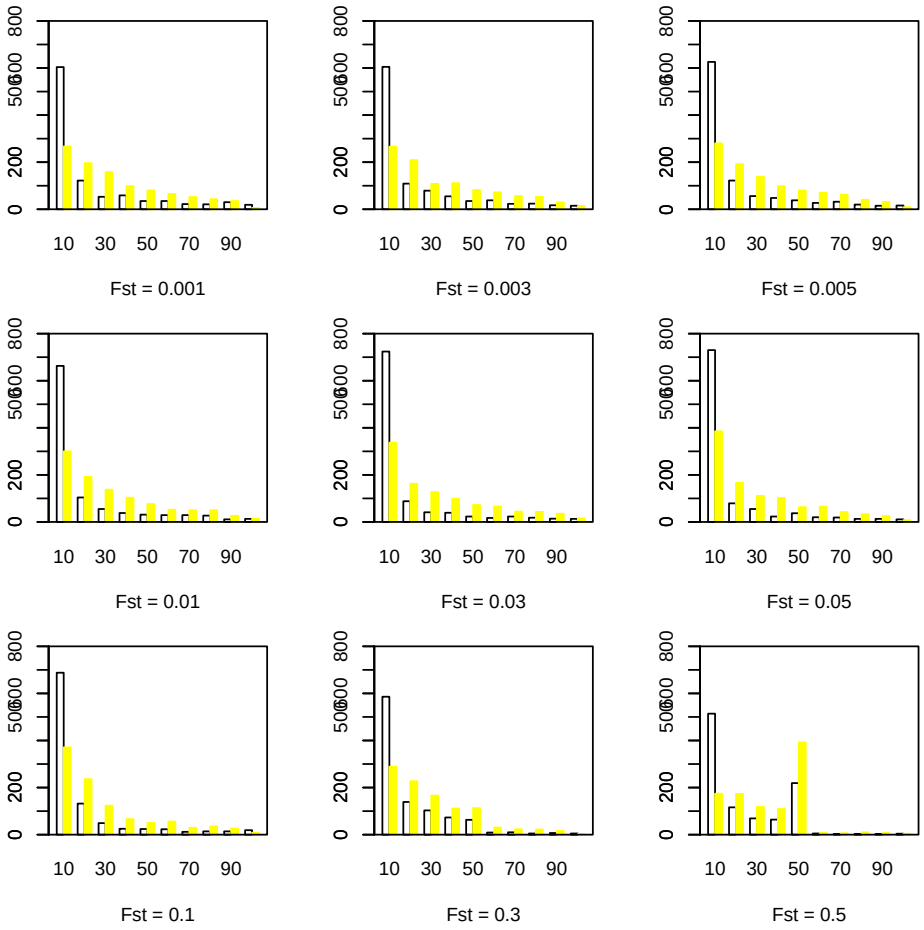


Figure 4.8: Allele frequency spectra from 1000 SNPs simulated from the isolation model without any ascertainment (white bars) and with an ascertainment sample of size 2 from each of the two subpopulations (yellow bars) for the range of F_{st} values.

under both models. Plots of components from PCA, as well as allele frequency spectra with and without ascertainment, are provided. Since the PCA method described in section 2.1.3 finds the components that are significant in capturing the structure in data, we explore how well the data projected onto the significant components are able to preserve the similarities and difference between the models.

Three symmetric migration rates, $\{m_1, m_2, m_3\}$ and three divergence times, $\{T_1, T_2, T_3\}$ were specified in the models portrayed in figure 4.10. Migration rate m_3 was chosen at a

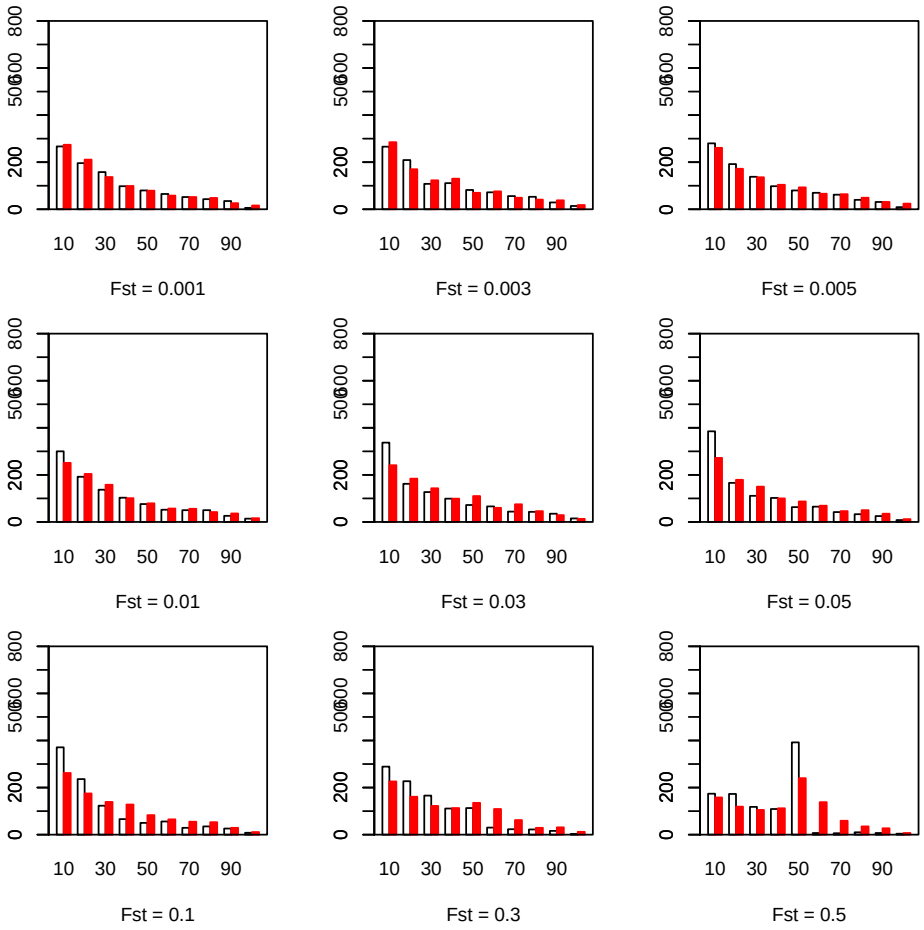
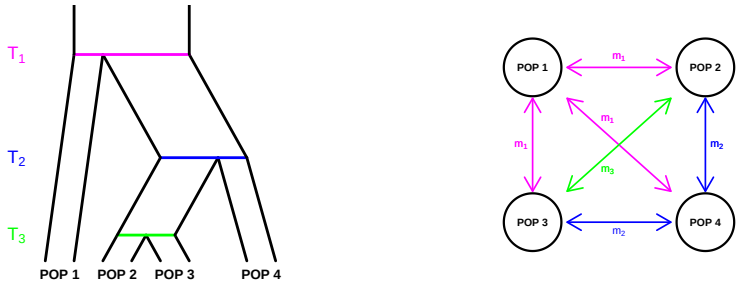


Figure 4.9: Allele frequency spectra from 1000 SNPs simulated from the isolation model (white bars) and migration model (red bars) with an ascertainment sample of size 2 from each of the two subpopulations for the range of F_{st} values.

high value, m_2 at an intermediate value and m_1 at a low value, in order to imitate the structure of figure 4.10(a), which has a recent divergence time T_3 between subpopulations 2 and 3, an intermediate divergence time T_2 between subpopulation $\{2, 3\}$ and 4 and the longest time T_1 between subpopulation 1 and the others.

1500 SNPs were simulated under each model, with sample size 50 from each of the subpopulations, $\{T_1, T_2, T_3\} = \{0.4, 0.03, 0.003\}$ and $\{m_1, m_2, m_3\} = \{0.0006, 0.008, 0.08\}$. Principal components analysis performed on the simulated data. The first two compon-



(a) (b)

Figure 4.10: Example of the migration model and the isolation model with 4 subpopulations as described in text.

ents were found to be significant and are given, for each model separately, in figure 4.11. There is very little difference between 4.11(a) and 4.11(b), suggesting that, although the derived components, from PCA, capture the population structure present in the data, they are unable to provide evidence of the underlying history of that data.

In addition, 1500 SNPs were simulated from both models with ascertainment as described in section 4.4.1 with $b = 8$, 2 from each of the 4 subpopulations. Plots of the first two components from principal components analysis are given in figures 4.11(c) and 4.11(d). Ascertainment has little effect on the first two components in this example.

Pairwise F_{st} values were computed and are reported in table 4.1. Both data sets produce similar pairwise F_{st} values with the smallest between subpopulations 2 and 3, $F_{st} = 0.02$ (isolation model) and 0.03 (migration model). The largest values lie between subpopulation 4 and the remaining three subpopulations, $F_{st} \approx 0.25$ (isolation model) and 0.18 (migration model). Similar patterns are found in the pairwise F_{st}'s from the ascertained data. Comparing F_{st} from non-ascertained and ascertained data, since ascertainment effects the allele frequencies, the values are F_{st} calculated from the ascertained data are larger than those calculated from data with no ascertainement.

Lastly, the allele frequency spectra of the four data sets, from the isolation and migration

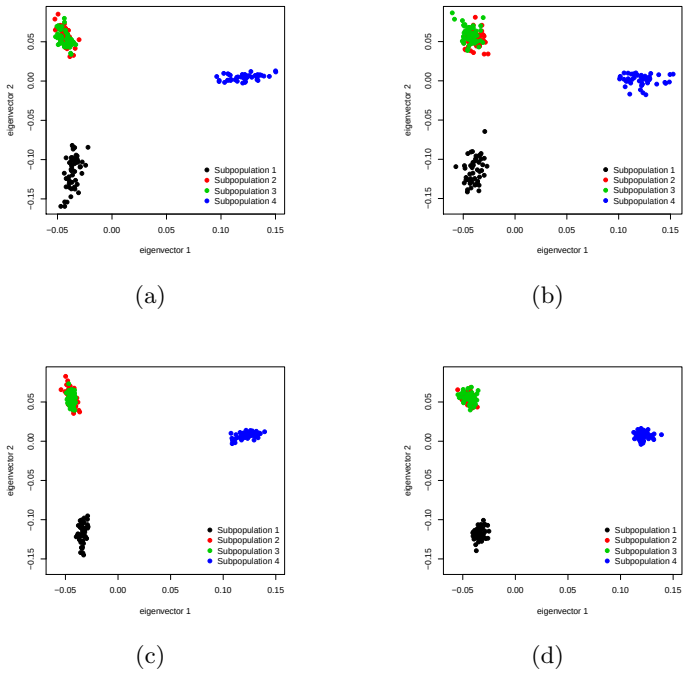


Figure 4.11: Plots of the first two components from principal components analysis from simulated data from the isolation model (left) and the migration model (right) without (top) and with (bottom) ascertainment.

models with and without ascertainment, were computed and are shown in figure 4.12. Considering the case with no ascertainment (left), the distributions are rather similar. However, the isolation model displays more SNPs with allele count less than 10 compared to the migration rate. Also, the isolation model displays a steeper decline in allele frequencies. A similar pattern is present in the case with ascertainment (right).

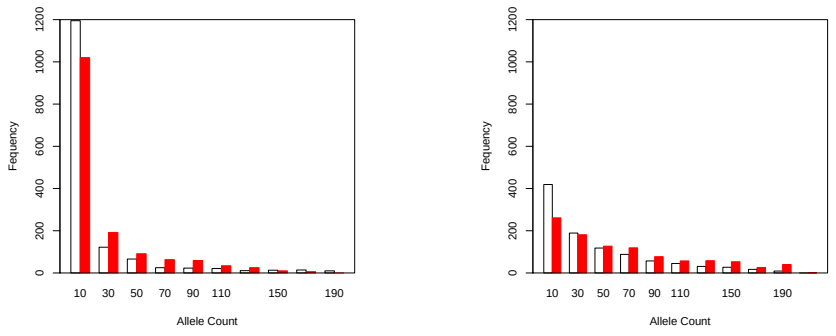
4.5.1 Projected data

Attempts have been make to infer demographic history based on projected data. For example, McVean (2009) shows that many different demographic scenarios result in similar biplots, as demonstrated in section 3.5. However, this may not be the case under certain scenarios. Patterson et al. (2006) simulated data with four subpopulations, A, B, C and D, with C an admixture of A and B and performed principal components analysis. The plot of

	1	2	3	4
1	- -	0.11 (0.13) 0.10 (0.20)	0.10 (0.14) 0.10 (0.20)	0.25 (0.35) 0.20 (0.34)
2	0.11 (0.13) 0.10 (0.20)	- -	0.02 (0.02) 0.03 (0.04)	0.26 (0.35) 0.17 (0.31)
3	0.10 (0.14) 0.10 (0.20)	0.02 (0.02) 0.03 (0.04)	- -	0.25 (0.35) 0.18 (0.31)
4	0.25 (0.35) 0.20 (0.34)	0.26 (0.35) 0.17 (0.31)	0.25 (0.35) 0.18 (0.31)	- -

Table 4.1: Pairwise F_{st} values of the four subpopulations. Values in blue are from the isolation model and values in red are from the migration model. Values in brackets are taken from data simulated with an ascertainment sample of size 2 from each subpopulation.

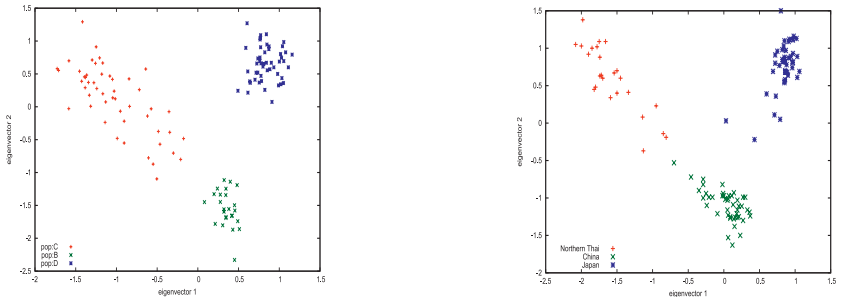
the first two components are compared to principal components analysis from three Asian populations from the Hapmap project (the International HapMap Consortium (2003)) namely China, Japan and Thailand. The results from the paper are displayed in figure 4.13. The authors find two significant components and comment on the likeness of the plot produced from the simulated data and that from the data from the HapMap project although this may not be the only demographic scenario to produce such components. In



(a) (b)

Figure 4.12: Allele frequency spectra of data simulated under the migration (red bars) and isolation (white bars) model without any ascertainment (a) and with ascertainment (b).

the simulated data, population A was not included in the principal components analysis but the authors suggest population A would correspond to a cluster lying along the same line as the red and green clusters.



(a) Simulated data (b) Data from HapMap project

Figure 4.13: Reconstruction of figures produced by Patterson et al. (2006) described in text.

In order to establish if the first few components capture the differences present in the allele frequency spectra from the data simulated from both models with four subpopulations, singular value decomposition was used to truncate the data matrix, C. The matrix C can be decomposed such that

$$C = USV^T,$$

where the columns of U are the eigenvectors of CC^T, the columns of V are the eigenvectors of C^TC and S is a diagonal matrix consisting of the ordered singular values of CC^T. By only considering a small number, k, of the largest singular values, the matrix S can be estimated. Let S_k denote the matrix with the first k diagonal entries non zero, corresponding to the first k singular values, and the remaining $n^T - k$ entries set to zero.

That is,

$$S = \begin{pmatrix} \lambda_1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & 0 & 0 & \dots & 0 \\ 0 & 0 & \lambda_k & 0 & \dots & 0 \\ 0 & 0 & 0 & \lambda_{k+1} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_{n^T} \end{pmatrix} \approx \begin{pmatrix} \lambda_1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & 0 & 0 & \dots & 0 \\ 0 & 0 & \lambda_k & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} = S_k.$$

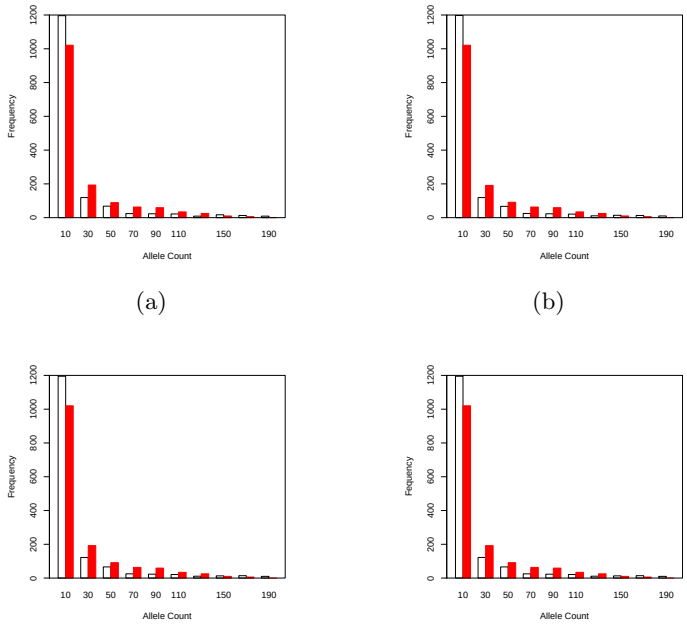
Consequently,

$$\begin{aligned} C &= USV^T \\ &\approx US_kU^T. \end{aligned}$$

Figure 4.14 illustrates the allele frequency spectra produced through estimating the matrix C with $k = 1, 2$ and 3. There is very little difference between the allele frequency spectra when $k = 1, 2, 3$ or using all the singular values. In addition, the differences in allele frequencies between the two models is mirrored in the projected data.

4.6 Summary

This chapter has presented some methods of distinguishing the two demographic models, presented by Wakeley (1996) and Nielsen and Wakeley (2001), as well as drawing attention to some problematic areas such as ascertainment. By comparing allele frequency spectra from both models for a range of F_{st} values, both with and without ascertainment, it is possible to distinguish the models by comparing particular aspects of these distributions. The remainder of this thesis aims to provide adequate estimation of the parameters of these models and then explores some descriptive statistics which, combined, are shown to be powerful in distinguishing these models.



(a) (b)

(c) (d)

Figure 4.14: Allele frequency spectra of projected data under the isolation (white bars) and migration (red bars) models. The number of components considers is 1 (a), 2 (b) and 3 (c).

Chapter 5

Estimating population parameters

In both the isolation and migration models, there are unknown parameters of interests, respectively, the population divergence times and the migration rates (figure 4.1). Attention here is primarily given to estimation of the parameters in an isolation model. Methods for estimating population divergence range from estimators based on summary statistics, maximum likelihood estimators, method of moments estimators, as well as Bayesian inference from the posterior distribution.

This chapter begins by examining estimates of the population divergence time of two subpopulations looking firstly at estimators based on F_{st} and then exploring an approximate Bayesian computation approach.

5.1 Estimating population divergence time

Given data from two subpopulations, each of haploid population size N, that diverged at some unknown time τ in the past with time measured in units of $2N$ generations, there are several ways to estimate τ.

5.1.1 Estimating population divergence using F_{st}

It is possible to estimate τ given F_{st}, as previously described in equations (2.3) and (3.9). However, there are several ways to define F_{st} and several estimators of it. One such estimator was proposed by Reynolds et al. (1983) and used by Nielsen et al. (1998), for two subpopulations with equal sample size, n:

$$\hat{F}_{st_1} = \frac{\sum_{i=1}^{L}\left[\frac{1}{2}\sum_{j=1}^{2}(p_{ij_1}-p_{ij_2})^2 - \frac{1}{2(2n-1)}\left(2-\sum_{j=1}^{2}(p_{ij_1}^2+p_{ij_2}^2)\right)\right]}{\sum_{i=1}^{L}\left[1-\sum_{j=1}^{2}p_{ij_1}p_{ij_2}\right]}, \quad (5.1)$$

where p_{ij_k} is the allele frequency of allele j at locus i in subpopulation k. Reynolds et al derived (5.1) under a "drift model". This model is an isolation model whereby allele frequencies are only affected by drift, so that is assumes no mutation occurred after time τ. Therefore, this estimator may only be suitable if it is thought that the population divergence time is relatively recent.

Another estimator of F_{st} was presented by Hudson et al. (1992):

$$\hat{F}_{st_2} = 1 - \frac{H_w}{H_T}, \quad (5.2)$$

where H_w is the average heterozygosity within a subpopulation and H_T is the average heterozygosity in the total population. $\hat{F}_{st_2}$ is used in the packages DIYABC (Cornuet et al. (2008)), described in section 5.3.2.

The last commonly-quoted estimator to be considered is:

$$\hat{F}_{st_3} = \frac{1}{L}\sum_{i=1}^{L}\frac{\hat{\sigma}_i^2}{\bar{p}_i(1-\bar{p}_i)}., \quad (5.3)$$

where $\bar{p}_i$ is the mean allele frequency across subpopulations at SNP i and $\hat{\sigma}_i^2$ is the variance of allele frequencies across subpopulations at that SNP, for $i = 1, \ldots, L$.

Figure 5.1 presents central 95% confidence bands for $\hat{F}_{st}$ for a range of τ values using the

three estimators. The three methods produce similar values of F_{st}. $\hat{F}_{st_2}$ has the largest estimated standard error. For τ close to zero, $\hat{F}_{st_1}$ and $\hat{F}_{st_2}$ produce similar values with $\hat{F}_{st_3}$ sitting slightly higher. For values of τ around one, $\hat{F}_{st_1}$ begins a steady plateau whereas $\hat{F}_{st_3}$ and $\hat{F}_{st_2}$ both continue to increase. Values of F_{st} one would expect to find within the human population may be in the range $[0.0001, 0.4]$ as seen in table 2.1. Within this range, the three estimators are fairly consistent.

Given an estimate of F_{st}, it is possible to estimate τ in at least two separate ways. Firstly, the method outlined in section 3.5 gives:

$$\hat{\tau} \quad = \quad \frac{\hat{F}_{st}}{1 - \hat{F}_{st}}. \tag{5.4}$$

Secondly, suppose there are D subpopulations, each of haploid population size N, that diverged at time t in the past (in generations so that $t = DN\tau$) and since evolved independently under the Wright-Fisher model as illustrated in figure 5.2.

At a particular locus in a single population suppose two alleles, A and a, are present. Let P_j and p_j denote the number of copies of allele A and the corresponding allele frequency in generation j and let $P_0 = i$ and $p_0 = \frac{i}{N}$. In the first generation, the number of copies of allele A is a random sample, of size N, from the population in generation 0. Hence,

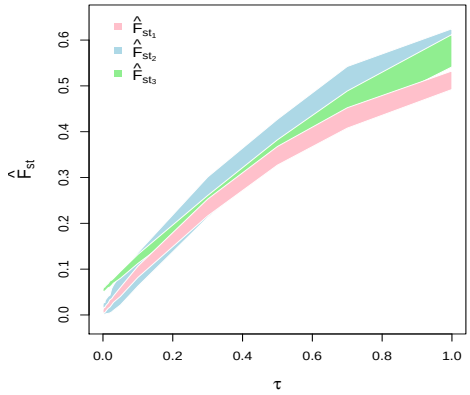


Figure 5.1: Central 95% confidence bands for the three estimates, $\hat{F}_{st_1}$, $\hat{F}_{st_2}$ and $\hat{F}_{st_2}$, of F_{st} for a range of values of $\tau \in [0, 1]$.

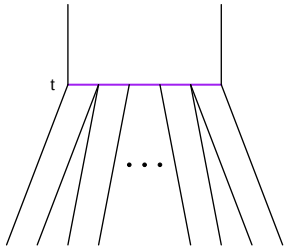


Figure 5.2: Example of D subpopulations diverging at time t in the past.

$P_1 \sim Bi(N, p_0)$. Generally, $P_k \sim Bi(N, p_{k-1})$ and

1. $E(p_k) = p_0$,
2. $Var(p_k) = p_0(1-p_0)\left(1-(1-\frac{1}{N})^k\right)$ for $k = 0, 1, 2, \ldots$.

Consider the case $k = 1$, then $E(p_1) = E\left(\frac{P_1}{N}\right) = \frac{1}{N}(Np_0) = p_0$. When $k = 2$,

$$E(p_2) \quad = \quad E(E(p_2|p_1)) = E\left(E\left[\frac{P_2}{N}|p_1\right]\right) = E(p_1) = p_0.$$

Iteratively, $E(p_k) = E(E(p_k|p_{k-1})) = E(p_{k-1}) = \cdots = p_0$.

Generally,

$$Var(p_k) = p_0(1-p_0)\left(1-\left[1-\frac{1}{N}\right]^k\right). \tag{5.5}$$

For instance, when $k = 1$, $Var(p_1) = \frac{1}{N^2}Var(P_1) = \frac{1}{N}p_0(1-p_0)$. When $k = 2$, since $Var(x) = E(Var[x|y]) + Var(E[x|y])$,

$$\begin{aligned} Var(p_2) &= E\left(Var(p_2|p_1)\right) + Var\left(E(p_2|p_1)\right) \\ &= E(\frac{1}{N^2}Var(P_2|p_1)) + Var(p_1) \end{aligned}$$

$$\begin{aligned} &= \frac{1}{N}E(p_1 - p_1^2) + Var(p_1) \\ &= \frac{1}{N}(E(p_1) - Var(p_1) - E(p_1)^2) + Var(p_1) \\ &= \frac{1}{N}p_0(1-p_0) + \frac{1}{N}\left(1 - \frac{1}{N}\right)p_0(1-p_0) \\ &= p_0(1-p_0)\left(1 - \left[1 - \frac{1}{N}\right]^2\right). \end{aligned}$$

Assume (5.5) to be true for $k = k - 1$, and so

$$Var(p_{k-1}) = p_0(1-p_0)\left(1 - \left[1 - \frac{1}{N}\right]^{k-1}\right).$$

Therefore,

$$\begin{aligned} Var(p_k) &= E\left(Var(p_k|p_{k-1})\right) + Var\left(E(p_k|p_{k-1})\right) \\ &= \frac{1}{N}p_0(1-p_0) + \left(1 - \frac{1}{N}\right)Var(p_{k-1})) \\ &= \frac{1}{N}p_0(1-p_0) + \left(1 - \frac{1}{N}\right)p_0(1-p_0)\left(1 - \left[1 - \frac{1}{N}\right]^{k-1}\right) \\ &= p_0(1-p_0)\left(1 - \left[1 - \frac{1}{N}\right]^{k}\right) \\ &\approx p_0(1-p_0)(1 - e^{-k/N}). \end{aligned}$$

Therefore, given p_0 in any subpopulation at some locus at generation t:

$$\begin{aligned} E(p_t) &= p_0, \\ Var(p_t) &\approx p_0(1-p_0)(1 - e^{-\frac{t}{N}}), \end{aligned}$$

since each subpopulation evolves according to the Wright-Fisher model and the subpopulations evolve independently,

$$\begin{aligned} E(\bar{p}_t) &= p_0, \\ Var(\bar{p}_t) &\approx \frac{1}{D}p_0(1-p_0)(1 - e^{\frac{t}{N}}), \end{aligned} \tag{5.6}$$

where $\bar{p}_t$ is the mean of p_t across the subpopulations. Consider the expectation of estimator (5.3):

$$E(\hat{F}_{st_3}) = E\left[\frac{1}{L}\sum_{i=1}^{L}\frac{\hat{\sigma}_i^2}{\bar{p}_i - \bar{p}_i^2}\right] \approx \frac{1}{L}\sum_{i=1}^{L}\frac{E(\hat{\sigma}_i^2)}{E(\bar{p}_i - \bar{p}_i^2)}$$

where $\hat{\sigma}_i^2 = \sum_{d=1}^{D}\frac{p_{id}^2}{D} - \bar{p}_i^2$. At a single locus at time t, from (5.6),

$$\begin{aligned}
E(\bar{p}t) - E(\bar{p}_t^2) &= E(\bar{p}_t) - E(\bar{p}_t)^2 - Var(\bar{p}_t) \\
&\approx p_0(1-p_0)\left(1 - \frac{1}{D}(1 - e^{-\frac{t}{N}})\right). \\
E(\hat{\sigma}_t^2) &= \sum_{d=1}^{D}\frac{1}{D}E(p_{d_t}^2) - E(\bar{p}_t^2) \\
&= E(p_t^2) + Var(p_t) - E(\bar{p}_t)^2 - Var(\bar{p}_t) \\
&= p_0(1-p_0)(1 - e^{-\frac{t}{N}})\left(1 - \frac{1}{D}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
E(\hat{F}_{st_3}) &\approx \frac{p_0(1-p_0)(1 - e^{-\frac{t}{N}})(1 - \frac{1}{D})}{p_0(1-p_0)\left(1 - \frac{1}{D}(1 - e^{-\frac{t}{N}})\right)} \\
&= 1 - \frac{e^{-\frac{t}{N}}}{1 - \frac{1}{D}(1 - e^{-\frac{t}{N}})}.
\end{aligned}$$

As $D \to \infty$, $E(\hat{F}_{st_3}) \to 1 - e^{\frac{t}{N}}$, as described by Cavalli-Sforza (1969). Measuring time in DN generations, at time τ,

$$E(\hat{F}_{st_3}) = 1 - \frac{e^{-D\tau}}{1 - \frac{1}{D}(1 - e^{-D\tau})}.$$

Rearranging for τ leads to the following estimator in terms of an estimator of F_{st}:

$$\hat{\tau} = -\frac{1}{D}ln\left[\frac{(1 - \frac{1}{D})(1 - F_{st})}{(1 - \frac{1}{D}) + \frac{F_{st}}{D}}\right].$$

In particular, when $D = 2$,

$$\hat{\tau} \quad = \quad -\frac{1}{2}\log\left[\frac{1-\hat{F}_{st}}{1+\hat{F}_{st}}\right]. \tag{5.7}$$

For a range of F_{st} values, figure 5.3 illustrates the degree to which the two estimators (5.4) and (5.7) coincide. For small F_{st} values, the two functions are comparable but they diverge from each other as F_{st} increases with (5.4) producing higher values of $\hat{\tau}$.

In order to assess how well the two F_{st}-based estimators perform in estimating τ, data were simulated under the isolation model with $D = 2$ for a range of τ's. F_{st} was estimated using (5.1) and τ estimated using (3.10) and (5.7). For each τ, 100 data sets were simulated and F_{st} estimated. The results are presented in figure 5.4. The plots show central 95% confidence bands for $\hat{\tau}$ for each combination of F_{st}-based estimator and the two estimators of τ. Figure 5.4(a) shows τ estimated by (5.7) and figure 5.4(b) shows τ estimated using (3.10). Estimator (5.4) follows more closely with the true τ values with F_{st} estimated using (5.1) but, it diverges for larger values of τ as may be expected since Reynolds et al. (1983) explains this estimate is appropriate for recent population divergence times.

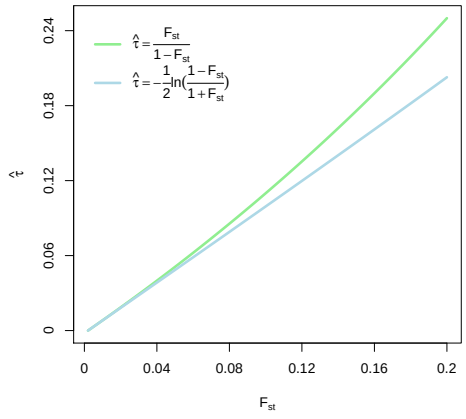
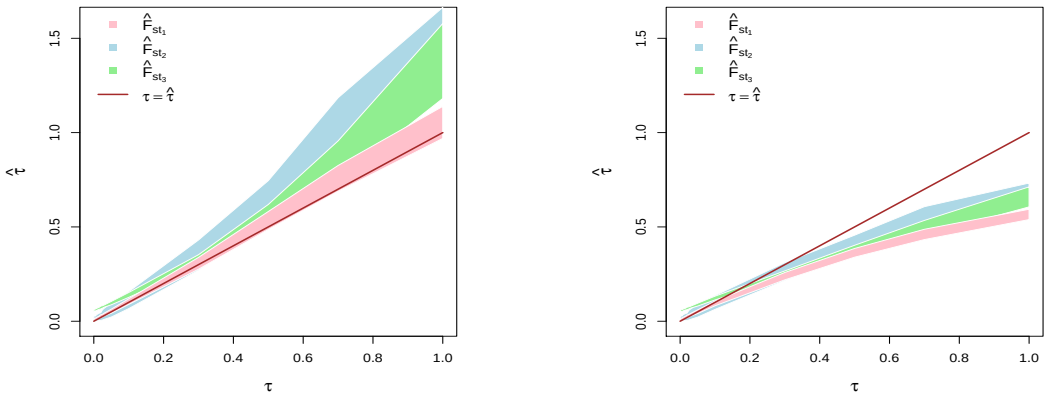


Figure 5.3: Comparison of two F_{st}-based estimators of τ described in equations (3.10) and (5.7).



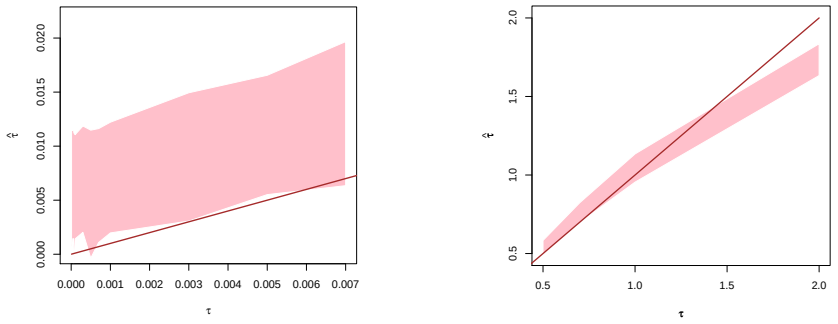
(a) (b)

Figure 5.4: 95% central confidence bands of τ for a range of true τ values in $[0, 1]$ from data simulated under the isolation model. τ is estimated used (5.7) (a) and (3.10) (b). The brown line shows $\tau = \hat{\tau}$.

5.1.2 Difficulties with F_{st}-based estimators

There are limitations in F_{st}-based estimators. More precisely, the proposed estimator performs poorly when τ is close to zero as illustrated in figure 5.5(a). The brown line shows $\tau = \hat{\tau}$ and lies (almost) completely below the confidence bands. In addition to overestimating τ for relatively recent population split times, more ancient population split times are underestimated, as shown in figure 5.5(b). On the other hand, the range of F_{st} values actually encountered between human populations is approximately $(0, 0.35]$ (as shown in table (2.1)). In this range, the corresponding $\hat{\tau}$ values are approximately $(0, 0.5]$ as exemplified by the pink bands in figure 5.1. As a result, it is unproductive to focus on values outwith this range, and certainly $\tau > 1$.

To understand why F_{st}-based estimators perform in this manner, think of range of F_{st} compared to that of τ or m. F_{st} is in the range $[0, 1]$, whereas the migration rate m and divergence time τ are in the range $[0, \infty)$. Figure 5.6 demonstrates the relationships using equations (3.10) and (3.11). As $\tau \to \infty$ there is little change in F_{st} with F_{st} approaching one as shown in figure 5.6(a). Likewise, figure 5.6(b) shows as $m \to \infty$ there is little change in F_{st} with F_{st} approaching zero. Therefore, for increasing values of τ there is

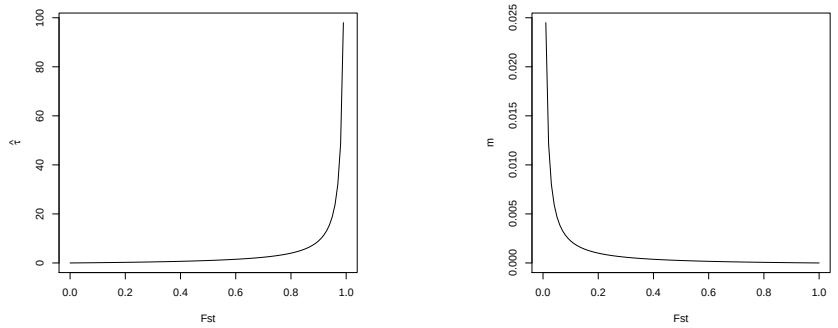


(a) (b)

Figure 5.5: 95% central confidence bands for $\hat{\tau}$ in the range (0,0.007] (a) and [0.5,2] (b). The brown line shows $\tau = \hat{\tau}$ and τ estimated using (5.4).

little change in F_{st}, hence, little change in $\hat{\tau}$. This may account for the underestimation of $\hat{\tau}$ in figure 5.5(b).

Generally, F_{st}-based estimators do not provide the best accuracy but are convenient due to their simplicity. Maximum likelihood estimators, as described in section 1.2, have been shown by Nielsen et al. (1998) to provide more accurate estimates based on a comparison of the estimated standard errors for the two estimators.



(a) (b)

Figure 5.6: Relationship between F_{st} and (a) $\hat{\tau}$ and F_{st} and (b) $\hat{m}$ using equations (3.10) and (3.11).

5.2 Software for estimating population parameters

There are several computer programs available that estimate population parameters such as migration rates and population divergence times. For example, MIGRATE, originally introduced by Beerli and Felsenstein (2001), estimates migration rates from sequence data. IM (with extensions IMa and IMa2) which fits the isolation with migration model, originally with two subpopulations, as described by Nielsen and Wakeley (2001), and was later extended to incorporate more than two subpopulations, as described by Hey (2010). In this model, Hey assumes that the structure can be represented in a phylogenetic tree, with migration occurring between all pairs populations in existence at each time point as described in section 4.2. However, the remainder of this chapter will concentrate on Bayesian approaches to parameter estimation.

5.3 Bayesian approaches to parameter estimation

Taking a Bayesian approach, estimating parameters from a model parameterised by ϕ given observed data x_{obs} involves, for example, sampling from the posterior distribution $p(\phi|x_{obs})$. If it is not possible to do this directly, Gelman et al. (2004) provide details of methods used to simulate from $p(\phi|x_{obs})$. One method simulates draws from $p(\phi|x_{obs})$ by evaluating $p(\phi|x_{obs})$ on a grid of values $\phi_1, \ldots, \phi_{N_{sim}}$ covering a broad range of the parameter space of ϕ. Another method is to sample from a distribution $g(\phi)$ for which there exists $M > 0$ such that $p(\phi|x_{obs})/g(\phi) \leq M$ for all ϕ. The idea behind this method is depicted in figure 5.7. The curve $Mg(\phi)$ completely contains $p(\phi|x_{obs})$. The algorithm samples ϕ' from $g(\phi)$ and accepts it as a draw from $p(\phi|x_{obs})$ with probability $p(\phi'|x_{obs})/Mg(\phi')$.

Markov-chain Monte-Carlo methods devise a Markov chain with stationary distribution equal to the target distribution $p(\phi|x_{obs})$. Starting from an initial value ϕ_0, for example, drawn from the prior distribution $\pi(\phi)$, the algorithm iteratively sample ϕ', given the previous draw, for long enough until it is thought that the draws are from the target

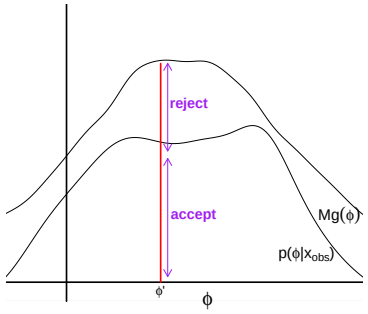


Figure 5.7: Example of rejection sampling.

distribution.

Bayesian inference relies on the computation of the likelihood function $p(x_{obs}|\phi)$, but, as shown in section 1.2, this may not always be feasible given a reasonably large sample of SNP data. One method of overcoming this issue is to replace the full data by summaries of the data. The remaining sections in this chapter discuss a Bayesian approach to estimating population parameters using summary statistics.

5.3.1 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) has been developed over the last 10 years in population genetics in order to estimate parameters of interest when the likelihood cannot be estimated or is computationally intractable. Given a sample of SNP data with an unknown genealogy then the likelihood function requires integrating over all possible genealogies. Figure 1.3 shows that in a sample of size 6 from a neutral model, there are 2700 possible branching structures. For parameter ϕ to be estimate, this method computes a set of summary statistics denoted by S_{obs} from observed data and then, loosely speaking, simulates data under the proposed model for a range of values of ϕ computing S_{sim} and accepts values of ϕ when $\eta(S_{obs}, S_{sim}) \approx 0$ for some distance measure $\eta(\cdot, \cdot)$ and so replacing the full likelihood function $p(x|\phi)$ by $p(S(x)|\phi)$ for observed data x.

More precisely, Robert et al. (2011) provide the underlining ABC procedure. Given observed data $x \in \mathcal{X}$, summary statistics S and distance measure η, then samples $\{\phi_1, \ldots, \phi_m\}$ are simulated from

$$f(\phi, S(z)|x) \quad \propto \quad \pi(\phi) f(S(z)|\phi) \mathbb{I}_{A^{\eta}_{\epsilon,x}}(z),$$

where

$$A^{\eta}_{\epsilon,x} \quad = \quad \{z \in \mathcal{X} | \eta\{S(z), S(x)\} \leq \epsilon\},$$

and

$$\mathbb{I}_{A^{\eta}_{\epsilon,x}}(z) \quad = \quad \begin{cases} 1, & \text{if } z \in A^{\eta}_{\epsilon,x} \ ; \\ 0, & \text{otherwise,} \end{cases}$$

is an indicator function and ϵ chosen to be 'close' to zero. The performance of this method depends on the choice of η and S. Stephens (2007) suggests that using statistics that are directly affected by the parameter of interest will increase the efficiency of this method. However, it is useful to consider not only how informative a statistic is in parameter estimation but also how informative the statistic is given the set of statistics already included in the analysis. Joyce and Marjoram (2008) assess the question: given statistics $S_1, \ldots, S_{k-1}$, then how beneficial will an additional statistic S_k be? They compare the log likelihood function $l\{S_1, \ldots, S_{k-1}|\phi\}$ to $l\{S_1, \ldots, S_{k-1}, S_k|\phi\}$, the difference being the latter function contains the additional term $l\{S_k|S_1, \ldots, S_{k-1}, \phi\}$. They attached a score δ_k to each new statistic S_k given $S_1, \ldots, S_{k-1}$ and ϕ. Once a score falls below a pre-specified threshold then the corresponding statistic is not included in the inference. Their methodology was to find a set of approximately sufficient statistics. For example, if $S_1, \ldots, S_{k-1}$ are sufficient statistics then $l(S_k|S_1, \ldots, S_{k-1}, \phi) = l(S_k|S_1, \ldots, S_{k-1})$ and so S_k adds no additional information about ϕ. Barnes et al. (2011) take an information theory approach to finding the set of statistics that minimise the consequent loss of information that results from incorporating summary statistics into any inference as opposed to using the full data. Given

a larger set $\mathcal{S}$ of statistics they devise an algorithm to find the minimal set $U \subset \mathcal{S}$ such that $\mathcal{S}$ contains no more information about ϕ given U. In order to assess the performance of the set of statistics $\mathcal{S}$, Nunes and Balding (2010) implemented the square root of the sum of squared errors, RSSE;

$$\text{RSSE} = \left(\frac{1}{n_a}\sum_{i=1}^{n_s} I_i|\phi_i - \phi_{obs}|^2\right)^{1/2},$$

where n_s is the number of simulations in the ABC algorithm, n_a is the number of accepted draws and

$$I_i = \begin{cases} 1, & \text{if } \phi_i \text{ is an accepted draw;} \\ 0, & \text{otherwise.} \end{cases}$$

For each subset $U \subseteq \mathcal{S}$, the ABC algorithm is repeated n_o times and RSSE computed with the intention to find the set U^* which minimises

$$\text{MRSSE} = \sum_{j=1}^{n_o} \text{RSSE}_j.$$

With attention given to methods of selecting the optimal set of statistics, the focus now moves to describing the evolution of ABC algorithms. Beaumont et al. (2002) proposed a method of inference based on summary statistics as an extension to a method described by Tavaré et al. (1997) who were interested in estimating the time to the most recent common ancestor of a sample conditioning on the observed number of segregating sites S_n. They specified priors π_N and π_μ on the population size N and mutation rate μ per site per generation. In a genealogy of length L, $S_n \sim Poi(\frac{1}{2}L\theta)$. The algorithm used was:

1. Simulate N and μ from π_N and π_μ.
2. Simulate data under the standard coalescent model to find coalescent waiting times $\{W_1, \ldots, W_n\}$.
3. Find the time to the most recent common ancestor T_{MRCA} and the total length of the genealogy T_{total} using $\{W_1, \ldots, W_n\}$.

4. Keep T_{MRCA}, N and μ with probability

$$\begin{aligned} u &\propto Pr(S_n = k | T_{total} = L) \\ &= \frac{1}{k!} e^{-\frac{L\theta}{2}} \left(\frac{L\theta}{2}\right)^k. \end{aligned}$$

This process simulates samples $\{T_{MRCA_1}, \ldots, T_{MRCA_{n_{sim}}}\}$ from distribution of T_{MRCA}, given N and $\mu | S_n = k$ and an estimate of T_{MRCA} is $\hat{T}_{MRCA} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} T_{MRCA_i}$.

Pritchard et al. (1999) studied an ancestral population of size N_A that underwent a period of exponential expansion at rate r that began at time t_0. They used three summary statistics, namely the average heterozygosity $\bar{H}$, the number of distinct haplotypes, n, and the mean of the variance in repeat numbers (using microsatellite data), $\bar{V}$. The rejection algorithm implemented was:

1. Calculate observed statistics, $s = \{n, \bar{H}, \bar{V}\}$.
2. Simulate ϕ' from the prior distribution π_ϕ with $\phi' = \{N'_A, t'_0 \text{ and } r'\}$.
3. Given ϕ', simulate a genealogy under the required model and then microsatellite data.
4. Computed the statistics $s' = \{n', \bar{H}', \bar{V}'\}$ on the simulated data.
5. If $|s'_i - s_i| < \delta$ for all $i = 1, 2$ and 3, then accept ϕ'.

Again, this process produces draws from the joint distribution of the parameters of interest ϕ given the observed summary statistics s. Beaumont et al. (2002) extended this method by following the first four steps to produce draws (ϕ_i, s_i) for $i = 1, \cdots, n_s$ and weighting draw ϕ_i by $|s_i - s|$. They assume that ϕ_i can be modelled by

$$\phi_i = \alpha + (s_i - s)^T \beta + e_i \qquad \text{for } i = 1, \cdots, n_s.$$

for some coefficients α and β to be estimated and independent errors e_i Since $E(\phi|S = s) = \alpha$, the authors write

$$\phi^* = \phi - (s_i - s)\hat{\beta}$$

as a random sample from $p(\phi|S = s)$. Although the relationship between $\{\phi_i\}$ and $\{s_i\}$ may not be linear, the authors assume that the relationship is linear locally to s. They estimate (α, β) by minimising

$$\sum_{i=1}^{n_s} (\phi_i - \alpha - (s_i - s)\beta)^2 K_\delta(|s_i - s|)$$

for band–width δ (synonymous with the last step of the above algorithm) and kernel function K_δ.

Fearnhead and Prangle (2012) describe ABC algorithms based on rejection and MCMC sampling methods. If interest lies in estimating a parameter ϕ given observed data x_{obs}, the Metropolis-Hastings algorithm, detailed by Gelman et al. (2004), aims to make draws $\phi_1, \ldots, \phi_{N_{sim}}$ from the target distribution $p(\phi|x_{obs})$. Given a prior distribution $\pi(\phi)$, the algorithm begins by drawing ϕ_0 from $\pi(\phi)$ and, for $i = 1, \ldots, N_{sim}$, a draw $\phi^{'}$ is made from a proposal distribution $q(\phi^{'}|\phi_{i-1})$ and is accepted with probability $\min(\alpha, 1)$ where

$$\alpha = \frac{p(\phi^{'}|x_{obs})/q(\phi^{'}|\phi_{i-1})}{p(\phi^{i-1}|x_{obs})/q(\phi^{i-1}|\phi^{'})}.$$

The algorithm described by Fearnhead and Prangle requires specification of statistics $S(x)$ with $s_{obs} = S(x_{obs})$ and a density $K(x)$. It is initiated by simulating ϕ_0 from $\pi(\phi)$ and calculating s_{obs}, and for $i = 1, \ldots, n_{sim}$,

1. Simulate $\phi^{'}$ from proposal distribution $q(\phi^{'}|\phi_{i-1})$.

2. Simulate data x_{sim}, from the required model, and calculate $S(x_{sim}) = s_{sim}$.

3. With probability

$$\alpha = \min\left[1, \frac{K\{(s_{sim} - s_{obs})/h\}}{K\{(s_{i-1} - s_{obs})/h\}} \frac{\pi(\phi^{'})g(\phi_{i-1}|\phi^{'})}{\pi(\phi_{i-1})g(\phi^{'}|\phi_{i-1})}\right]$$

set $\phi_i = \phi^{'}$, otherwise set $\phi_i = \phi_{i-1}$, for some pre-specified bandwidth $h > 0$.

This algorithm may be used to estimate the population divergence time τ by setting

- $S = F_{st}(=F$ say),
- a $N(\tau_{i-1}, \sigma^2)$ proposal distribution,
- a $Uni(0.00001, 0.7)$ prior distribution for τ, and,
- $K(x) = \begin{cases} 1, & \text{if } |x| \leq 0.5; \\ 0, & \text{otherwise.} \end{cases}$

Therefore,

$$\alpha = \min\left[1, \frac{K\{(F_{sim} - F_{obs})/h\}\pi(\tau)g(\tau_{i-1}|\tau)}{K\{(F_{i-1} - F_{obs})/h\}\pi(\tau_{i-1})g(\tau|\tau_{i-1})}\right]$$

and

$$K\{(F_{sim} - F_{obs})/h\} = \begin{cases} 1, & \text{if } |(F_{sim} - F_{obs})/h| < \frac{1}{2}; \\ 0, & \text{otherwise.} \end{cases}$$

Necessarily, $K\{(F_{i-1} - F_{obs})/h\} = 1$ and

$$\frac{\pi(\tau)}{\pi(\tau_{i-1})} = \begin{cases} 1, & \text{if } \tau \in (0.00001, 0.7); \\ 0, & \text{otherwise.} \end{cases}$$

Since the proposal distribution is normal, $g(\tau_{i-1}|\tau) = g(\tau|\tau_{i-1})$ and $\alpha = 1$ or 0. The ABC_MCMC algorithm details how to estimate τ.

Algorithm 1 (ABC_MCMC)**.**

For $i = 1, \ldots, N_{sim}$:

1. *Simulate* τ *from* $N(\tau_{i-1}, \sigma^2)$.
2. *Simulate data using* τ *and calculate* $F_{st} = F_{sim}$.
3. *If* $\alpha = 1$ *then set* $\tau_i = \tau$ *and* $F_i = F_{sim}$. *If* $\alpha = 0$ *set* $\tau_i = \tau_{i-1}$ *and* $F_i = F_{i-1}$.

This algorithm produces a set of draws $\tau_1, \tau_2, \ldots, \tau_{N_{sim}}$. Figure 5.8 shows density estimates of simulated τ's for a range of true τ values using this algorithm. On each plot, the red dot shows the true value of τ. For most values of τ, the density peaks around the true values with a few exceptions. By taking the lower and upper 2.5% of each distribution, figure 5.9 shows 95% credible bands for the range of τ values. This algorithm estimates τ well; the bands contain of the line $\tau = \hat{\tau}$ and are narrower than those produced by the F_{st}-based estimator (figure 5.4).

5.3.2 ABC packages

There is a battery of summary statistics that are extensively used in population genetics. Several packages use ABC to infer aspects of demographic history, for example, DIY ABC (Cornuet et al. (2008)), PopABC (Lopes et al. (2009)) and ABCtoolbox (Wegmann et al. (2010)). The statistics employed in these packages are discussed in chapter 6. Although each program aims to make inferences about population parameters via summary statistics, the methods adopted in each are different.

Cornuet et al. (2008) model the demographic history of a sample by firstly specifying the population size, population divergence times (backwards in time) and population admixture (backwards in time, a population splits into two other populations in the sample). Data is then simulated under this pre-specified history and a set of summary statistic

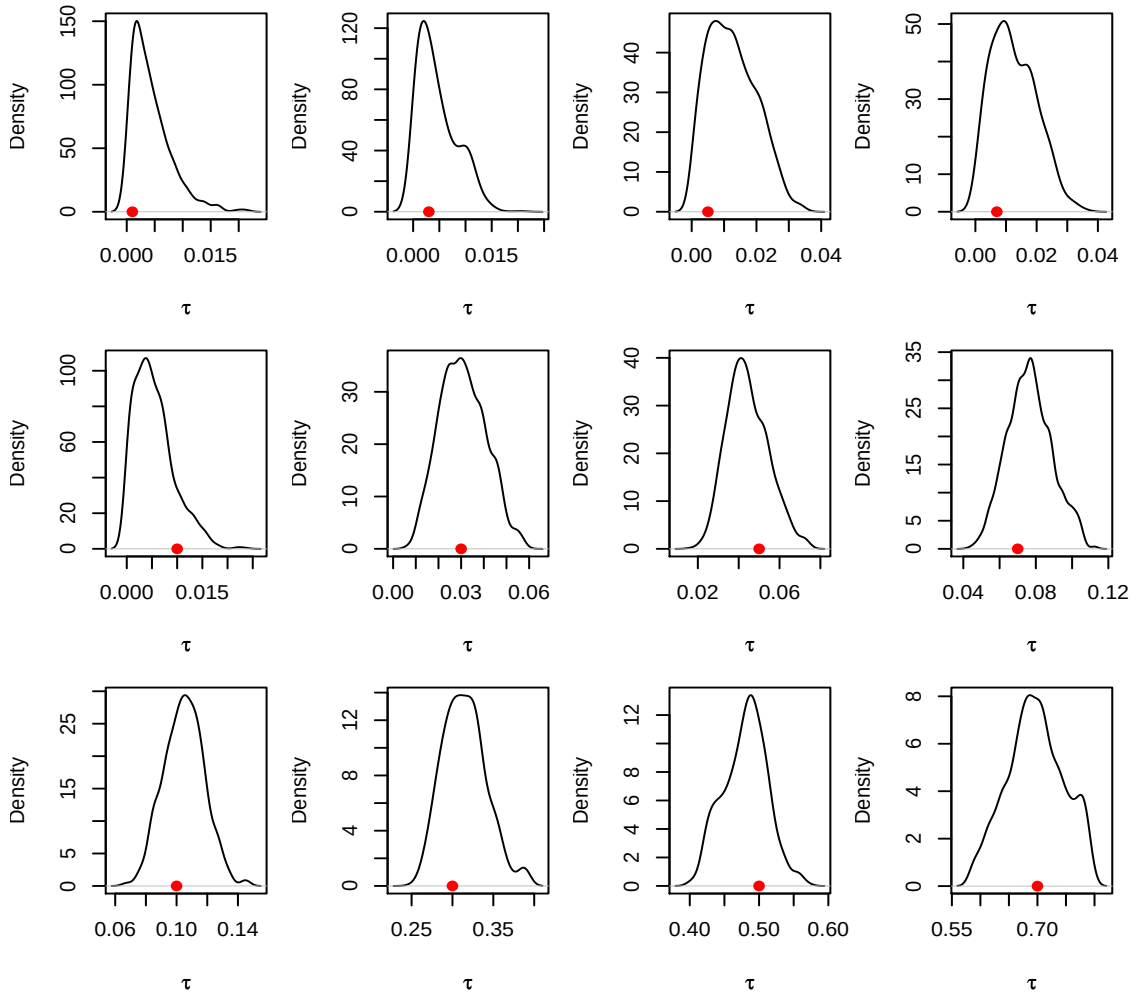


Figure 5.8: Density plots of simulated τ's for a range of true τ values (red dot) using ABC_MCMC algorithm.

computed. Simulated data set i is compared to the observed data using distance measure

$$d_i = \sqrt{\sum_{j=1}^{m} \frac{(s_{ij} - s_{obs_j})^2}{Var_j}},$$

where m is the number of statistics, Var_j the variance of the jth statistic across statistics, s_{ij} is the value of statistic j in simulation i and s_{obs_j} is the observed value of statistic j. This program then uses the algorithm given by Beaumont et al. (2002) to estimate the

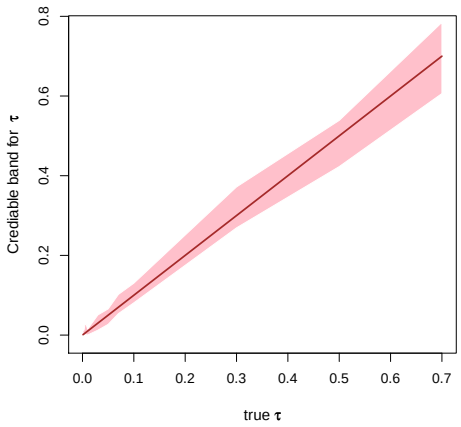


Figure 5.9: 95% credible bands for τ, and the line of equality.

parameters.

Lopes et al. (2009) fit the isolation with migration model presented originally by Nielsen and Wakeley (2001) and described in section 4.2. It aims to estimate the tree topology (treated as a categorical variables with several possible topologies), population size, population split times, migration, mutation and recombination rates, by implementing a rejection based algorithm.

Once specifying a model (to simulate from) and a set of summary statistics, Wegmann et al. (2010) use partial least squares (PLS) to make linear combinations of the summary statistics in order to find an optimal set of statistics. This methods was motivated by Joyce and Marjoram (2008) who showed that although it may be beneficial to include as many summary statistics thought to be informative about the parameters of interest, adding too many contribute more noise. A further discussion of the matter is given in section 6.2.4. PLS regression has two main steps as described by Boulesteix and Strimmer (2007). The first is a dimension reduction step. It is assumed that there are q continuous response variables $Y_1, \ldots, Y_q$ and p continuous explanatory variables $X_1, \ldots, X_p$ with observed data $y_i = \{y_{i_1}, \ldots, y_{i_q}\}$ and $x_i = \{x_{i_1}, \ldots, x_{i_p}\}$ for $i = 1, \ldots, n_{sim}$. Wegmann et al. (2010) consider the summary statistics as the explanatory variables and the parameters of interest

the response variables. The general underlying model of PLS is to write

$$\begin{aligned} X &= TP^T + E \text{ and} \\ Y &= TQ^T + F, \end{aligned}$$

where T is a matrix of latent components, P and Q are matrices of dimension $p \times c$ and $q \times c$, respectively, and E and F are error matrices. As with principal components analysis, linear combinations of the columns of the matrix X, of dimension $n \times p$, can be found that are independent and contain most of the variability in the data. For example, in the notation of Boulesteix and Strimmer (2007),

$$T_j = w_{1_j} X_1 + \ldots + w_{p_j} X_p, \qquad \text{for } j = 1, \ldots, c,$$

where $T_1, \ldots, T_c$ are the components, c is the chosen number of components that are thought to explain most of the variation in the data and the columns of the matrix $W = \{w_{ij}\}$ of dimension $p \times c$ are such that the latent components explain the variation in the explanatory and response data. Therefore,

$$T = XW, \text{ and hence } X = TW^T.$$

The second stage is to model the data. As in the case of multiple linear regression, the matrix Q^T can be estimated by $Q^T = (T^T T)^{-1} T^T Y$. In particular, Y can be modelled by

$$Y = TQ^T + F = XWQ^T + F$$

and so a least squares estimation of the matrix of regression coefficients $WQ^T = W(T^T T)^{-1} T^T Y$. This approach was implemented to find a minimal number of independent statistics, with the authors also suggesting this procedure recovers an optimal set of summary statistics.

5.4 Model selection

The probability of a model given data provides the natural Bayesian tool to assess which model provides the better fit to the data. To illustrate, data were simulated from the isolation model with two subpopulations diverging at time 0.7. The ABC_MCMC algorithm was used to estimate $p(\tau|F_{st})$, as illustrated in figure 5.8, but also $p(m|F_{st})$, the posterior distribution of the migration rate between the two subpopulations given F_{st} under the (misspecified) migration model. Figure 5.10 shows the posterior density estimate of m given F_{st} with the red dot showing the posterior mean value.

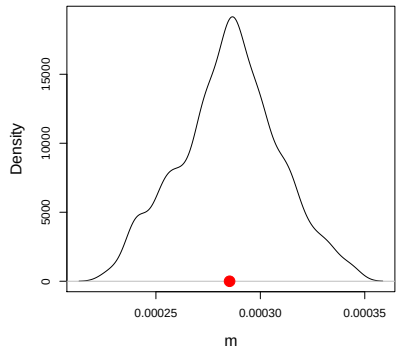


Figure 5.10: Estimate of $p(m|\hat{F}_{st})$ using the ABC_MCMC algorithm.

Robert et al. (2011) provide an algorithm to calculate the probability of a model given the data, or summaries of the data. In this application, let M_1 denote the isolation model and M_2 denote the migration model. The algorithm produces a vector $m = (m_1, \ldots, m_{N_{sim}})$. At the ith step, m^* is generated from $\pi(M)$ the prior distribution on the models, for example $p(M = M_1) = p(M = M_2) = 0.5$, and, using a draw $\phi_{m^*} \sim \pi(\phi_{m^*})$, data are simulated under model m^* and the summary statistics S_{sim} computed. These steps are repeated until the distance between S_{obs} and S_{sim} is less that ϵ and they set $m_i = m^*$. They estimate the probability of model j, for $j = 1, 2$, given S_{obs} as

$$Pr\{M_j|S_{obs}\} = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \mathbb{I}_{m_i=j}.$$

Using the algorithm, with $N_{sim} = 1000$ and both models a priori equally likely, the probabilities of the isolation and migration models are estimated to be:

$$\begin{aligned} Pr\{\text{migration}|F_{st}\} &= 0.24 \\ Pr\{\text{isolation}|F_{st}\} &= 0.76. \end{aligned}$$

If M_1 and M_2 correspond to the isolation and migration models, respectively, then the Bayes factor

$$\begin{aligned} B_{12} &= \frac{Pr\{M_1|S_{obs}\}p(M = M_1)}{Pr\{M_2|S_{obs}\}p(M = M_2)} \\ &= 3.2, \end{aligned}$$

providing evidence in favour of the isolation model over the migration model.

In population genetics, it is often the case that the statistics implemented in ABC (for example, Tajima's D) are not sufficient which presents problems when estimating the likelihood function $p(x|\phi)$ since the observed data x are replaced by a statistic $S(x)$. In particular, Bayesian model selection methods require the likelihood function to be evaluated as discussed by Robert et al. (2011) and Barnes et al. (2011).

The issue is made explicit by Barnes et al. (2011). The authors define sufficiency in terms of the likelihood function. A statistic S is sufficient if

$$f\Big(x|S(x),\phi\Big) = g\Big(x|S(x)\Big),$$

where $f\Big(x|S(x),\phi\Big)$ is the likelihood of data x given parameter ϕ and statistic $S(x)$ and $g(x|S(x))$ is the probability of the data given the statistic, independent of ϕ. For two prospective models, M_1 and M_2, the posterior probabilities of the models $p(M_i|x)$ for $i = 1, 2$ given the data are estimated. The model comparison considers the ratio

$$B_{12} = \frac{p(x|M_1)}{p(x|M_2)} = \frac{p(M_1|x)\pi(M_1)}{p(M_2|x)\pi(M_2)}.$$

In particular, in the case of k models, the posterior probability of the ith model is

$$Pr(M_i|x) = \frac{\pi(M_i)\int_{\Theta_{M_i}} p(x|\phi_{M_i})\pi(\phi_{M_i})d\phi_{M_i}}{\sum_{j=1}^{k}\pi(M_j)\int_{\Theta_{M_j}} p(x|\phi_{M_j})\pi(\phi_{M_j})d\phi_{M_j}}, \tag{5.8}$$

where $\pi(M_i)$ is the prior probability that the data are from model M_i with parameters $\phi_{M_i} \in \Theta_{M_i}$. If the observed data x is replaced by simulated data y then as, $\epsilon \to 0$,

$$\begin{aligned}\int_{y\in\mathcal{X}} p(y|\phi_{M_i})\pi(\phi_{M_i})\mathbb{I}_{\{\eta(x,y)<\epsilon\}}dy &\propto \int_{y\in\mathcal{X}} \pi(\phi_{M_i}, y|x)\mathbb{I}_{\{\eta(x,y)<\epsilon\}}dy \\ &\to p(\phi_{M_i}|x),\end{aligned} \tag{5.9}$$

where

$$\mathbb{I}_{\{\eta(x,y)<\epsilon\}} = \begin{cases} 1, & \text{if } \eta(x,y) < \epsilon; \\ 0, & \text{otherwise.}\end{cases}$$

As a results, as $\epsilon \to 0$, (5.8) can be approximated by

$$\frac{\pi(M_i)\int_{\Theta_{M_i}}\int_{y\in\mathcal{X}} p(y|\phi_{M_i})\pi(\phi_{M_i})\mathbb{I}_{\{\eta(x,y)<\epsilon\}}dyd\phi_{M_i}}{\sum_{j=1}^{k}\pi(M_j)\int_{\Theta_{M_j}}\int_{y\in\mathcal{X}} p(y|\phi_{M_i})\pi(\phi_{M_i})\mathbb{I}_{\{\eta(x,y)<\epsilon\}}dyd\phi_{M_j}}.$$

In the context of ABC where the observed data x is replaced by a summary of the simulated data $S(y)$, Barnes et al. (2011) note that (5.9) becomes

$$\int_{S(\mathcal{X})} p(S(y)|\phi_{M_i})\pi(\phi_{M_i})\mathbb{I}_{A^{\eta}_{\epsilon,x}}(y)dS(y).$$

If S is a sufficient statistic, then $p(S(y)|\phi_{M_i}) \propto p(y|\phi_{M_i})$. However if S is not sufficient, as is often the case, then $Pr(M_i|x)$ cannot be approximated by

$$\frac{\pi(M_i)\int_{\Theta_{M_i}}\int_{S(\mathcal{X})} p(S(y)|\phi_{M_i})\pi(\phi_{M_i})\mathbb{I}_{A^{\eta}_{\epsilon,x}}(y)dS(y)d\phi_{M_i}}{\sum_{j=1}^{k}\pi(M_j)\int_{\Theta_{M_j}}\int_{S(\mathcal{X})} p(S(y)|\phi_{M_j})\pi(\phi_{M_j})\mathbb{I}_{A^{\eta}_{\epsilon,x}}(y)dS(y)d\phi_{M_j}}.$$

Therefore, although using non–sufficient statistics to estimate the joint distribution of $f(\phi, z|x)$ for $\{z \in \mathcal{X}|\eta(S(x), S(z)) < \epsilon\}$ is valid, there are problems in computing Bayes factors. Barnes et al. (2011) argue that the problems with using insufficient statistics in model selection are reflected in parameter estimation evidenced in an example estimating the mean from a $N(\mu, 1)$ distribution. They consider four separate statistics to estimate μ namely the sample mean, variance, the minimum value and the maximum value. They show using the sample mean produces the most accurate results.

5.4.1 Model misclassification

Csilléry et al. (2012) wrote the abc package in R, (R Development Core Team (2008)), which performs parameter estimation and model selection. In particular, the model selection procedure estimates the posterior probability of a model given the observed summary statistics. They measure the success of ABC in model selection by estimating misclassification rates. Fitting the observed data to each of the prospective models, they store the summary statistics' values at each iteration. For each model they select one set of summary statistics, to be treated as the observed data, and find the posterior probability of each model using the values from the leftover simulations and assign the data to the model with the highest probability. This is repeated a specified number of times. The function output consists of a confusion matrix with the diagonal entries showing the number of correctly classified simulations and the off-diagonal entries showing the number of misclassifications and also a matrix of posterior probabilities. The diagonal entries are the average posterior probabilities of the simulations where the correct model gave the highest posterior probability.

Repeating this process 1000 times for each model, the following matrices are produced:

	Isolation	Migration
Isolation	648	352
Migration	390	610

(a) Confusion matrix.

	Isolation	Migration
Isolation	0.58	0.42
Migration	0.48	0.52

(b) Average posterior probabilities.

The isolation and migration model were correctly identified 648 and 610 out of the 1000 repetitions. The average posterior probability of those simulations from the isolation model with $Pr\{\text{isolation} \mid S_{obs}\} > Pr\{\text{migration} \mid S_{obs}\}$ was 0.58 and the corresponding probability from the migration model was 0.52. This provides evidence, although not overwhelming, that the ABC model selection procedure using only F_{st} can assign the correct model the highest posterior probability. Inevitably, the performance of this method depends on how informative the summary statistics are about the full data. For example, Nielsen et al. (1998) showed F_{st} was not sufficient in estimating the population divergence time and so these results may not be too surprising.

5.5 Summary

This chapter has presented different estimators of the population divergence time τ between two subpopulations in the isolation model. In particular, the ABC_MCMC algorithm described by Fearnhead and Prangle (2012) estimates τ well as shown in figure 5.9 but also the F_{st}-based estimator (3.10), with F_{st} estimated by (5.1), was shown in figure 5.4 to also estimate τ well although this method produces wider 95% confidence bands. This chapter also touched upon distinguishing between the migration and isolation models taking an ABC approach. If it is possible to implement sufficient statistics then Bayes factors may be used to calculate the posterior probabilities of both models but Robert et al. (2011) showed that if the summary statistics used are not sufficient then this method may be invalid. It is clear that some carefully chosen statistics do contain valuable information in relation to parameter estimation and model selection and so this line of enquiry is adapted in the next few chapters.

Chapter 6

A hypothesis test for demography

There are two general approaches to statistical inference. As described in the previous chapter, the Bayesian approach depends on the evaluation of the likelihood function, $p(x|\phi)$, and when this function is computationally unattainable for example, in the presence of complex missing data such as the genealogy of a locus. One strategy is to replace the data x by summaries. However, problems have been highlighted with this ABC approach to model comparison (e.g by Robert et al. (2011)), for example when the summaries are not sufficient. On the other hand, by considering a frequentist approach, our attention is deflected to testing the hypothesis that the observed data are consistent with a particular demographic model. This chapter begins by finding a set of test statistics that may maximise the power of a frequentist hypothesis test to distinguish the isolation and migration models. With a suitable set of test statistics, of size m, having been identified, the observed value of each is individually compared to values simulated under the null model (one of the two models under consideration). Formally, m null hypotheses, $H_{0_1}, \ldots, H_{0_m}$, are tested, one concerning the value of each test statistic, with the overall intention of testing the global hypothesis equal to the intersection of the individual m hypotheses:

$$H_0 = \bigcap_{i=1}^{m} H_{0_i}.$$

This chapter also provides a brief discussion of incorporating a multiple testing correction to control the type I error rate.

6.1 Summary statistics

Hypothesis testing about the demography of samples of SNP data based on summary statistics obviously requires statistics that contain as much information as possible to distinguish the models. We have already seen that the allele frequency spectrum may be useful in distinguishing the isolation and migration models particularly for larger values of F_{st}. The computer packages described in section 5.3.2 employ a wide variety of summary statistics. However, some of them are only applicable to certain types of data. Examples of these statistics are listed below:

Statistic 1. The mean number of alleles across loci.

Statistic 2. The variance in allele length (relevant to loci such as microsatellites).

Statistic 3. The number of segregating sites, S (relevant to sequence data).

Statistic 4. The number of different haplotypes (relevant to phased sequence data).

Statistic 5. Heterozygosity. Let k be the number of alleles and p_i be the frequencies of the ith allele at a particular locus (for SNP data $k = 2$), then

$$\bar{H} = 1 - \sum_{i=1}^{k} p_i^2.$$

Statistic 6. Estimator of Nm based on heterozygosity. This is used in popABC by Lopes et al. (2009) and uses the heterozygosity of all the populations H_a and the

heterozygosity within subpopulations H_w:

$$Nm = \frac{H_w}{1 + H_a - H_w}.$$

Statistic 7. F_{st}.

Statistic 8. The mean pairwise difference π. Let k_{ij} be the number of difference between the ith and jth chromosome, then

$$\pi = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i}^{n} k_{ij}.$$

The mean pairwise difference is often calculated within and between subpopulations denoted by π_W and π_B, respectively.

Statistic 9. Tajima's D_t statistic. Let n be the total sample size and $\sum_{i=1}^{n-1} a_1 = \frac{1}{i}$, then

$$D_t = \frac{\pi - S/a_1}{\sqrt{\widehat{Var}(\pi - S/a_1)}}.$$

Statistic 10. Number of SNPs with allele count 1, η_1. This statistic may be plausible in distinguishing the two models as shown in figure 4.4. The isolation model had higher η_1 than the migration model for values of $F_{st} > 0.01$ (approximately).

Statistic 11. Variance of allele counts. In figure 4.4, the migration model shows greater variability in allele counts than the isolation model.

Statistic 12. The average allele count. Taken to be the average of the counts in the allele frequency spectrum. Figure 4.4 shows isolation model has a higher value than the migration model.

Statistic 13. $\eta_{max} = max\{\eta_i : i = 1, \ldots, n-1\}$, where η_i is the number of SNPs with allele count i. Again, fig 4.4 shows, for larger value of F_{st}, there is an increase in allele counts equal to n, the sample size of each subpopulation, in both models.

In particular, the isolation model shows a larger allele count equal to n than the migration model.

Statistic 14. Fu and Li's F^* and D^* statistics defined as

$$\begin{aligned} F^* &= \frac{\pi - (n-1)\eta_1/n}{\sqrt{\widehat{Var}(\pi - (n-1)\eta_1/n)}}, \\ D^* &= \frac{S/a_1 - (n-1)\eta_1/n}{\sqrt{\widehat{Var}(S/a_1 - (n-1)\eta_1/n)}}. \end{aligned}$$

Statistics 1–4 are calculated in popABC and DIY ABC (Cornuet et al. (2008)) but are not applicable to SNP data. There are a few other statistics not mentioned above that are also not applicable to SNP data. Since the parameter in the model of interest is estimated using a function of F_{st}, and this estimate used to simulate data under the null model, F_{st} is not included. Also, F^* and D^* are similar to D_t in that they are all function of S, π and η_1, so the test will exclude the former two statistics and include the latter four. The remaining statistics are considered, with attention next given to whether they successfully distinguish the migration and isolation models.

6.1.1 Initial comparison of summary statistics

In order to compare the distributions of the summary statistics listed, a small migration rate and a large migration rate were considered. The distributions under the migration model were compared to the distributions of the statistics under an isolation model with population split time $\hat{\tau}$ estimated to produced similar F_{st} values to the migration models.

For a small migration rate between the two subpopulation, $m_{small} = 0.0001$, 1000 SNPs were simulated with samples from the two subpopulations each of sample size 10 and statistics 8–13 and 5 were computed. F_{st} was estimated using (5.1) and a population split time was estimated using (3.10). Data were simulated under the isolation model, with the same sample sizes as for the migration model, and the set of summary statistics were

computed. This process was repeated 100 times and the distributions of the summary statistics plotted separately.

Figure 6.1 gives the allele frequency spectra of data simulated under the two models with the isolation model shown in green and the migration model shown in red.

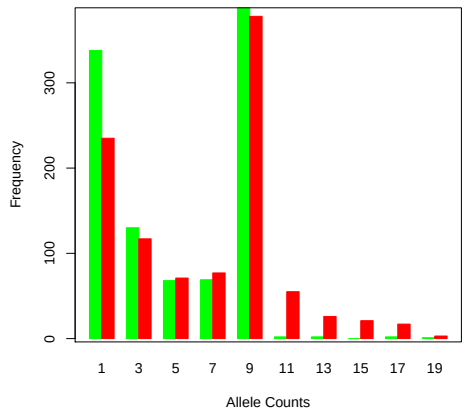


Figure 6.1: Allele frequency spectra from 1000 SNPs simulated from the isolation model (green bars) and the migration model (red bars).

In this sample, the isolation models exhibits a greater η_1 and η_{max} with $\eta_1 = 224$ and $\eta_{max} = 363$ in the isolation model and $\eta_1 = 153$ and $\eta_{max} = 326$ in the migration model. The isolation model has a lower average frequency of 5.6 compared to 7.1 in the migration model and the migration model slightly has a higher variance, of 18.4 compared to 15.1 in the isolation model. By repeating the simulating process 100 times, we explore whether the distributions of the statistics are different under the two models. For each of the 100 simulations, F_{st} was computed under both models and their distributions presented in figure 6.2. The two distributions are very similar. The distributions of the other summary statistics from both models are given in figure 6.3. Of the nine statistics, the average heterozygosity appears least able to distinguish the models. This is perhaps not surprising as F_{st} is a function of $\bar{H}$.

This investigation was repeated with a larger migration rate, $m = 0.1$. As previously

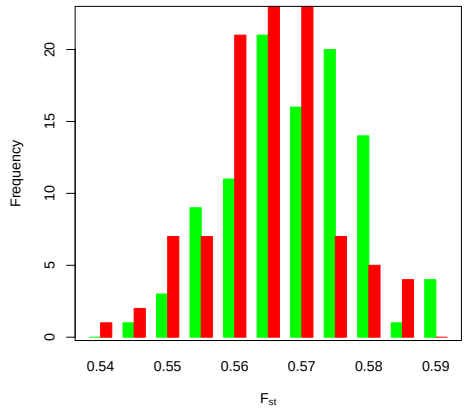


Figure 6.2: Histograms of F_{st} values from data simulated under the isolation (green bars) and migration (red bars) models with m=0.0001.

noted, as the migration rate increases the data resembles that from a single panmictic population more and more. The distributions of the nine summary statistics are given in figure 6.4. As expected, all of the statistics produced similar distributions under both models and so, distinguishing the models is more difficult if there is a higher migration rate between the two subpopulations or, equivalently, a recent population divergence time.

6.2 Hypothesis test

Given observed SNP data from two subpopulations, it is of interest to test if the data show signs of a population divergence followed by isolation or shows signs of recurrent migration events between subpopulations. Assuming the population of interest experienced one of the two events, the following hypotheses are tested:

$$\begin{aligned} H_0 &: \text{data from the isolation model.} \\ H_1 &: \text{data not from the isolation model.} \end{aligned} \tag{6.1}$$

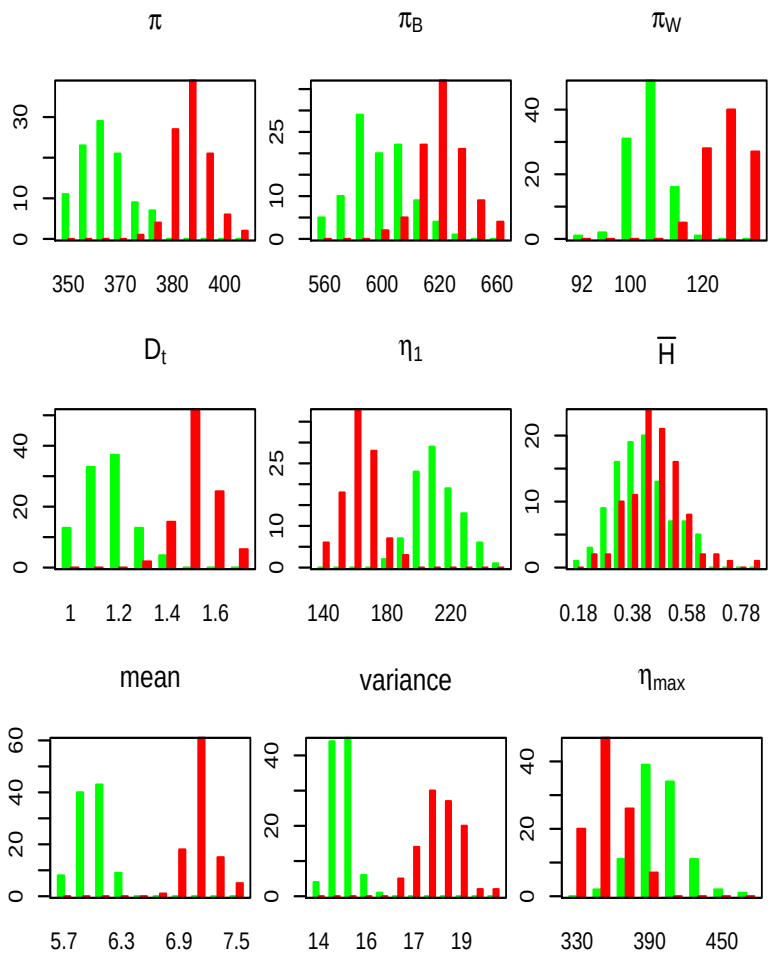


Figure 6.3: Histograms of summary statistics from data simulated under the isolation (green bars) and migration (red bars) models.

Given the prior restriction to the two models of isolation and migration, H_1 is equivalent to "data from the migration model". The choice of which model plays the role of H_0 is essentially arbitrary.

In order to test H_1 against H_0, the set of summary statistics $S = \{S_1, \ldots, S_m\}$ form the basis of m separate tests. For a given statistic, S_i, let S_{obs_i} denote the observed value and $\bar{S}_{iso_i}$ denote the expected value of the summary statistic consistent with an isolation

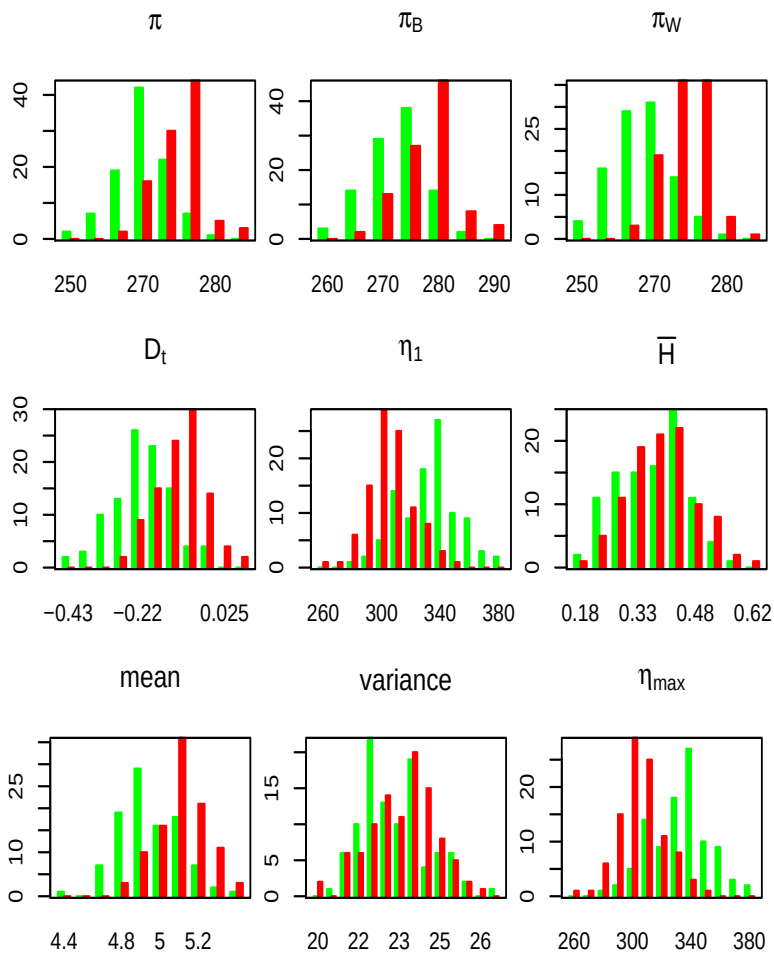


Figure 6.4: Histograms of summary statistics from data simulated under the isolation (green bars) and migration (red bars) models with the migration rate fixed as m=0.1 (in $2N$ generations).

model for some τ. The ith hypothesis to be test is:

$$H_{0_i} \quad : \quad S_{obs_i} = \bar{S}_{iso_i}$$
$$H_{1_i} \quad : \quad S_{obs_i} \neq \bar{S}_{iso_i}.$$

To test the global null hypothesis that the data are consistent with an isolation model,

the following global hypothesis is tested:

$$H_0 = \bigcap_{i=1}^{m} H_{0_i}.$$

Therefore, in order to accept the isolation model, each H_{0_i} needs to be accepted.

6.2.1 Calculating p-values

For each statistic, S_i, an empirical p-value is calculated. That is, given S_{obs_i}, the probability of observing a value more extreme than S_{obs_i} under the null model is estimated by simulating data under the isolation model. An example of a possible density estimation of the distribution of S_i is given in figure 6.5. In a two-tailed test at a 5% significance level, the red lines denote the boundaries of a rejection region in this hypothesis test. Suppose $S_{obs_i} = 254.8$, shown by the light green dot. Then, if $S_{obs_i} > \bar{s}$, y is found such that $f(S_{obs_i}) = f(y)$ (shown as the dark green dot) and the empirical p-value is

$$p = \frac{\#\{s_i > S_{obs_i}\}}{N_{sim}} + \frac{\#\{s_i < y\}}{N_{sim}},$$

where N_{sim} is the number of simulations. If $S_{obs_i} < \bar{s}$, then

$$p = \frac{\#\{s_i < S_{obs_i}\}}{N_{sim}} + \frac{\#\{s_i > y\}}{N_{sim}}.$$

In this example, $y = 233.6$ and $p = 0.02$.

In this thesis, each hypothesis is a two sided test. There might be a case for using one-sided tests. In particular, note that the distributions of π, π_W, π_B, D_t, the mean allele count and the variance of allele counts from the migration model lie above the distributions from the isolation model whereas the distributions of the η_1 and η_{max} under the migration models lie below the isolation model as shown in both figure 6.3 and figure 6.4.

Since several hypotheses are tested at the same time, a multiple–testing correction is made.

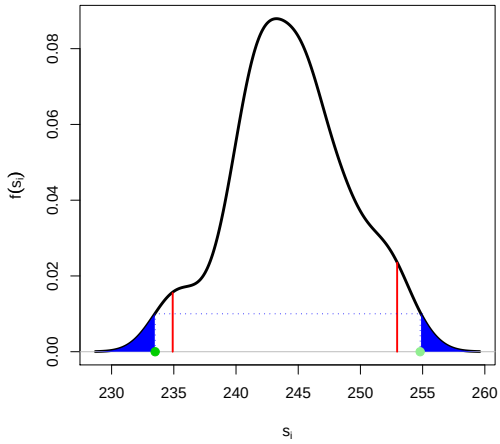


Figure 6.5: Density of statistic S_i. Red lines denote the lower and upper 2.5% of the distribution.

The next section compares several methods of correction.

6.2.2 Multiple comparisons procedures

Typically, with any hypothesis test, the type I error rate is controlled by a pre-determined significance level α. When testing a single hypothesis at significance level α, a p-value, p, is calculated to be the probability of observing a value more extreme that the observed statistic and, if $p < \alpha$, the null hypothesis is rejected. To assess the performance of the test, the probability of falsely rejecting a true null hypothesis, predetermined by α, and the probability of correctly rejecting a false null hypotheses are evaluated. Simultaneously testing multiple hypotheses will inflate the type I error rate, unless each test becomes more stringent.

There are several methods for multiple hypothesis testing, many of which are outlined by Bretz et al. (2011), Chang (2011) and Hommel et al. (2011), that simultaneously test p-values from multiple test statistics. There are also different approaches to multiple testing including single-step and stepwise procedures. Stepwise procedures either begin with the

smallest ordered p-value (step–down procedure) or the largest (step–up procedure). In the case of the step–down procedure, each hypothesis is tested from smallest to largest (adjusted) p-values and once a hypothesis is accepted, using a predetermined threshold, the remaining hypotheses (with larger corresponding p-values) are accepted. Likewise, the step–up procedure begins with the largest p-value and once a hypothesis is rejected, the remaining hypotheses are rejected.

Since interest lies with testing the global hypotheses, as opposed to taking into account which particular hypotheses are rejected, a detailed review of single-step procedures is provided.

Let $p_1, \ldots, p_m$ denote the p-values from the m hypotheses and $p_{(1)} \leq \cdots \leq p_{(m)}$ be the ordered p-values. It is conventional firstly to test the global hypothesis and then consider each hypothesis individually to find significant results. However, in this case, interest primarily lies in the global hypothesis. Hommel et al. (2011) describe a general global test in which H_0 is rejected if

$$p_{(k)} \quad \leq \quad b_k \text{ for at least one } k,$$

where $b_1, \ldots, b_m$ are chosen such that $mb_1 + \sum_{i=2}^{m} (b_i - b_{i-1}) \frac{m}{i} = \alpha$. For example, if $m = 1$ then $b_1 = \alpha$. If $m > 2$ the each $b_k < alpha$. One of the more classical approaches to multiple comparisons is the Bonferroni correction. In this case, $b_1 = \cdots = b_m = \frac{\alpha}{m}$ and H_0 is rejected if

$$p_i \quad \leq \quad \frac{\alpha}{m} \qquad \text{for any } i = 1, \ldots, m.$$

Simes (1986) proposed a modified version of the Bonferroni correction by rejecting H_0 if

$$p_{(i)} \quad \leq \quad \frac{i\alpha}{m} \qquad \text{for any } i = 1, \ldots, m,$$

or equivalently, if any $\frac{mp_{(i)}}{i} \leq \alpha$. Therefore, the global hypothesis is reject if $\tilde{p} < \alpha$ where

$$\tilde{p} = \min_{1 \leq i \leq m} \left\{ \frac{m}{i} p_{(i)} \right\}.$$

Bretz et al. (2011) demonstrate how the Simes test is more powerful than the Bonferroni test by considering the case $m = 2$. In this instance

$$\begin{aligned} H_0 &= H_{0_1} \cap H_{0_2} \\ H_1 &= H_{1_1} \cup H_{1_2}. \end{aligned}$$

The Simes method rejects H_0 if either $p_{(1)} < \frac{\alpha}{2}$ or $p_{(2)} < \alpha$. The Bonferroni method rejects H_0 if either p_1 or $p_2 < \frac{\alpha}{2}$. Therefore, the Simes rejection region contains the Bonferroni rejection region.

Another variation of this test was proposed by Hommel et al. (2011) which rejects H_0 if

$$p_{(i)} \leq \frac{i\alpha}{mC_m} \qquad \text{for any } i = 1, \ldots, m,$$

where $C_m = \sum_{i=1}^{m} (i^{-1})$. Figure 6.6 compares the rejection regions of the three corrections when $m = 2$ and $p_{(1)} \leq p_{(2)}$. The left hand side plot shows the Bonferroni rejection region, the middle plot show the Simes rejection region and the right hand side plot shows the Hommel rejection region. In each, the shaded area shows the combination of p_1 and p_2 for which the null hypothesis is rejection. The Simes' method shows the largest rejection region.

In this context, the statistics tested have an unknown correlation structure and there are no known distributional results. Therefore, the multivariate procedure employed should be valid for dependent statistics. For the three hypothesis tests presented in this section, results hold for specific cases of dependent variables under some distributional results.

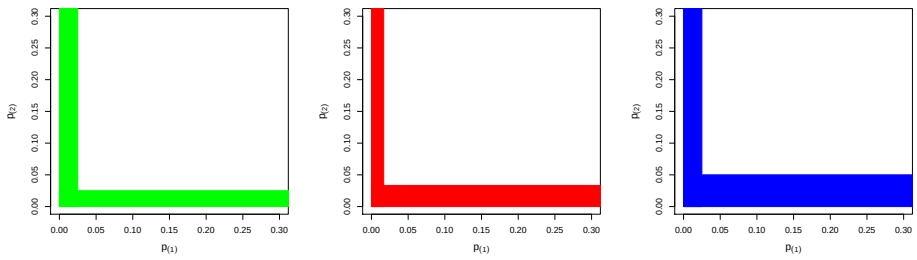
The Bonferroni method is the most simplistic of the proposed procedures but is the most

conservative. In his paper, Simes (1986) showed, for independent p-values $p_1, \ldots, p_m$,

$$Pr\left(p_{(j)} > \frac{j\alpha}{m} \mid j = 1, \ldots, m\right) \quad = \quad 1 - \alpha.$$

By simulating test statistic values $T_1, \ldots, T_m$ from a multivariate normal distribution $N_m(0, \Omega)$ with Ω an $m \times m$ covariance matrix with diagonal entries equal to one and off–diagonal entries equal to ρ (for some $0 < \rho < 1$), Simes found that his improved method preserved the type I error for $\rho \approx 0.3$ but for $\rho > 0.9$, the error rate falls below α. This test was shown to be more powerful than the Bonferroni method for highly correlated statistics.

It may therefore be useful to employ independent statistics to distinguish between the two models. The most intuitive way is to find independent linear combinations of the original set of statistics by means of principal components analysis. The next section describes in detail the step required to test the hypotheses as in (6.1) followed by methods of testing principal components rather than the original statistics.



(a) Bonferroni rejection region. (b) Hommel rejection region. (c) Simes rejection region.

Figure 6.6: Comparison of the three rejection regions.

6.2.3 Parametric bootstrap

The basic principles of bootstrap methods is to find the distribution of some statistic $T(X)$ given independent and identically distributed data $\{X_1, \ldots, X_n\}$. This can be approached in two ways: either parametrically or non-parametrically. If an explicit probability model of the data exists, then the observed data can be used to estimate the parameters in the model and the distribution function, denoted by $\hat{F}$ by Davison and Hinkley (1997), is used for inference. In this context, the parametric bootstrap can be implemented to simulate data under the null model.

It is hypothesised that the observed data X, comprising of n haploid individuals from each of the two subpopulations genotyped at L SNPs, is from an isolation model with population split time τ. Observations $X_i = \{x_{i_1}, \ldots, x_{i_{2n}}\}$ for $i = 1, \ldots, L$ are assumed to be independent. For each statistic, the distribution under the null hypothesis is estimated. Implementing bootstrap methods, data sets $X_1^*, \ldots, X_{N_{sim}}^*$ are simulated under $f_{\hat{\tau}}$, the isolation model with $\tau = \hat{\tau}$, with $X_j^* = \{X_{j_1}^*, \ldots, X_{j_L}^*\}$ for each $j = 1, \ldots, N_{sim}$. The simulated data produces estimates $S_1^*, \ldots, S_{N_{sim}}^*$ of statistic S under the null hypothesis. Given observed data, the hypotheses outlined in (6.1) are tested using the following steps:

Algorithm 2 (Test I)**.**

1. *Calculate observed values of the summary statistics* $S_{obs} = \{S_{obs_1}, \ldots, S_{obs_m}\}$.
2. *Calculate* $\hat{F}_{st}$ *and estimate* $\hat{\tau}$ *using equations (5.1) and (3.10).*
3. *For* $j = 1, \ldots, N_{sim}$:

   4. *Simulate L SNPs under the isolation model with two subpopulations, each of sample size n,diverging at time* $\hat{\tau}$ *and calculate summary statistics* $S_j = \{S_{j,1}, \ldots, S_{j,m}\}$.
5. *Let* $\bar{S}_i = N_{sim}^{-1} \sum_{j=1}^{N_{sim}} S_{j,i}$. *For each statistic separately, test the hypothesis*

$$H_{0_i} : S_{obs_i} = \bar{S}_i$$
$$H_{1_i} : S_{obs_i} \neq \bar{S}_i$$

by calculating a p-value using the method described in section 6.2.1.

6. Correct the p-values using one of the methods described in section 6.2.2. If all hypotheses are accepted then accept the global null hypothesis, H_0, otherwise reject the global hypothesis.

Test I tests whether the data are consistent with an isolation model. However, it is equally reasonable to test if the data are from a migration model with migration rate m, estimating m from F_{st} and simulating data under the migration model.

6.2.4 Incorporating principal components analysis

In addition to the dependence structure presented in the summary statistics, there are additional complications in finding the 'best' set of statistics to exploit in the hypothesis test. As with ABC, it may seem intuitive to include as many summary statistics as possible. However, Joyce and Marjoram (2008) investigated the effects of using a large number of statistics in ABC. They suggest that including additional, uninformative, statistics merely adds noise and so devise a system that includes an additional statistic only if it "improves the quality of inference" although the authors do not account for the order in which the statistics are considered.

The use of principal components analysis reduces the number of dimensions and produces independent linear combinations of the original statistics and therefore may better fit the conditions of some multiple hypothesis tests. Bazin et al. (2010) for example examined a set of loci for natural selection and used PCA to reduce the dimension of the data to 30 components from an initial set of 60 statistics.

Let N_{sim} be the number of simulated data sets under H_0. For simulation i, the set of

statistics $\{S_{i,1}, S_{i,2}, \ldots, S_{i,m}\}$ are computed and the following matrix is produced:

$$\boldsymbol{\mathcal{S}} = \begin{pmatrix} S_{11} & S_{21} & \ldots & S_{N_{sim},1} \\ S_{12} & S_{22} & \ldots & S_{N_{sim},2} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1m} & S_{2m} & \ldots & S_{N_{sim},m} \end{pmatrix}. \tag{6.2}$$

The objective is to make linear combinations of the original m statistics. However, there is some disagreement on whether it is appropriate to perform principal component analysis on the covariance matrix or the correlation matrix. In particular, Jolliffe (2002) illustrates that, in some case, the methods produce very different component coefficients. When using the covariance matrix, Jolliffe demonstrates that if there is a sizeable difference in the variances of the original variables, then those variables with the largest variances tend to dominate. Since each statistic is not necessarily on the same scale, as exemplified in figures 6.3 and 6.4, principal components analysis is performed on the correlation matrix of dimension $m \times m$

$$\begin{pmatrix} 1 & \rho_{21} & \ldots & \rho_{m1} \\ \rho_{12} & 1 & \ldots & \rho_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1m} & \rho_{2m} & \ldots & 1 \end{pmatrix},$$

producing a matrix of coefficients

$$\begin{pmatrix} a_{11} & a_{21} & \ldots & a_{m1} \\ a_{12} & a_{22} & \ldots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \ldots & a_{mm} \end{pmatrix},$$

such that

$$PC_i = \sum_{j=1}^{m} a_{ij} S_j.$$

The number of components utilized may be defined in several ways. For instance, only one or two components may be employed, accounting for the most and second most variance in the data, respectively. Alternatively, a level $0 \leq \nu \leq 1$ may be specified such that the number of components used explains $100\nu\%$ of the variation, or, formally, a test of each eigenvalue of the correlation matrix can be made and the significantly large eigenvalues employed, as described by Patterson et al. (2006).

Integrating principal components analysis into our testing procedure involves additional work after step 3 in Test I, lead to

Algorithm 3 (Test II)**.**

1. *Calculated observed values of the summary statistics* $S_{obs} = \{S_{obs_1}, \ldots, S_{obs_m}\}$.
2. *Calculate* $\hat{F}_{st}$ *and estimate* $\hat{\tau}$ *using equations (5.1) and (3.10).*
3. *For* $j = 1, \ldots, N_{sim}$:

   4. *Simulate L SNPs under the isolation model with two subpopulations, each of sample size n, diverging at time* $\hat{\tau}$ *and calculate summary statistics* $S_j = \{S_{j,1}, \ldots, S_{j,m}\}$.
5. *Construct matrix* $\mathcal{S}$ *in (6.2).*
6. *Compute correlation matrix and identify components* $PC_1, \ldots, PC_{n_{comp}}$ *for some determined* $n_{comp} \in \{1, \ldots, m\}$.
7. *For each component separated, test the hypothesis*

$$H_{0_i} : PC_{obs_i} = \overline{PC_i}$$
$$H_{1_i} : PC_{obs_i} \neq \overline{PC_i}$$

by calculating a p-value using the method described in section 6.2.1.

8. *Correct the p-values using one of the methods described in section 6.2.2. If all hypotheses are accepted then accept the global null hypothesis,* H_0*, otherwise reject the global hypothesis.*

6.3 Type I and Type II errors

Testing a single hypothesis at significance level α, two possible errors can occur: either a Type I error where the null hypothesis is rejected when it is true or a Type II error where the null hypothesis is accepted when it is false. A hypothesis test should aim to minimise both types of errors. The same conditions should hold in a multiple hypothesis setting.

Generally, let m_0 denote the number of true null hypotheses from the collection $\{H_{0_1}, \ldots, H_{0_m}\}$. Table 6.1 provides all eventualities in a multiple hypothesis test.

Hypothesis	Not Rejected	Rejected	Total
True	U	V	m_0
False	T	S	$m - m_0$
Total	W	R	m

Table 6.1: Counts of Type I and Type II errors in multiple hypothesis testing reconstructed from Bretz et al. (2011), table 2.1.

6.3.1 Type I error

In the case of testing multiple hypothesis, there are many definitions of the Type I error rate. Most common is the familywise error rate (FWER), defined as

$$\text{FWER} \quad = \quad Pr(V > 0).$$

In order to control FWER, it is required that

$$Pr(V > 0) \quad \leq \quad \alpha. \tag{6.3}$$

For example, the Bonferroni procedure rejects hypothesis H_{0_i} if $p_i < \frac{\alpha}{m}$. Under the null hypothesis, the p-values are draws from a $Uni(0, 1)$ distribution and $Pr(p_i \leq \frac{\alpha}{m}) = \frac{\alpha}{m}$ for

all $i = 1, \dots m$ and the Bonferroni inequality ensures

$$\begin{aligned} Pr\Big(\bigcup_{i=1}^{m}\Big(p_i \leq \frac{\alpha}{m}\Big)\Big) &\leq \sum_{i=1}^{m} Pr\Big(p_i \leq \frac{\alpha}{m}\Big) \\ &= \alpha, \end{aligned}$$

with equality if the p-values are independent.

Bretz et al. (2011) differentiates between controlling FWER in a weak sense and in a strong sense. Controlling FWER in the weak sense corresponds to the probability that at least one hypothesis is declared false given that they are all true, that is

$$Pr(V > 0| \text{ the global hypothesis is true }) \leq \alpha.$$

If not all the null hypotheses are true, strongly controlling the FWER requires

$$\max_{I \subseteq M} Pr\Big(V > 0| \bigcap_{i \in I} H_i\Big) \text{ is true } \leq \alpha, \tag{6.4}$$

for $M = \{1, \dots, m\}$ and non-empty subsets I of M. M is the set of hypotheses and so (6.4) requires that for every subset $I \subseteq M$, the probability of falsely rejecting a true hypothesis, given that the set of I hypotheses are true, is less than (or equal to) α. Bretz et al. (2011) explains the difference between the two. Weakly controlling the Type I error rate is synonymous to controlling the probability of rejecting a null hypothesis whereas controlling in the strong sense is to control the probability of rejecting at least one true hypothesis.

Secondly, as the number of hypotheses tested increases, it becomes more likely that at least one null hypothesis will be rejected and so in some scenarios, it may not be required that all rejections be correct. The false discovery rate FDR is defined to be

$$\text{FDR} = E\Big(\frac{V}{R} \mid R > 0\Big) Pr\Big(R > 0\Big),$$

and is the expected proportion of falsely rejected hypothesis among the rejected hypothesis. In this case, under the null model all the individual tested hypotheses are true. Therefore, the probability of rejecting the null hypothesis given that it is true is equal to the probability that at least one hypothesis is rejected given that they are all true, weakly controlling the FWER.

To test whether TestI and II weakly control the FWER, a range of population divergence times was considered and 'observed' data comprising of 1000 SNPs and a sample of size 10 in each subpopulation were simulated under the isolation model and the eight statistics computed (the average heterozygosity was excluded from each test since distributions under the null and alternate model appeared indistinguishable). Test I was implemented 100 times, applying Hommel's and Simes' corrections for multiple comparisons. In addition, Test II was considered employing different numbers of components from the PCA analysis, namely 1 or 2 components and the number of components that collectively accounted for more that 97% of the variation in the data (chosen arbitrarily). Figure 6.7 displays the results. For each value of τ, the probability of rejecting H_0 was estimated as the number of the rejected global hypotheses divided by the total number of tests:

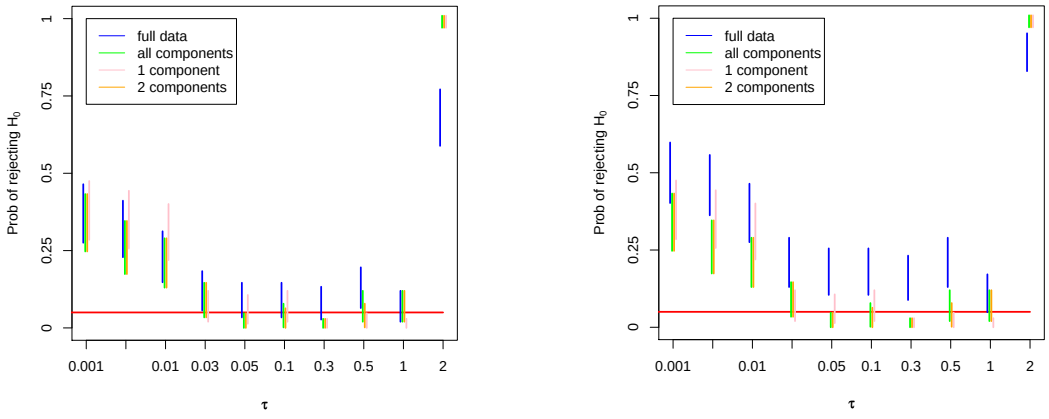
$$\hat{\rho} \quad = \quad \frac{\#\text{ of rejected hypotheses}}{100},$$

with the endpoints of an approximate 95% confidence interval for the probability calculated from

$$\hat{\rho} \pm 1.96\sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{100}}.$$

The red lines in figure 6.7 correspond to the probability of rejecting $H_0 = 0.05$. The vertical lines show the approximate 95% confidence intervals for the probability for each value of τ.

Generally, Hommel's method performs slightly better than Simes' since, for the full data, the Simes' confidence intervals lie above the red line for all but one of the τ values. This



(a) (b)

Figure 6.7: Estimated probability of Type I error rate for a range of τ using (a) Hommel's and (b) Simes' corrections for multiple comparisons.

pattern is repeated using Test II for the range of components considered. The most obvious problem is the dramatic increase in Type I error for $\tau > 1$. However, τ is measured in $2N$ generations and $\tau > 1$ corresponds approximately to $F_{st} > 0.5$, which, in human populations, is unrealistic, as demonstrated in table 2.1 where the F_{st} values are in the range $(0.005, 0.35)$. Likewise, smaller τ's show an increase in the probability of a Type I error. It appears that this methods begins to fail below $\tau \approx 0.03$, which corresponds to $F_{st} < 0.03$ (applying equation (3.9)). The reason for this result is the poor estimation of the population divergence time in this range. Data were simulated under an isolation model for a range of τ values in the interval $[0.00001, 0.08]$ and $\hat{\tau}$ calculated. Figure 6.8 illustrates how this estimator tends to overestimate population divergence times, more so as $\tau \to 0$. The pink confidence bands show the 2.5th and the 97.5th percentiles of the distribution of each estimate for the range of τ values. Logically, as $\tau \to 0$, the isolation model begins to approach an unstructured model. Therefore, one might propose that if $\hat{\tau} < \delta$, for some $\delta > 0$, then the data may be more consistent with a neutral model and so it may be appropriate to set $\hat{\tau} = 0$. More precisely, let $\tau^* = \frac{F_{st}}{1-F_{st}}$ then the estimator is revised to be

$$\hat{\tau} = \begin{cases} \tau^*, & \text{if } \tau^* > \delta; \\ 0, & \text{if } \tau^* \leq \delta. \end{cases}$$

In words, an isolation model with an estimated population divergence time less than δ is taken as an unstructured model.

6.3.1.1 Isolation model vs neutral model

In order to assess when the isolation model with population divergence $\tau \leq \delta$ is more similar to a neutral model that an isolation model with population divergence τ^*, the distributions of the summary statistics are simulated under the appropriate model. The three models considered are

1. the isolation model with divergence time τ,
2. the isolation model with divergence time τ^*, and
3. the unstructured model. For completeness, for statistics that rely on population labels, for instance π_W and π_B, population labels are randomly allocated to the sample.

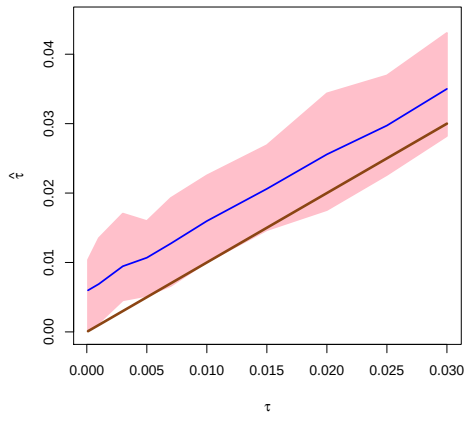


Figure 6.8: Confidence bands of τ for a range of values from data simulated under isolation model

Using a small divergence rate, $\tau = 0.00001$, data were simulated under the isolation model with a sample of size 10 from each subpopulation at 1000 SNPs and the statistics were computed. In each simulation, τ^* was estimated and data were simulated under the isolation model with τ^* and the set of statistics computed. 1000 SNPs were simulated under the unstructured model with a sample size of 20 and the statistics computed. This process was repeated 100 times and the resulting distributions of the statistics are presented in figure 6.9.

Subjectively, the distributions of the statistics under the three models are quite similar. However, arguably the blue and red bars are more in agreement compared to the green bars suggesting that the true model, corresponding to the red bars, is best estimated by the neutral model and so $\hat{\tau} = 0$.

In order to estimate δ, we considered the distribution of allele frequencies under the unstructured model. Griffiths and Tavaré (1998) derived an expression for the expected allele frequency spectrum under an unstructured model by considering the probability of observing x copies of a mutant allele in the sample. Suppose in a single population of sample size n that, at a particular locus, a single mutation occurred before the most common recent ancestor of the sample. Then for a small mutation rate,

$$Pr(X = x) = \frac{\sum_{k=2}^{n} k p_{n,k}(x) E(T_k)}{\sum_{k=2}^{n} k E(T_k)}, \qquad 0 < x < n,$$

where $p_{n,k}(x)$ is the probability that a mutation occurs during the time there are k lineages that results in x copies of the mutant allele in a sample of size n. Suppose there are k lineages present in the sample, this probability was derived by Feller (1950) by considering the number of ways of placing n balls into k cells such that no cell is empty, which was found to be $\binom{n-1}{k-1}$. If there are x copies of the mutant allele, this is the same as placing x balls into 1 cell and placing $n - x$ balls in the remaining $k - 1$ cells. Hence the required

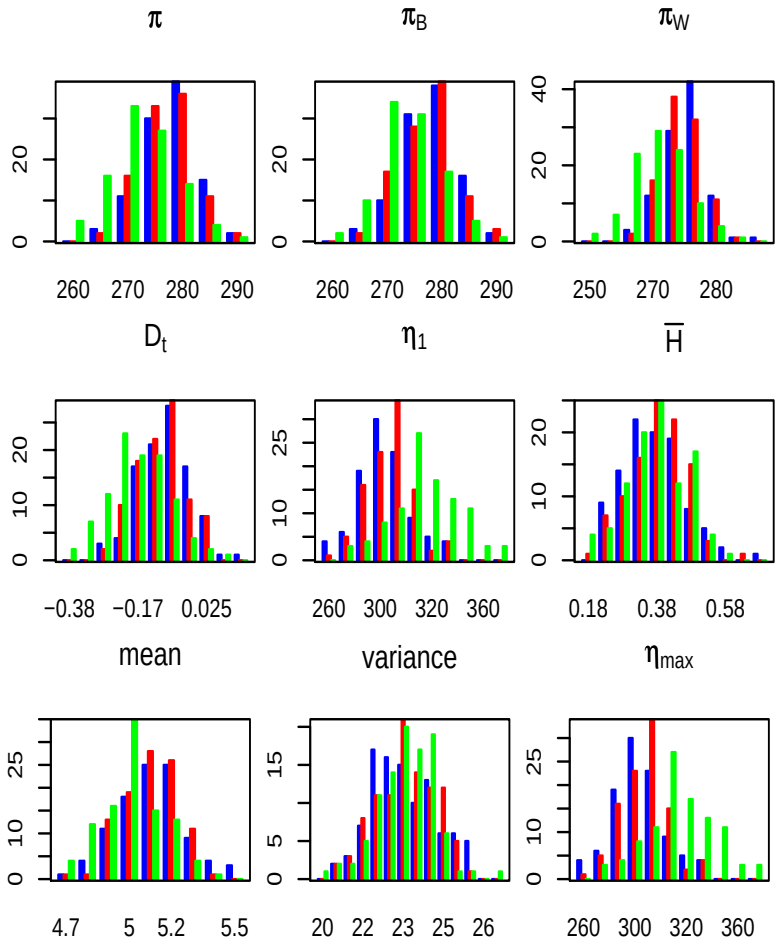


Figure 6.9: Histograms of summary statistics from data simulated under the isolation model with $\tau = 0.00001$ (red bars) and τ^* (green bars) and under an unstructured model (blue bars).

probability was given by Griffiths and Tavaré (1998) as

$$p_{n,k}(x) \quad = \quad \frac{\binom{n-x-1}{k-2}}{\binom{n-1}{k-1}}.$$

Under the standard coalescent model $E(T_k) = \frac{2}{k(k-1)}$ for $k = 2, \ldots, n$ and so Griffiths and Tavaré (1998) showed

$$Pr(X = x) = \frac{\frac{1}{x}}{\sum_{j=1}^{n-1} \frac{1}{j}}, \qquad 1 < x < n-1. \tag{6.5}$$

This distribution assumes that it is known which allele is ancestral. Figures 6.10 illustrates the allele frequency spectrum under this model with a sample size of 40 using equation (6.5).

Given the expected frequencies under an unstructured model, it is possible to test whether observed allele frequencies are consistent with allele counts under a neutral model by testing

H_0 : allele frequencies derived from an unstructured model

H_1 : allele frequencies not derived from an unstructured model.

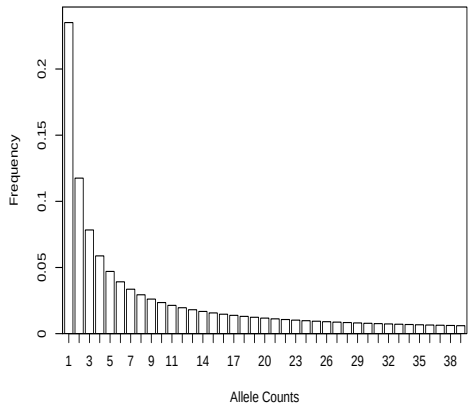
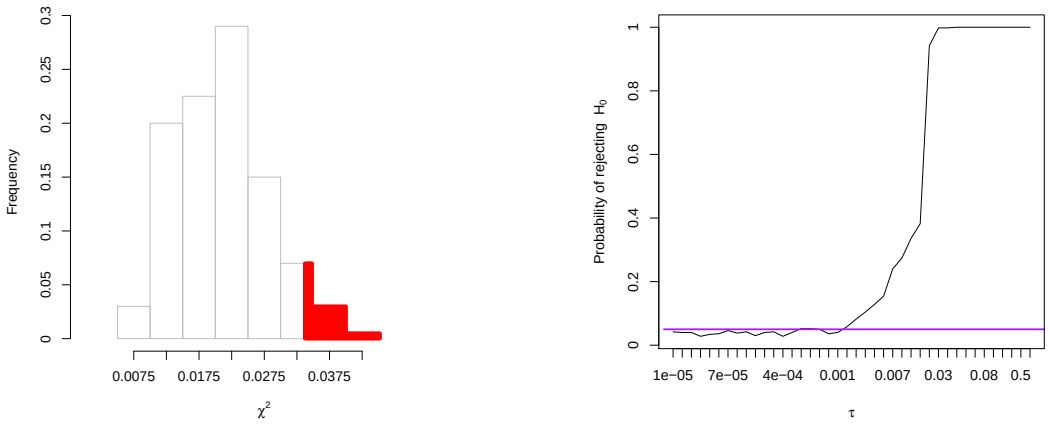


Figure 6.10: Example of allele frequency spectrum under a unstructured model.

Let $\{p_1, \ldots, p_{n^T-1}\}$ denote the expected frequencies under the neutral model calculated using equation (6.5) and $\{p_{obs_1}, \ldots, p_{obs_{n^T-1}}\}$ denote the observed allele frequencies, then



(a) (b)

Figure 6.11: (a) Distribution of χ^2 under the null model. (b) Probability of rejecting the unstructured model for a range of values of τ.

the test statistic

$$\chi^2 = \sum_{i=1}^{n^T-1} \frac{(p_{obs_i} - p_i)^2}{p_i},$$

is computed. Data were simulated under a neutral model with total sample size n^T and χ^2 computed, repeating this process produced the distribution of χ^2 under the null hypothesis given in figure 6.11(a). The upper 5% of the distribution is highlighted in red and if the observed test statistic lies within this tail of the distribution, the null hypothesis is rejected. In order to gauge the value of δ, data were simulated for a range of values of τ and the χ^2 statistic was computed. This was repeated 500 times and for each simulation, $\hat{\tau}$ was recorded. The probability of rejecting the null hypothesis was computed as

$$\frac{\#\ \chi^2 \text{ in the upper 5\% of null distribution}}{\#\ \text{simulations}}.$$

The results are shown in figure 6.11(b). The null hypothesis is not rejected for values of τ that produced a probability of less than 0.05. Therefore, from figure 6.11(b), the distribution of allele frequencies under the isolation model with $\tau < 0.002$ is not signific-

antly different from the allele frequencies under a neutral model. Given $\tau = 0.002$, the corresponding value τ^* needs to be estimated. From the data simulated using $\tau = 0.002$, the average value of τ^* was 0.009 and taking the 2.5th and 97.5th percentiles, a confidence interval of $(0.003, 0.013)$. Setting $\delta = 0.009$, the resulting estimator is shown in figure 6.12, which compares the estimator $\hat{\tau}$ when $\delta = 0, 0.009, 0.013$ and 0.015. Using $\delta = 0.009$ does improve the estimation of τ, however, it appears using $\delta = 0.013$ produces the most satisfactory estimate and so the estimator

$$\hat{\tau} = \begin{cases} \tau^*, & \text{if } \tau^* > 0.013; \\ 0, & \text{if } \tau^* \leq 0.013 \end{cases} \tag{6.6}$$

is used in step 1 of Test I and Test II. The type I error, given this estimator, is shown in figure 6.13. As $\tau \to 0$, there is a reduction in the number of false rejections. However, for values of τ around 0.1, there is an increase in the type I error. Therefore, although this estimator is helpful for $\tau \approx 0$, the error rate is still greater than 0.05 for $\hat{\tau} \in [0.001, 0.01]$ (approximately).

Figure 6.12 shows equation (6.6) improves the estimation of τ, however, there is still a high amount of variation shown by the pink confidence bands. Generally, the brown line showing $\tau = \hat{\tau}$ lies close to the lower band, more so when $\tau = 0.01$. In order to account for the variation, instead of using only a single estimate of τ, some noise is included as described in Test III.

Algorithm 4 (Test III)**.**

1.Calculate observed values of summary statistics S_{obs} and estimate $\hat{F}_{st} = F_{st_1}$ and then $\hat{\tau}$ using equation (6.6).

2. Set $N_{acc} = 0$ and while $N_{acc} < N_{sim}$:

3. Simulate $T \sim N(\hat{\tau}, \sigma^2)$. If $T \geq 0$:

4. Simulated data under isolation model with two subpopulations diverging at time T and calculate $\hat{F}_{st} = F$.

5. If $|F - F_{st_1}| < \epsilon$, calculate summary statistics and set $N_{acc} = N_{acc} + 1$.

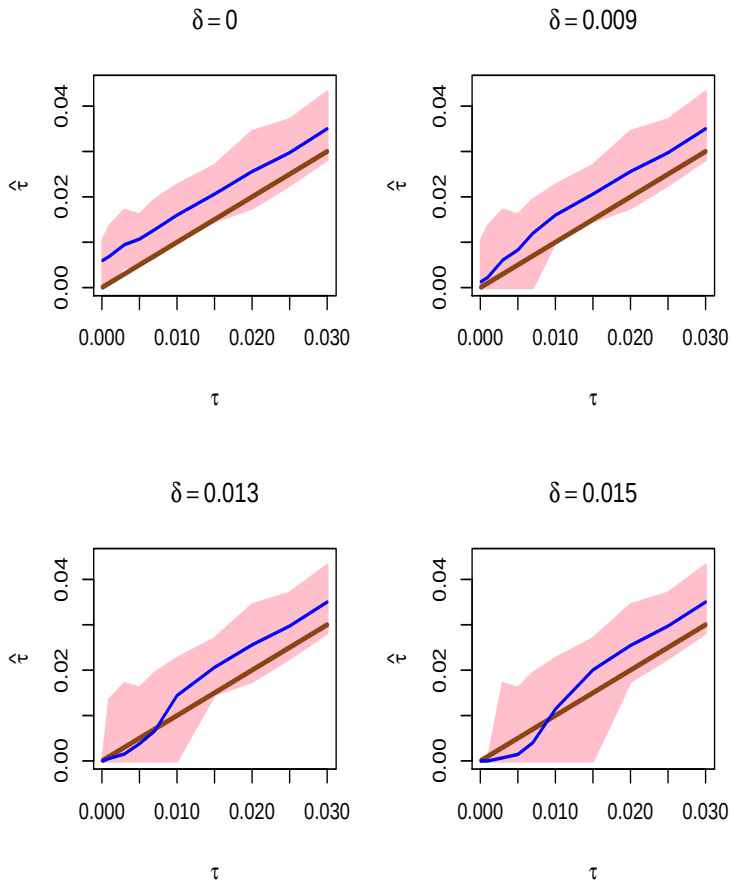


Figure 6.12: Confidence bands of $\hat{\tau}$ using $\delta = 0$ (top left hand side), $\delta = 0.009$ (top right hand side), $\delta = 0.013$ (bottom left hand side) and $\delta = 0.015$ (bottom right hand side). The solid brown line shows the line $\tau = \hat{\tau}$.

6. For each statistic separately, test the hypothesis

$$H_{0_i} : S_{obs_i} = \bar{S}_i$$
$$H_{1_i} : S_{obs_i} \neq \bar{S}_i.$$

7. Correct the p-values using one of the methods described in section 6.2.2. If all hypotheses are accepted then accept the global null hypothesis, H_0, otherwise reject the global hypothesis.

This test requires specification of two parameters namely σ^2 and ϵ. The value of ϵ should be small enough to accept only values of T that are close to $\hat{\tau}$ and the value of σ^2 could

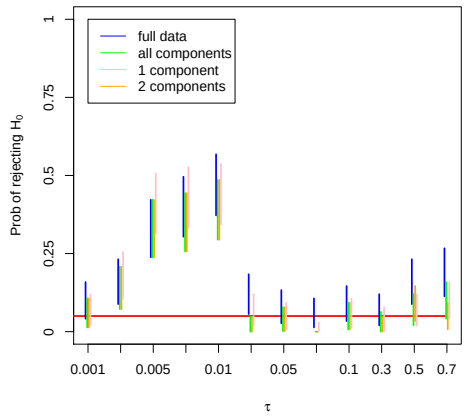


Figure 6.13: Type I error for a range of τ values using equation 6.6 to estimate τ.

be an indication of how well it is thought $\hat{\tau}$ estimates τ. The values of σ^2 and ϵ should not be too large or small. If too large, the test accepts values of T inconsistent with the observed data and if too small, the test is computationally expensive.

A range of values of ϵ and σ^2 were considered and it was found that setting $\epsilon = \sigma^2 = 0.01$ controls the type I error as illustrated in figure 6.14 as does setting $\epsilon = 0.001$ and $\sigma^2 = 0.01$. Figure 6.14 also shows the error rate using $\epsilon = 0.02$, $\sigma^2 = 0.015$ and $\epsilon = 0.001$, $\sigma^2 = 0.005$. When $\sigma^2 = 0.005$ there is an increase in the error, suggesting 0.005 is not an adequate value of σ^2. Setting $\epsilon = 0.02$ and $\sigma^2 = 0.015$ shows a similar pattern to figure 6.13.

6.3.2 Power of hypothesis test

As with the Type I error rate, the Type II error rate and so the power is an essential consideration in the construction of a hypothesis test. In multiple comparisons, the individual power of each test, denoted as $\beta_i^{ind} = Pr\{\text{reject } H_{0_i} | H_{0_i} \text{ is false}\}$ by Bretz et al. (2011), is used to define the power of the global hypothesis. In particular, the disjunctive power

$$\beta^{dis} = Pr\{S \geq 1\},$$

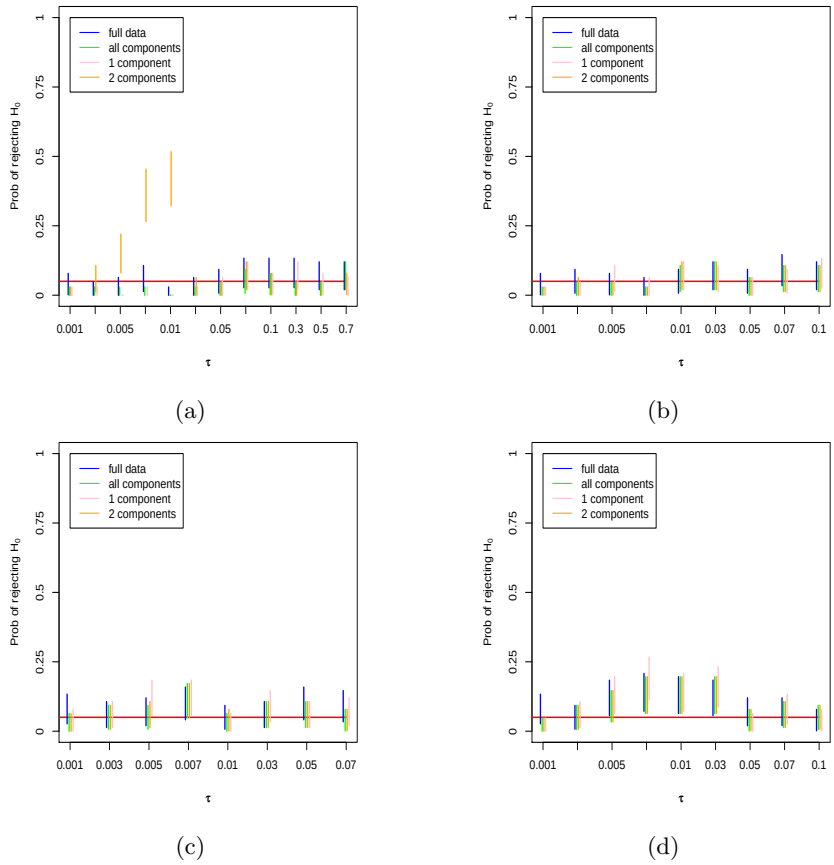


Figure 6.14: Type I error for a range of τ values using Test III with (a) $\epsilon = 0.02$, $\sigma^2 = 0.015$, (b) $\epsilon = \sigma^2 = 0.01$, (c) $\epsilon = 0.001$, $\sigma^2 = 0.01$ and (d) $\epsilon = 0.001$, $\sigma^2 = 0.005$.

is the probability of rejecting at least one false hypothesis, with S defined in table 6.1 as the number of rejected false hypotheses. If the observed data are truly from a migration model then in order to reject the isolation model, at least one hypothesis needs to be rejected and so the disjunctive power is an appropriate measure of power.

To gauge the power of Test III, 1000 SNPs were simulated from the migration model with two subpopulations each of sample size 10 for a range of migration rates and the null hypothesis was tested. Repeating this process 100 times, the proportion of rejected hypotheses was estimated and confidence intervals for the proportion of rejected hypotheses was computed. The results are given in figure 6.15. As the migration rate approach zero, the power tends to one, whereas as the migration rate becomes large, the model begins

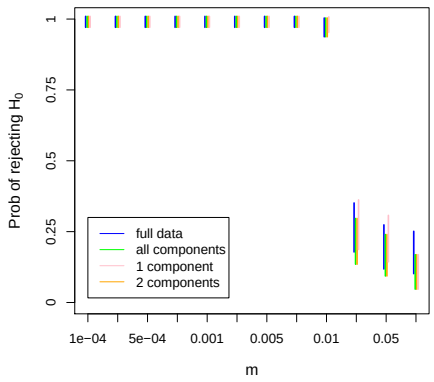


Figure 6.15: Power of hypothesis test for a range of migration rates.

to approach an unstructured model and so the two models are not only indistinguishable from an unstructured model but also from each other and so figure 6.15 shows a decline in power.

6.4 Discussion

This chapter aims to set out a powerful hypothesis test to distinguish between the isolation and migrations models with two subpopulations. Test III provides steps to test whether observed data are consistent with an isolation model. The method utilises a set of summary statistics and tests each one individually before testing the global null hypothesis correcting for simultaneously testing multiple hypotheses. Principal components analysis was used to make independent linear combinations of the statistics but this was shown to have little effect in the type I error rate and power of the test. However, this test had a total of only eight statistics, and reducing the dimension of the summaries may be more beneficial in studies involving a larger number of statistics. The test requires estimation of the population divergence time and consequently the specification of two parameters. Using F_{st} to estimate τ, a value σ^2 was introduced to address poor estimation and the parameter ϵ ensured that simulated data had approximately the same F_{st} as the observed data.

These additional steps are required in order to control the Type I error rate. Although computationally more expensive than using the simple point estimator, the ABC_MCMC algorithm produced more accurate results and so may improve the quality of Test III in that it may control the Type I error rate without the need to include these extra steps.

Chapter 7

Extensions to the hypothesis test

This chapter aims to incorporate further aspects of the treatment of real data into the hypothesis test developed in the previous chapter. Firstly, SNP loci that have been ascertained in some manner, prior to being genotyped in the samples of interest are considered. Secondly, due to the increasing dimensionality of SNP data sets, attempts are made to test projected data instead of the full data, to make feasible a hypothesis test that would otherwise be too computationally demanding.

7.1 Ascertained data

In order to make inference using ascertained data, it may be possible either:

1. to find methods that are robust against ascertainment;
2. to build in directly a model of ascertainment in the simulation steps of the bootstrap;
3. to apply a correction to the parameter estimates as described by Nielsen and Signorovitch (2003) and Albrechtsen et al. (2010). These methods correct for ascertainment by finding the maximum likelihood estimates of the true allele frequencies

given the ascertainment scheme.

Since the global hypothesis test is based on statistics that are functions of the allele frequencies, this test, as developed in chapter 6, is not robust against ascertainment. Figures 4.7 and 4.8 exemplified this by comparing the allele frequency spectra of simulated data from both models with and without ascertainment. Therefore, ascertainment needs to be accounted for at some point during the test.

Three different ways to account for ascertainment in the hypothesis test are proposed. The first involves testing the consistency of the 'corrected' observed data with an isolation model simulated with no ascertainment. The second is to test the observed data using Test III (from section 6.3.1.1) but simulate data in step 4 under the isolation model with the same ascertainment scheme as the observed data. Note, however, that the ascertained data may not allow adequate estimation of F_{st}, and hence of the population divergence time τ. The last method involves using the corrected allele frequencies to estimate F_{st}, and subsequently τ, but then testing the observed data using Test III by simulating ascertained data in step 4.

This chapter analyses how well the test statistics distinguish the two models, analogously to section 6.1.1, by simulating data with ascertainment from both models and comparing the distributions of the statistics and also estimating τ using ascertained data, with and without any corrections of allele frequencies.

In the first instance, the ascertainment scheme considered is one in which a sample of size n_a is taken from each subpopulation and if variability is found within this sample, then a larger sample is taken and the final data set includes the original ascertainment sample.

7.1.1 Initial comparison of statistics

Section 4.4 illustrated the dramatic effects ascertainment has on the allele frequencies. In particular, figure 4.9 not only highlighted the differences between allele frequencies from

data with and without ascertainment but also showed the distributions of allele frequencies of ascertained data from both models are more similar compared to unascertained data. As a result, it initially appears to be more difficult to distinguish these models by the methods introduced in chapter 6.

In order to investigate how much of an effect ascertainment has on distinguishing the models, the set of summary statistics, from the previous chapter, is considered under both models. For τ values close to zero, it becomes increasingly difficult to distinguish the two models, hence, a large enough value of τ (and small enough corresponding migration rate m) are chosen to produce an initial impression of the power of the test. 1000 SNPs were simulated from two subpopulations each of sample size 10 and an ascertainment sample size in each subpopulation of size 2 and the set of statistics computed. This was repeated 100 times and the results, for each statistic separately, are given in figure 7.1. Most of the statistics still appear useful for distinguishing the models although the distributions of π_W and η_{max} overlap considerably and so may need to be excluded from the test.

7.1.2 Correcting for ascertainment

Nielsen et al. (2004) derived, using Bayes' theorem, an expression for the probability that, in a sample of size n^T, there are x copies of a mutant allele, given variability in the ascertainment sample of size m, assuming the ascertained sample is included in the final sample:

$$\begin{aligned} Pr(X = x|\text{Asc}) &= \frac{Pr(\text{Asc}|x)Pr(x)}{Pr(\text{Asc})} \\ &= \left\{1 - \frac{\binom{x}{m} + \binom{n^T - x}{m}}{\binom{n^T}{m}}\right\} \frac{\frac{1}{x}}{\sum_{j=1}^{m-1} \frac{1}{j}}, \qquad 0 < x < n^T, \end{aligned} \tag{7.1}$$

where 'Asc' is the event that the ascertainment sample is variable. Equation (7.1) is

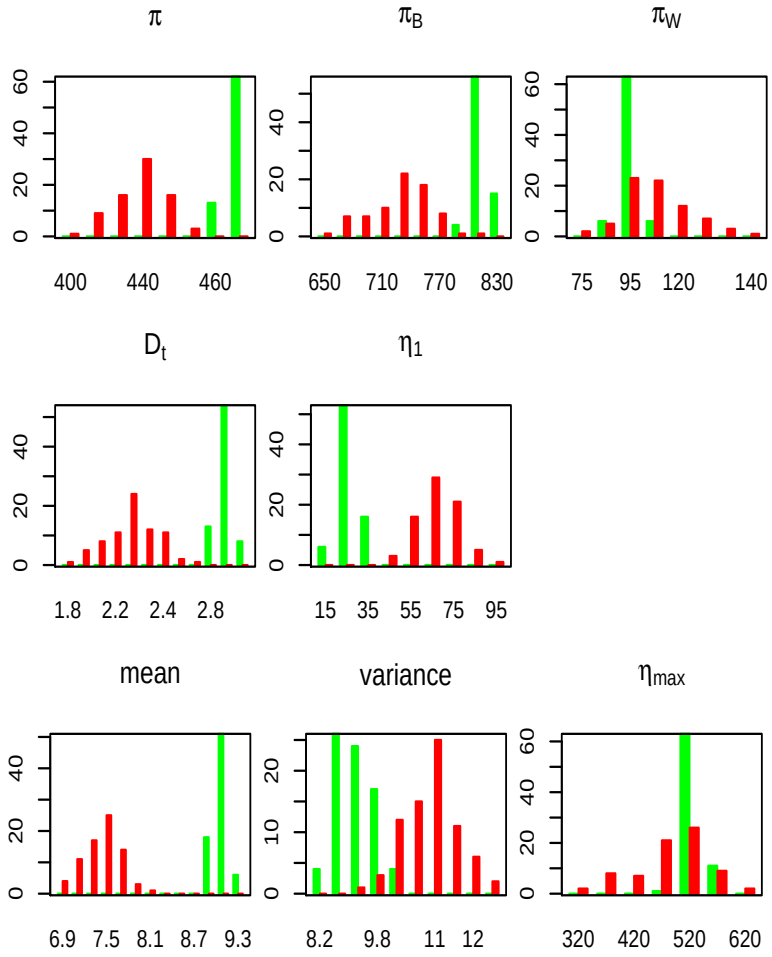
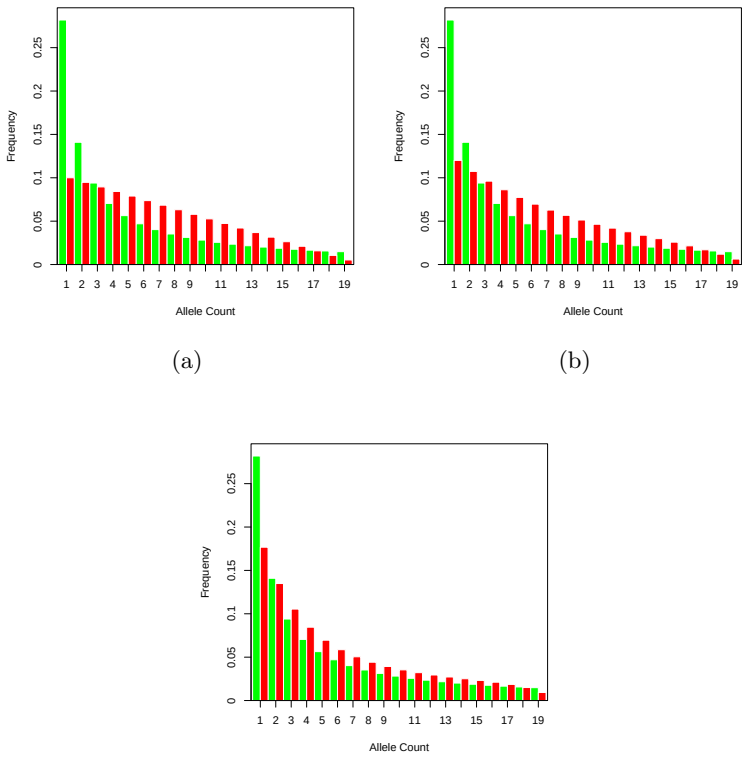


Figure 7.1: Histograms of summary statistics from data simulated under the isolation (green bars) and migration (red bars) models with ascertainment

the probability that there are x copies of the mutant allele multiplied by one minus the probability that the alleles in the ascertainment panel are all of the same type (that is, the probability that the ascertainment panel is variable). The authors derived similar expressions for other ascertainment schemes. This equation shows that the size of the ascertainment panel impacts the allele frequencies. For $n = 10$ and $m = 2, 5$ and 10, figure 7.2 illustrates the distribution of allele counts. The distribution of allele counts with no ascertainment is given by (6.5). As the ascertainment sample increases to 10, the allele counts begin to imitate those of the sample with no ascertainment. For lower

ascertainment sample sizes, rare alleles are under–represented, with a corresponding over–representation of intermediate frequency alleles.



(a) (b)

(c)

Figure 7.2: Allele frequency spectra under the standard coalescent model, shown in the green bars, compared with allele frequency spectra with an ascertainment process (red bars) of sample size 2 (a), 5 (b) and 10 (c).

7.1.2.1 Correcting allele frequencies in the whole population

Nielsen et al. (2004) computed the likelihood of $P = \{p_1, \ldots, p_{n^T-1}\}$, where p_i is the frequency of SNPs with allele count i in a sample with no ascertainment. Their method makes no parametric assumptions and therefore is valid under any genetic model. Let $X = \{X_1, \ldots, X_L\}$, where X_i is the observed allele count of SNP i. The authors show

$$
\begin{aligned}
L(P) &= Pr\{X|P\} \\
&= \prod_{i=1}^{L} Pr\{X_i = x_i|P, \text{Asc}_i\} \qquad \text{(assuming independent loci)} \\
&= \prod_{i=1}^{L} \frac{Pr\{X_i = x_i, \text{Asc}_i|P\}}{Pr\{\text{Asc}_i|P\}} \qquad \text{(by applying Bayes' theorem)} \\
&= \prod_{i=1}^{L} \frac{Pr\{X_i = x_i|P\}Pr\{\text{Asc}_i|X_i = x_i, P\}}{Pr\{\text{Asc}_i|P\}} \qquad \text{(again, applying Bayes' theorem)} \\
&= \prod_{i=1}^{L} \frac{p_i Pr\{\text{Asc}_i|X_i = x_i\}}{Pr\{\text{Asc}_i|P\}} \qquad \text{(replacing } Pr\{X_i = x_i|P\} = p_i\text{)} \\
&= \frac{\prod_{k=1}^{n^T-1} p_k^{\eta_k} Pr\{\text{Asc}_i|X_k = x_k\}^{\eta_k}}{\prod_{i=1}^{L} Pr\{\text{Asc}_i|P\}} \qquad \text{(where } \eta_k \text{ is the number of SNPs with frequency } p_k\text{.)} \\
&= \frac{\prod_{k=1}^{n^T-1} p_k^{\eta_k} Pr\{\text{Asc}_i|X = x_k\}^{\eta_k}}{\prod_{i=1}^{L} \sum_{k=1}^{n^T-1} Pr\{\text{Asc}_i, X_i = k|P\}} \\
&= \frac{\prod_{k=1}^{n^T-1} p_k^{\eta_k} Pr\{\text{Asc}_i|X = x_k\}^{\eta_k}}{\prod_{i=1}^{L} \sum_{k=1}^{n^T-1} p_k Pr\{\text{Asc}_i|X = x_k\}}
\end{aligned}
$$

An expression for $Pr\{\text{Asc}_i|X_i = x_i\}$ is required under the particular ascertainment scheme. In the previously described situation,

$$Pr\{\text{Asc}_i|X_i = x_i\} = 1 - \frac{\binom{x_i}{m} + \binom{n^T - x_i}{m}}{\binom{n^T}{m}}.$$

The authors maximise the log-likelihood function

$$l(P) = \sum_{k=1}^{n^T-1} \eta_k \log(p_k Pr\{Asc|X = x_k\}) - L \log\left(\sum_{k=1}^{n^T-1} p_k Pr\{Asc|X = x_k\}\right)$$

under the constraints $0 \leq p_k \leq 1$ and $\sum_{k=1}^{n^T-1} p_k = 1$ to find

$$\hat{p}_i = \frac{\eta_i}{Pr\{\text{Asc}_i|X=i\}} \left[\sum_{j=1}^{n^T-1} \frac{\eta_j}{Pr\{\text{Asc}_i|X=j\}} \right]^{-1}. \quad (7.2)$$

Simulating data under an isolation model with two subpopulations each of size $n = 10$ without ascertainment and with an ascertainment sample of size 2 from each subpopulation, figure 7.3 illustrates how well (7.2) estimates the allele frequencies. This method recovers the true allele frequencies well.

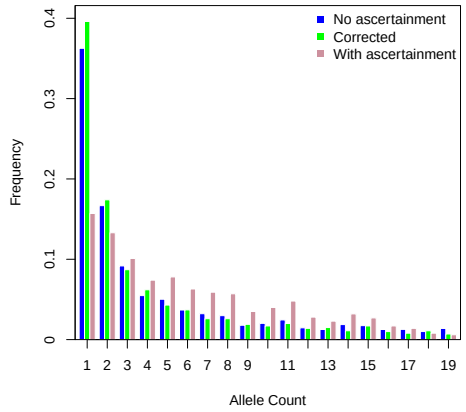


Figure 7.3: Allele frequency spectra of data simulated under an isolation model without ascertainment (blue bars), with ascertainment (pink bars) and maximum likelihood frequencies given ascertainment (green bars).

7.1.2.2 Correcting allele frequencies within subpopulations

In order to calculate the set of summary statistics required to perform the hypothesis test, the frequencies within each subpopulation, denoted by P_1 and P_2, need to be estimated. Let p_{dj} be the frequency of SNPs with allele count j in the dth subpopulation and η_{dj} be the observed number of SNPs in subpopulation d with count j. When $d = 2$, let η_{j_1,j_2} be the number of SNPs with count j_1 in subpopulation 1 and j_2 in subpopulation 2.

Nielsen (2004) illustrated that, if the ascertainment sample size is equal in both subpopulations, the allele frequency spectrum is, in expectation, equal in the two subpopulations. Hence, one can either maximise the likelihood of $\{P_1, P_2\} = \{p_{10}, \ldots, p_{1n_1}, p_{20}, \ldots, p_{2n_2}\}$ or assume the allele frequencies are the same in the two subpopulations, hence, reducing the number of parameters to be estimated by maximising the likelihood of $P_w = \{p_{w0}, \ldots, p_{wn}\}$, where $n = n_1 = n_2$.

The likelihood of the within-subpopulation frequencies P_w is

$$\begin{aligned} L(P_w) &= \prod_{i=1}^{L} Pr\{X_{1i} = x_{1i}, X_{2i} = x_{2i} | P_w, \text{ Asc}\} \\ &= \prod_{i=1}^{L} \frac{Pr\{X_{1i} = x_{1i}, X_{2i} = x_{2i}, \text{Asc} | P_w\}}{Pr\{\text{Asc}|P_w\}} \\ &= \frac{\prod_{a=0}^{n}\prod_{b=0}^{n} p_{wa}^{\eta_{1a}} p_{wb}^{\eta_{2b}} Pr(\text{Asc}|X_1 = a, X_2 = b)^{\eta_{a,b}}}{[\sum_{j=0}^{n}\sum_{k=0}^{n} p_{wj}p_{wk}Pr\{\text{Asc}|X_1 = j, X_k = k\}]^L}, \end{aligned}$$

and so the log likelihood is

$$\begin{aligned} l(P_w) &= \sum_{a=0}^{n}\sum_{b=0}^{n}\Big[\eta_{1a}\log(p_{wa}) + \eta_{2b}\log(p_{wb}) + \eta_{a,b}\log(Pr\{\text{Asc}|X_1 = a, X_2 = b\})\Big] \\ &\quad -L\log\left(\sum_{j=0}^{n}\sum_{k=0}^{n} p_{wj}p_{wk}Pr\{\text{Asc}|X_1 = j, X_k = k\}\right). \end{aligned} \tag{7.3}$$

The maximum-likelihood estimate of P_w cannot be obtained analytically. Therefore, the log-likelihood function is numerically optimized. In this case, the parameters $p_{w0}, \ldots, p_{wn}$ are constrained such that

$$\sum_{i=0}^{n} p_{wi} = 1 \text{ and } 0 \leq p_{wi} \leq 1 \qquad \text{for } i = 0, \ldots, n.$$

The maximum-likelihood estimators were found using the 'alabama' package written by Varadhan (2011) in R (R Development Core Team (2008)). This method uses a Lagrangian adaptive barrier method for optimizing a nonlinear function, as described by Lange (1999).

More specifically, it allows the twice differentiable function $f(\theta)$ to be minimised with respect to $\theta = (\theta_1, \ldots, \theta_r)^T$ subject to constraints $A\theta = b$ and $B\theta - c \geq 0$ for matrices A and B of dimension $r' \times r$.

Maximising the likelihood of P_w for the simulated data leads to the results presented in figure 7.4. The corrected counts are more similar to the non-ascertained allele counts compared to the ascertained counts, but, there are more differences than were apparent in figure 7.3, suggesting this method may not as accurately estimate the true allele frequencies.

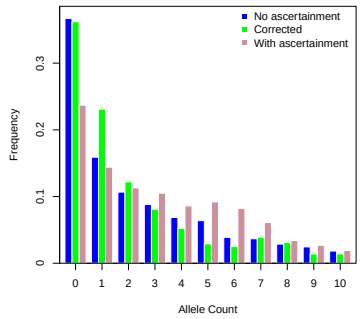


Figure 7.4: Allele frequency spectrum of one subpopulation from data simulated under an isolation model without ascertainment (blue bars), with ascertainment (pink bars) and maximum-likelihood frequencies given ascertainment (green bars).

7.1.2.3 Computing the summary statistics

The set of summary statistics can be estimated given the maximum-likelihood estimates of the allele frequencies. Ramírez-Soriano and Nielsen (2009) found estimators for θ (the scaled mutation rate) and D_t. The authors derive expressions for the expected number of segregating sites and pairwise difference given a particular ascertainment scheme. They also looked at the variance (and covariance) of the estimators and found that when the ascertainment sample size was much smaller than the sample size, estimates of θ and D_t given the ascertainment correction had a much higher variance than estimates from data with no ascertainment. As the ascertainment sample size increases towards the sample size,

the variance in the estimators decreases. This behaviour is natural, since the smaller the ascertainment sample size then the greater affect ascertainment has on allele frequencies and the more information is lost about the true allele frequencies. The approach taken here is slightly different. The statistics that are functions of the allele frequency spectrum, namely the variance of allele counts, the mean allele counts, η_1 and η_{max}, are computed directly from the reconstructed allele frequency spectrum. On the other hand, the pairwise differences, D_T and F_{st} require more attention.

Firstly, F_{st} needs to be estimated in order to estimate τ. The only information that can be retrieved is the number of SNPs with a particular frequency. Therefore, at each SNP, the allele frequencies in both subpopulations are unknown, and they are required in (5.1) to estimate F_{st}. However, equation (5.2) depends on the average heterozygosity of the whole population and the within-population heterozygosity. Given the allele counts,

$$\begin{aligned}
\bar{H} &= \frac{1}{L}\sum_{i=1}^{n^T-1} 2q_i(1-q_i)\eta_i, \\
\bar{H}_w &= \frac{1}{P}\sum_{j=1}^{P}\frac{1}{L}\sum_{i=0}^{n} 2q_{wi}(1-q_{wi})\eta_{ji} \\
&\approx \frac{1}{L}\sum_{i=0}^{n} 2q_{wi}(1-q_{wi})\eta_{wi},
\end{aligned}$$

where $\eta_i = Lp_i$ is the number of SNPs with allele count i in the total population, η_{ji} is the number of SNPs in subpopulation j with allele count i and $\eta_{wi} = Lp_{wi}$. Also, $q_i = i/n^T$ and $q_{wi} = i/n$. Therefore,

$$F_{st} \approx 1 - \frac{\sum_{i=0}^{n} 2q_{wi}(1-q_{wi})\eta_{wi}}{\sum_{i=1}^{n^T-1} 2q_i(1-q_i)\eta_i}. \tag{7.4}$$

Given the maximum-likelihood estimators of the total- and within-population allele counts, Wakeley (2009) provides an alternative formula for the mean pairwise difference π. In a

sample of size n^T,

$$\pi = \frac{1}{\binom{n^T}{2}} \sum_{i=1}^{[n^T/2]} i(n^T - i)\eta_i,$$

where

$$\left[\frac{n^T}{2}\right] = \begin{cases} \frac{n^T}{2}, & \text{if } n^T \text{ is even;} \\ \frac{n^T-1}{2}, & \text{if } n^T \text{ is odd.} \end{cases}$$

Wakeley explains that, if a locus divides the sample into i copies of one allele and $n^T - i$ copies of the other, by comparing two samples at this site, there is a difference in $i(n^T - i)$ of the $\binom{n^T}{2}$ possible comparisons. Similarly, when computing the within-population pairwise difference π_W, in the case of two subpopulations with n_i the size of subpopulation i, then

$$\pi_W = \frac{1}{\binom{n_1}{2} + \binom{n_2}{2}} \left\{ \sum_{i=1}^{[n_1/2]} i(n_1 - i)\eta_{1i} + \sum_{i'=1}^{[n_2/2]} i'(n_2 - i')\eta_{2i'} \right\},$$

and so

$$\pi_B = \frac{\sum_{i=1}^{[n^T/2]} i(n^T - i)\eta_i - \left[\sum_{i=1}^{[n_1/2]} i(n_1 - i)\eta_{1i} + \sum_{i'=1}^{[n_2/2]} i'(n_2 - i')\eta_{2i'} \right]}{\binom{n^T}{2} - \left[\binom{n_1}{2} + \binom{n_2}{2} \right]}.$$

To demonstrate how successful these methods of reconstruction are in estimating the set of summary statistics $\{\pi, \pi_B, \pi_W, D_t, \eta_1, \text{ mean}, \text{ variance}, \eta_{max}\}$, data were simulated under an isolation model without ascertainment and also with ascertainment using the same population divergence time τ. The set of summary statistics were computed under both models and also from the corrected data. This was repeated 100 times and the results are presented in figure 7.6.

Most of the distributions of statistics computed from the corrected allele frequencies are similar to those computed under no ascertainment, excluding π_B and π_W, which do show some improvement compared to the ascertained data without correction, but there is

more of a separation between the values derived from the data without ascertainment and the corrected data. This may be due to the maximum-likelihood estimates from equation (7.3) not estimating the true within-subpopulation allele frequencies well enough. These estimates may be improved by finding a better correction.

7.1.3 Estimating τ

Test III relies on adequately estimating τ and currently applies

$$\hat{\tau} = \begin{cases} \frac{F_{st}}{1-F_{st}}, & \text{if } \frac{F_{st}}{1-F_{st}} > 0.013; \\ 0, & \text{otherwise,} \end{cases}$$

with F_{st} estimated by (5.1). The value 0.013 was shown to improve the estimates produced in figure 6.12. Applying the same estimator to ascertained data leads to figure 7.5, which shows confidence bands for τ. 1000 SNPs were simulated from two subpopulations of sample size 10 and $n_a = 2$ and $\hat{\tau}$ computed. Repeating this 100 times, the endpoints of the confidence bands are taken as the lower and upper 2.5th percentiles. Evidently, ascertained data fail to lead to good estimates of τ, perhaps not surprisingly.

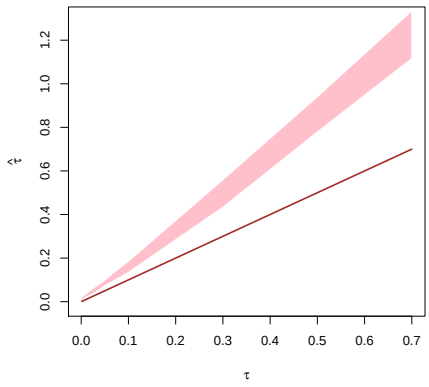


Figure 7.5: Confidence bands for τ from ascertained data.

Given the ascertainment scheme, the Markov-chain Monte-Carlo ABC algorithm used in

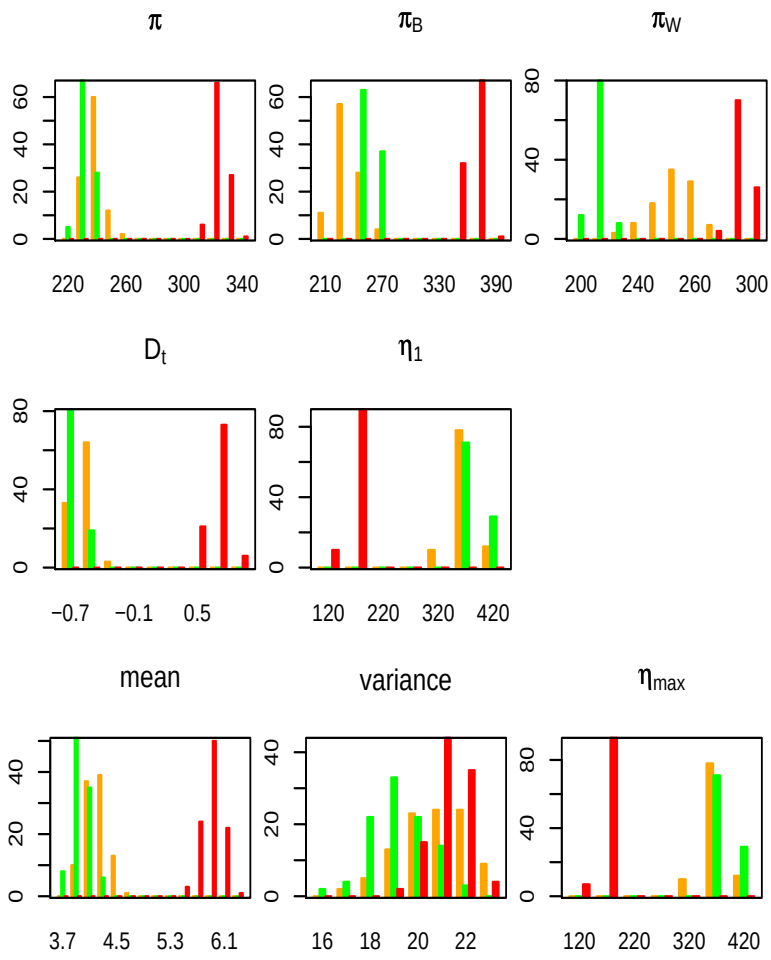
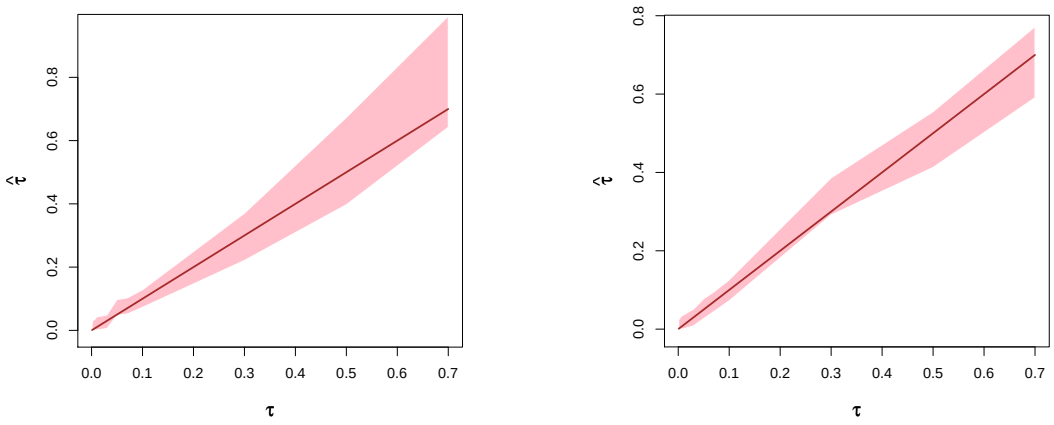


Figure 7.6: Histograms of summary statistics from data simulated under the isolation model without ascertainment (green bars), with ascertainment (red bars) and using the maximum-likelihood allele frequencies (orange bars).

section 5.3.1 can be used to estimate τ. Given an initial draw τ_0, this method iteratively simulates from a transition distribution, chosen to be Normal with mean τ_{i-1} and variance σ^2, and accepts a draw if the observed F_{st} is close to the simulated F_{st}. This algorithm can directly accommodate ascertainment by simulating from the isolation model with ascertainment. The statistic chosen in this algorithm, F_{st}, is calculated from the ascertained data and so it may be beneficial either to correct for ascertainment and calculate F_{st} using (5.2) or to use (7.4), not accounting for ascertainment. Figure 7.7 shows confidence bands for $\hat{\tau}$ calculated by taking the lower and upper 2.5 percentiles of the estimated densities

$p(\tau|F_{st})$ from the ABC_MCMC algorithm correcting and not correcting for ascertainment. Simply using the ascertainment data, confidence bands, which contain the true values of τ, are produced which are narrower compared to the corrected data. Therefore, this algorithm works well, directly accounting for ascertainment without any allele frequency corrections.



(a) (b)

Figure 7.7: Confidence bands for $\hat{\tau}$ from ascertained data (a) correcting for ascertainment and (b) not correcting for ascertainment.

7.1.4 Hypothesis test

This section aims to incorporate ascertainment directly into Test III. In its current state, ascertained data may be tested using Test III directly by correcting for ascertainment. Figure 7.6 demonstrated that, in most cases, the distributions of the statistics from non-ascertained data and corrected data are roughly similar. Excluding π_W and π_B from the set of statistics, the type I error rate is shown in figure 7.8. Most of the confidence intervals lie around 0.5, therefore, by correcting for ascertainment, Test III does not control the type I error rate.

Figure 7.1 compared the distributions of the summary statistics under the migration and

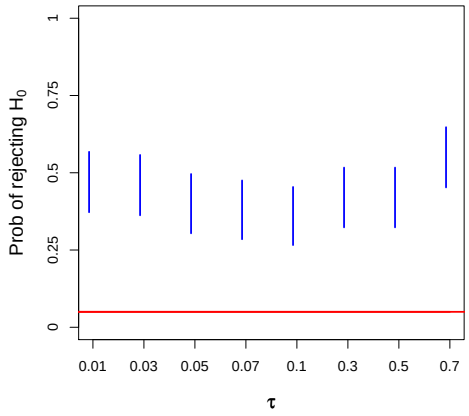


Figure 7.8: Type I error rate of Test III correcting for ascertainment.

isolation models with ascertainment and shows that most of the statistics are still able to distinguish the models excluding π_W and η_{max}. Figure 7.7 shows that the ABC_MCMC algorithm can be used to estimate the posterior distribution of τ given a summary of observed data without correcting for ascertainment. As a result, it appears practical to construct a hypothesis test in the presence of ascertainment in the following way:

Algorithm 5 (Test IV)**.**

1. Calculate the set of observed summary statistics S_{obs} and $F_{st} = F_{st_1}$ as described in the text.

2. Estimate the posterior distribution of τ given F_{st_1}, $p(\tau|F_{st_1})$ by ABC_MCMC.

3. For $k = 1, \ldots, N_{sim}$:

4. Take the kth simulated draws of τ from $p(\tau|F_{st_1})$ and simulate data under the isolation model under the same ascertainment scheme as the observed data.

5. Estimate the set of summary statistics $S_k = \{S_{k_1}, \ldots, S_{k_m}\}$.

6. For each statistic separately, test the hypothesis

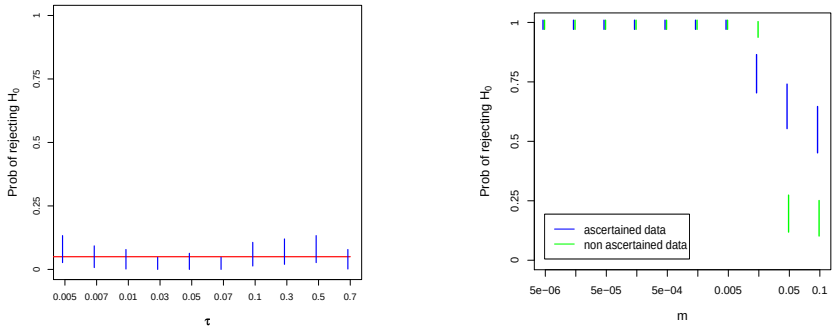
$$H_{0_i} : S_{obs_i} = \bar{S}_i$$
$$H_{1_i} : S_{obs_i} \neq \bar{S}_i,$$

where $\bar{S}_i = \sum_{k=1}^{N_{sim}} S_{k_i}$ for $i = 1, \ldots, m$.

7. Test the global null hypothesis, $H_0 = \bigcap_{i=1}^{m} H_{0_i}$.

1000 SNPs were simulated from the isolation model with two subpopulations, each of sample size 10 which diverged at time τ. The models were tested using Test IV for a range of τ values. This was repeated 100 times and the results are presented in figure 7.9(a). As in the case of non-ascertained data, the type I error is controlled. Figure 7.9(b) shows the power of Test IV, shown by the blue bars, for a range of migration rates compared to the power of Test III with non-ascertained data, shown by the green bars. As the migration rate decreases, this test shows high power in distinguishing these models, whereas as the migration rate increases, testing ascertained data through Test IV is more powerful that testing non-ascertained data through Test III. This apparent increase in power may be consequence of the choices of h and σ^2 in the ABC_MCMC algorithm. Recall that, at iteration i, a value $\tau^* \sim N(\tau_{i-1}, \sigma^2)$ and corresponding $F_{st} = F_{sim}$ are accepted if $|F_{sim} - F_{obs}|/h < 1/2$. As in any MCMC algorithm, the value of σ^2 ensures that one is neither in the situation where each proposed value of τ is accepted or each proposed value is rejected. Likewise, the value of h controls the acceptance rate of F_{st} values. Given a proposed value, h provides a balance between accepting F_{sim} too often or too infrequently. Fixing h (or similarly σ^2) too high causes difficulty in rejecting t^* when F_{obs} is much smaller than h. For instance, suppose $h = 0.1$ and $F_{st} = 0.01$. As the migration rate increases, F_{st} decreases and so the choice of in parameters becomes more crucial. The green bars in figure 7.9(b) show the power of Test III using data with no ascertainment. Data was tested from the migration model with $m = 0.05$ and no ascertainment using the ABC_MCMC algorithm to estimate τ. Repeating this 50 times, then 25 out of the 50 were rejected giving an interval for the probability of rejecting the null hypothesis of (0.36, 0.64) which seems more consistent was the ascertainment result in figure 7.9(b) (albeit this interval still lies slightly below the blue line when $m = 0.05$).

Another explanation of the increase in power may be due to the ascertainment process selecting SNPs with an intermediate allele frequency, which may be more informative in distinguishing between these two models than, for example, loci with a small minor allele



(a) (b)

Figure 7.9: Type I error rate (a) and power (b) of Test IV.

frequency. This can be explained by observing that SNPs with a low allele frequency are more likely to have a mutation on a terminal branch whereas mutations that occur on internal branches may be more revealing about the evolutionary history of the sample.

7.1.5 Other ascertainment schemes

This chapter has considered only one ascertainment scheme: with the ascertainment sample included in the final data set. However, there are many different possible schemes, some of which were discussed in section 1.3.1. In order to find the maximum-likelihood estimators of the true allele frequencies, the method of Nielsen et al. (2004) requires the probability of ascertainment given the allele count at a SNP. Therefore, this method is transferable to other ascertainment schemes given that an expression for these probabilities is available. Test IV can be applied to any ascertainment scheme that can be simulated.

7.2 Projected data

Most SNP data sets consist of a relatively small number of individuals who have been genotyped at a large number of SNPs. For example, the Human Diversity Panel described by Cann et al. (2002) now contains 1043 individuals genotyped at 600,000 SNPs, whilst the

HapMap project (International HapMap 3 Consortium (2010)) genotyped 1486 individuals and 1.4 millions SNPs. Such numbers present a challenge to the proposed test. Therefore, a method is proposed to test projected data rather than the full data.

Given a sample of n haploid genes genotyped at L SNPs with $L \gg n$, the data can be stored in a matrix C of dimension $L \times n$ with $C_{ij} = 0$ or 1 for all i, j corresponding to whether individual j carries the mutant allele at SNP i. Then, via singular value decomposition,

$$C = USV^T.$$

where

- U is a matrix of dimension $L \times L$ with the columns the eigenvectors of CC^T;
- V is a matrix of dimension $n \times n$ with the columns the eigenvectors of C^TC;
- S is a matrix of dimension of $L \times n$ with entries C_{ii} for $i = 1, \ldots, n$ equal to the n nonzero singular values of both CC^T and C^TC ordered from largest to smallest and the remaining entries equal to zero.

That is,

$$C = \begin{pmatrix} u_{11} & u_{12} & \ldots & u_{1L} \\ u_{21} & u_{22} & \ldots & u_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ u_{L1} & u_{L2} & \ldots & u_{LL} \end{pmatrix} \begin{pmatrix} s_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & s_n \\ 0 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} v_{11} & \ldots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{n1} & \ldots & v_{nn} \end{pmatrix}^T$$

$$= \begin{pmatrix} \sum_{k=1}^n s_k u_{1k} v_{1k} & \ldots & \sum_{k=1}^n s_k u_{1k} v_{nk} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^n s_k u_{nk} v_{1k} & \ldots & \sum_{k=1}^n s_k u_{nk} v_{nk} \end{pmatrix},$$

so that

$$C_{ij} = \sum_{k=1}^{n} s_k u_{ik} v_{jk},$$

and so, if only the first K components are included, where $K \leq n$, then

$$\begin{aligned} C_{ij} &\approx \sum_{k=1}^{K} s_k u_{ik} v_{jk} \qquad (7.5) \\ &= \tilde{C}_{ij}, \text{ say.} \end{aligned}$$

The value of K is chosen by formally testing which eigenvalues are significant in capturing the structure present in the data as described in section 2.1.3.

7.2.1 Estimating parameters using $\tilde{C}$

This section analyses how well $\tilde{C}$ allows estimates of the statistics needed to test the consistency of the data with an isolation model. 2000 SNPs were simulated under the isolation model with two subpopulations each of size 10 and $\tau = 0.3$. Principal components analysis was performed and the first two eigenvalues were significant in detecting population structure. In 100 simulations, the majority found two significant eigenvalues. Therefore, for each statistic, results are compared using both the full data and $\tilde{C}$ with $K = 2$ and also, for comparison, $K = 1$.

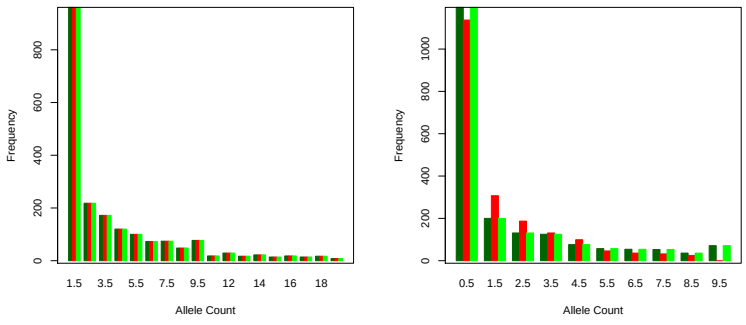
Figure 7.10 shows the allele frequency spectra from the three matrices, both for the total population and within subpopulations. The frequencies in the total population are well estimated when $K = 1$ and 2, whereas, using $K = 1$, the frequencies within subpopulation one are slightly different from using the full data and $K = 2$, which appear almost identical.

The population divergence time was estimated using the full data and also using $\tilde{C}$ with $K = 2$ and the results are given in figure 7.11. Using $K = 1$ proved to be inadequate for estimating τ. The distributions of $\hat{\tau}$ using the full data and $\tilde{C}$ with $K = 2$ are almost indistinguishable.

Given the estimated allele frequencies, the set of summary statistics π, π_W, π_B, D_t, η_1, mean, variance and η_{max} are estimated using the full data and using $\tilde{C}$ with $K = 1$ and 2. The results are presented in figure 7.12. For all the statistics, comparing the light and dark green bars, using $K = 2$ in $\tilde{C}$ leads to good estimates of the distributions of these statistics, whereas some of the statistics, in particular π_W and π_B, are badly estimated using $K = 1$.

7.2.2 Minimal data size

Patterson et al. (2006) define the data size to be the number of SNPs multiplied by the total number of (diploid) individuals. In a sample from two subpopulations each of haploid size n and L SNPs, the data size is $2(n/2)L = nL$. The test for population structure assumes the statistic x (using the notation of Patterson et al. (2006)), a function of the largest eigenvalue of CC^T, follows a Tracy-Widom distribution under the null hypothesis of no population structure. However, they discuss a minimal data size needed in order for population structure to be found. More precisely, if the two populations diverged at time τ with time measured in N^T generations, they find a minimal F_{st} for which it is possible to detect population structure and for F_{st} values which fall below the threshold, it is not



(a) (b)

Figure 7.10: Allele frequency spectra from 2000 SNPs simulated under the isolation model with two subpopulations using the full data (dark green bars), $K = 1$ (red bars) and $K = 2$ (light green bars) from the total population (a) and within subpopulations (b).

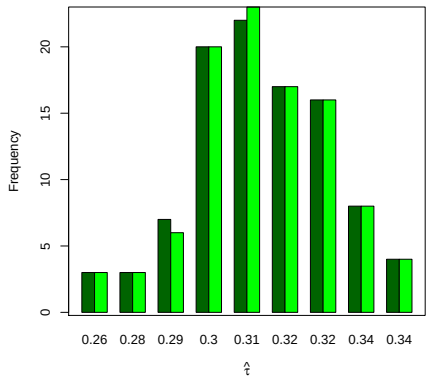


Figure 7.11: Histograms of $\hat{\tau}$ from data simulated under the isolation model using the full data (dark green bars) and using $K = 2$ (light green bars).

possible to detect population structure.

1. $F_{st} \approx \tau$, and

2. the threshold is reached when
$$\tau = \frac{1}{\sqrt{nL}}.$$

The authors explain that when $\tau < \frac{1}{\sqrt{nL}}$, then the largest eigenvalues of the theoretical and observed covariance matrices are different and so the distributional results may not hold.

This property affects the estimation of summary statistics in section 7.2.1. Fixing $\tau = 0.001$ and $L = 21,000$ then this would require $n > 47$. Consider π_W, simulating data under the isolation model with a range of haploid sample sizes in each subpopulation and estimating π_W using the estimated data from equation (7.5). The results are presented in figure 7.13.

For large sample sizes, the red line, showing π_W estimated from $\tilde{C}$, and the green line, showing π_W estimate using the full data, appear to be quite similar. However, when $n = 10$, the difference between the two estimates is 311 whereas when $n = 70$, the

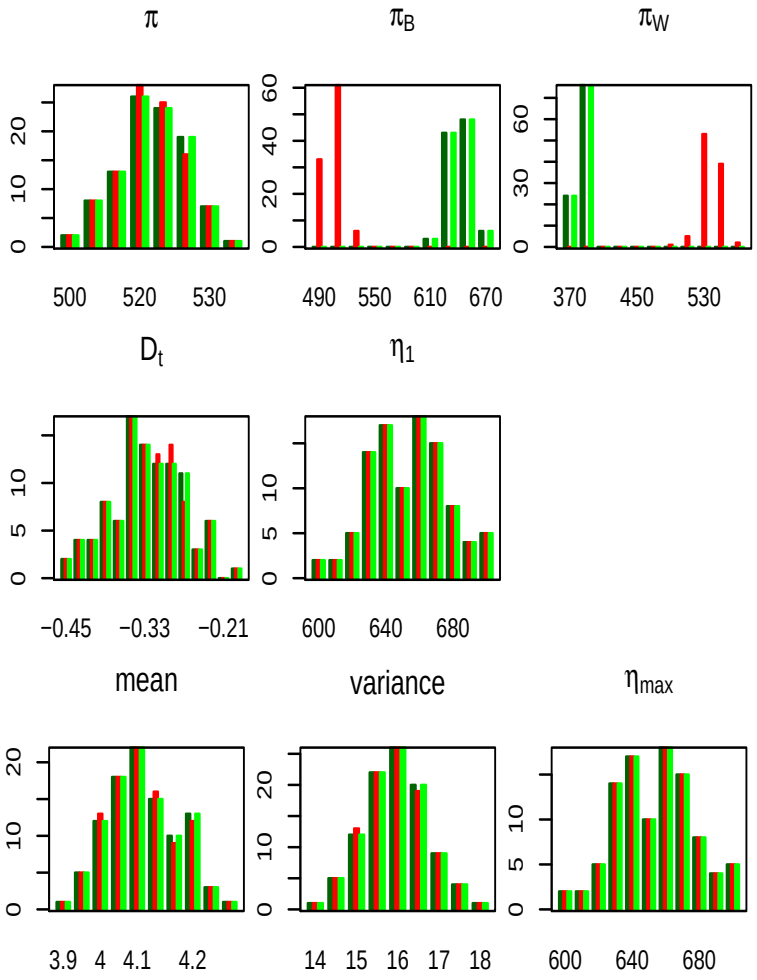


Figure 7.12: Histograms of summary statistics from data simulated under the isolation model using the full data (dark green bars), $K = 1$ (red bars) and $K = 2$ (green bars).

difference is only 12. Therefore, when performing the hypothesis test using $\tilde{C}$, one must ensure that the observed F_{st} lies above the threshold value, determined by the data size.

7.2.3 Hypothesis test

Assessing whether using the projected data decreases the power of Test III, the power and type I error rate from this test are compared to those produced by using Test V.

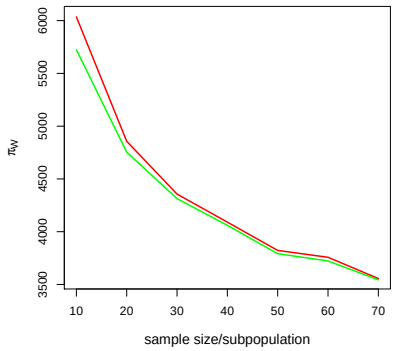


Figure 7.13: Estimating π_W from simulating data for a range of sample sizes. Red and green lines shows π_W using equation (7.5) and full data, respectively.

Algorithm 6 (Test V)**.**

1. *Given the observed data, find the number of significant eigenvalues* K*.*
2. *Estimate matrix* $\tilde{C}$ *given* K*.*
3. *Estimate* $F_{st} = F_{st_1}$ *and then* $\hat{\tau}$*.*
4. *Find summary statistics* S_{obs}*.*
5. *Set* $N_{acc} = 0$ *and while* $N_{acc} < N_{sim}$ *:*
   6. *Simulate* $T \sim N(\hat{\tau}, \sigma^2)$*. If* $T \geq 0$
   7. *Simulate data from the isolation model with two subpopulations diverging at time* T *and calculate* $\hat{F}_{st} = F$*.*
   8. *If* $|F - F_{st_1}| < \epsilon$*, calculate the set of summary statistics and set* $N_{acc} = N_{acc} + 1$*.*
9. *For* $i = 1, \ldots, m$*, test the hypotheses*

$$H_{0_i} : S_{obs_i} = \bar{S}_i$$
$$H_{1_i} : S_{obs_i} \neq \bar{S}_i$$

where $\bar{S}_i = \sum_{k=1}^{N_{sim}} S_{k_i}$ *for* $i = 1, \ldots, m$*.*

7. *Test the global null hypothesis,* $H_0 = \bigcap_{i=1}^{m} H_{0_i}$*.*

Data were simulated from two subpopulations each of haploid sample size 10 at 2000 SNPs. The power of both tests was assessed for a range of migration rates and the type I

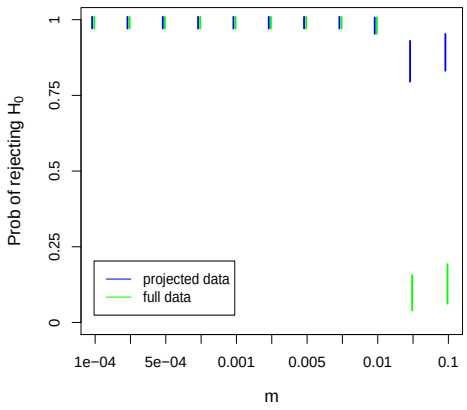


Figure 7.14: Comparison of the power of Test III and Test V.

error rate for a range of divergence times.

The power of this hypothesis test, compared to testing the full data, is given in figure 7.14. The green bars show the results using the full data and the blue bars show the results using the projected data. Test V preserves the power of Test III. However, as the migration rate increases, testing the projected data seems more powerful than using the full data. In this simulation of 2000 SNPs and haploid sample size of 20, and so a diploid sample size of 10, equation (3.11) may be used to give a rough estimated of the minimal value of M, the scaled migration rate, for which the distributional results hold. If $M = 2Nm$, since time is measured in $2N$ generations, then given

$$F_{st} \quad > \quad \frac{1}{\sqrt{10 \times 2000}} = 0.007,$$

then $M > 35$ (approximately). In this simulation, $M > 35$ corresponds to $m > 0.035$ which accounts for the increased power of Test V. Therefore, if the observed data produces a value of F_{st} below the threshold, the test is not reliable.

7.2.4 Higher dimensional data

The idea of performing the hypothesis test using the projected data is to better handle high dimensional data. As previously noted, many SNP data sets genotype in excess on 1 million SNPs, although many of the SNPs may not be independent if they are positioned close enough in the genome. The hypothesis test and ABC estimator of τ requires iteratively simulating data to the same size as the observed data, which is computational infeasible if the observed data contains around 1 million SNPs.

Suppose $L > 1,000,000$, then a value $L_1 \ll L$ may be defined such that, at each step in the hypothesis test, L_1 SNPs are simulated instead of L SNPs. Many of the test statistics are dependent on the number of SNPs, namely π, π_W, π_B, η_1 and η_{max}. In order to compare the observed statistics S_{obs} with the simulated statistics S_{sim}, it is possible to re-scale those statistics which are dependent on L. Given a statistic X, then

$$X' = \frac{X}{L} L_1.$$

In simulating L_1 SNPs, steps 7 and 9 in Test V are slightly altered. Step 7 requires simulating L_1 SNPs under the isolation model and if $|F - F_{st_1}| < \epsilon$, step 8 calculates

$$S'_{sim} = \{\pi', \pi'_W, \pi'_B, D_t, \eta'_1, \text{ mean, variance, } \eta'_{max}\}.$$

Lastly, high dimensional data can be handled when data has been ascertained by altering steps 4 and 5 in Test IV in a similar manner.

7.3 Discussion

This section has incorporated two aspects of real data into Test III given in chapter 6. In particular, Test IV was shown to be powerful in distinguishing the two models and control the Type I error rate as shown in figure 7.9. This test directly built in ascer-

tainment by altering steps 1 and 4 of Test III by estimating the population divergence time using the ABC_MCMC algorithm, given in section 5.3.1, and simulating data under an ascertainment model, similar to the observed data, and comparing the distributions of the simulated summary statistics to the observed values. Test V tests projected data, rather than the full data, in attempt to better handle high dimensional data. This test begins by estimating the data matrix C, using the notation of Patterson et al. (2006), using the components found to be significant in capturing the population structure in the data and then follows the steps of Test III. It is possible to extend these analyses by combing the steps of Test IV, which handle ascertainment, and the steps of Test V, which handle high dimensional data. This test would firstly estimate the matrix C using the significant components but then compare the reduced observed data to data simulated with ascertainment.

Chapter 8

An example from the HapMap project

To illustrate the use of the hypothesis test developed in the previous two chapters on real data, samples from populations involved in the HapMap project are subjected to the test in this chapter.

8.1 Description of data

A brief overview of the international HapMap project was given in section 1.4. The project was delivered in three phases between 2007 and 2010. A detailed description of each phase is given in this section.

The first phase consisted of samples with ancestry from parts of Africa, Asia and Europe and contained a total of 269 DNA sequences with 90 of the individuals from Utah thought to have European ancestry (CEU), 90 Yoruba individuals from Nigeria (YRI), 44 from Tokyo (JPT) and 45 from Beijing (CHB). Ancestry was determined differently for each sample. The Yoruba samples were required to have four Yoruba grandparents, the Chinese

donors required at least three Chinese grandparents, whereas the Japanese sample self–identified themselves as Japanese. The criteria used to define the Utah sample are unknown and so it is unclear how well this sample represents northern and western European ancestry.

The CEU samples were collected in 1980 when controlling for recent ancestry in population genetics studies was not as prevalent as it is now, as discussed by He et al. (2009), who compared the CEU sample to data from 81 populations with northern, eastern, southern and western European ancestry. Computing pairwise distances between populations, a measure related to pairwise F_{st}, they found the CEU sample to be more genetically similar to those samples from western Europe.

Each individual was genotyped at just over one million SNPs with the initial goal to genotype at least one SNP every 5kb across the genome with a minor allele frequency greater than 0.05 in the sample. In the second phase, the International HapMap Consortium (2007) improved the coverage of the genome by genotyping the 269 samples at 3.1 million SNPs. In the third phase, the International HapMap 3 Consortium (2010) genotyped samples from an additional seven populations, bringing the total sample size to 1184 from eleven populations. A summary of the eleven populations is given in table 8.1. The first column shown the location where the samples where taken, the second column the ancestral population of the samples, the third column the sample size and the last column the abbreviations chosen by the HapMap project. In total, in the sample of 1184 individuals, just over 1.4 million SNPs were found.

Not all individuals in the HapMap samples are unrelated. Samples from YRI, CEU, MKK, ASW and MEX are made up of trios. That is, the project included mother, father and child samples and so not all the samples are independent.

Principal components analysis, as described in section 2.1.3, was performed on a sample consisting of 50 individuals from each of the eleven phase 3 populations at just under 40,000 SNPs from chromosomes one to five. The first two components are plotted in figure 8.1.

Table 8.1: Summary of HapMap 3 samples.

Population	Ancestry	Sample Size	HapMap Abbreviation
Utah, USA	northern/western Europe	165	CEU
Beijing, China	China	85	CHB
Tokyo, Japan	Japan	86	JPT
Ibadan, Nigeria	Yoruba	167	YRI
USA	African	83	ASW
Colorado, USA	China	84	CHD
Texas, USA	Indian	88	GIH
Webuye, Kenya	Luhya	90	LWK
Kinyawa, Kenya	Maasai	171	MKK
California, USA	Mexico	77	MXL
Italy	Italy	88	TSI

The pattern mimics that shown in figure 2.4(a), which plots the first two components from the HGDP-CEPH panel. Generally, African populations are clustered together to the left, Asian populations to the right and European populations in the bottom centre of the plot. These continental labels are meant only illustratively to show the similarities in results between the two data sets: the International HapMap 3 Consortium (2010) are clear, in their supplementary information, that the samples from these populations are not meant to be representative of the larger populations or continent.

8.1.1 SNP discovery

SNPs were discovered in several different ways as documented by the International HapMap Consortium (2003). In the first instance, the dbSNP database, a catalogue of genomic variation in five species, namely human, mouse, rat, chimpanzee and the malaria parasite, Sherry et al. (2001), provided information on known variable positions in the human genome. At the beginning of the HapMap project, the database, release 118, contained around 2.8 million SNPs but the positions of the SNPs were not spread uniformly across chromosomes, leaving some sparse areas in the genome. Additional SNPs were discovered using whole genome and whole chromosome shotgun sequencing. The genome, or chro-

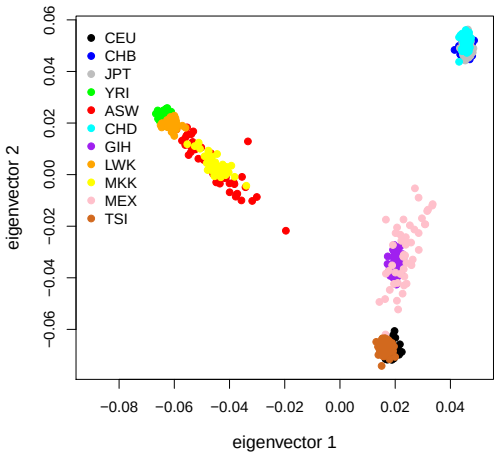


Figure 8.1: Plot of first two principal components from a subset of phase 3 HapMap samples and SNPs.

mosome, is broken up in random fragments, which are sequenced and reassembled by matching the overlaps in the fragments as described by Adams (2008), who provided a brief description of shotgun sequencing via a figure similar to that given in figure 8.2. In this figure, the purple line represents the part of the chromosome, or genome, to be sequenced. The fragment is divided into smaller fragments shown by the red lines and each red line is broken into smaller fragments to be sequenced and then reassembled through matching overlapping sequences. In the HapMap project, the samples used in the shotgun sequencing were from populations different from those included in the genotyping part of the project. The whole–genome shotgun sequencing used a pool of eight samples, whereas each chromosome was sequenced using only one or two individuals from a pool containing five individuals.

In addition, the HapMap project selected SNPs from 10 ENCODE regions. The ENCODE Project Consortium (2007) (ENCyclopedia Of DNA Elements) aimed to investigate in detail around 1% of the human genome. The idea of this project was to study randomly select regions of the human genome to assess the regions' functionality. In phase one of HapMap, 48 unrelated individuals from the original four populations lead to the ascer-

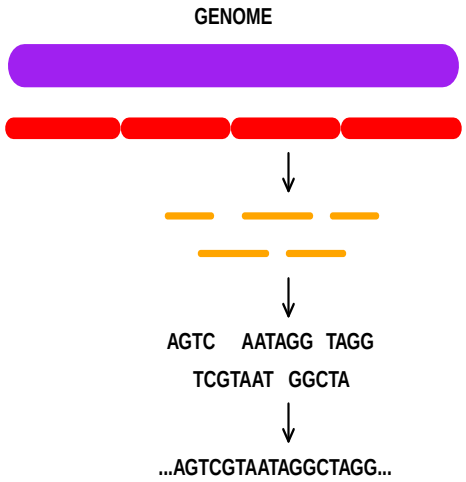


Figure 8.2: Schematic of shotgun sequencing.

tainment of around 20,000 SNPs. Data from the ENCODE regions and data gained from resequencing was compared by the International HapMap Consortium (2005) and it was found that the effects of ascertainment in the ENCODE regions were minimal due to the much larger size of the ascertainment panel. The second phase of HapMap increased the number of SNPs by (i) genotyping at SNPs identified by Hinds et al. (2005) who examined 71 unrelated individuals with African, European and Asian ancestry at around 1.5 millions SNPs and also by (ii) genotyping the additional SNPs in dbSNP release 122.

The complicated mechanisms of SNP discovery have a profound effect on the allele frequencies. Figure 8.3 shows the allele frequencies from data consisting of 50 individuals from the eleven populations at just under 40,000 randomly selected SNPs from chromosomes one to five compared to the expected frequencies under a neutral model. This demonstrates the inflated number of SNPs in the HapMap project with a high or intermediate allele frequency compared to what is expected under a neutral model, although these difference may also be due to departures from the neutral model, for example population structure.

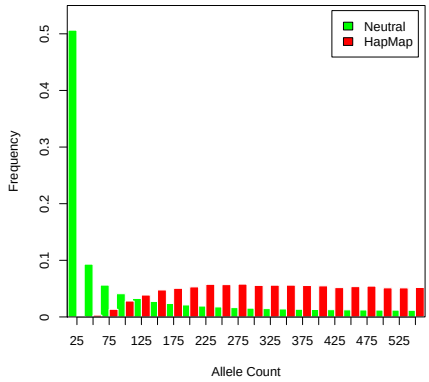


Figure 8.3: Allele frequency spectrum of a proportion of HapMap data the eleven populations compared to expected frequencies under a standard neutral model

8.1.2 Initial analysis of HapMap data

Data from HapMap 3 from the Yoruba, Japan and both Kenyan populations will be used. The two Kenyan populations will be analysed since they are geographically closer and then, as a contrast, the Yoruba and Japanese populations are analysed.

Since samples from the Yoruba and one of the Kenyan populations consist of trios, the child from each trio was removed. Once the dependent individuals were removed, the sample sizes of YRI and MKK were reduced to 114 and 143, respectively.

Principal components analysis was performed on data from the two Kenyan populations MKK and LWK using only chromosome 1, consisting of approximately 25,000 SNPs, and the results are given in figure 8.4(a). In addition, the allele frequency spectra for the combined populations and for each population separately are presented in figures 8.4(b),8.4(c) and 8.4(d). Similarly, principal components analysis was performed using chromosome 1 data from populations YRI and JPT and the allele frequency spectrum of each of the two populations is given in figure 8.5. The main conclusions from these summary plots are that the MKK samples show more variability in component values, but, more importantly, all four populations show signs of non-neutral model spectra, most likely as a result of SNP

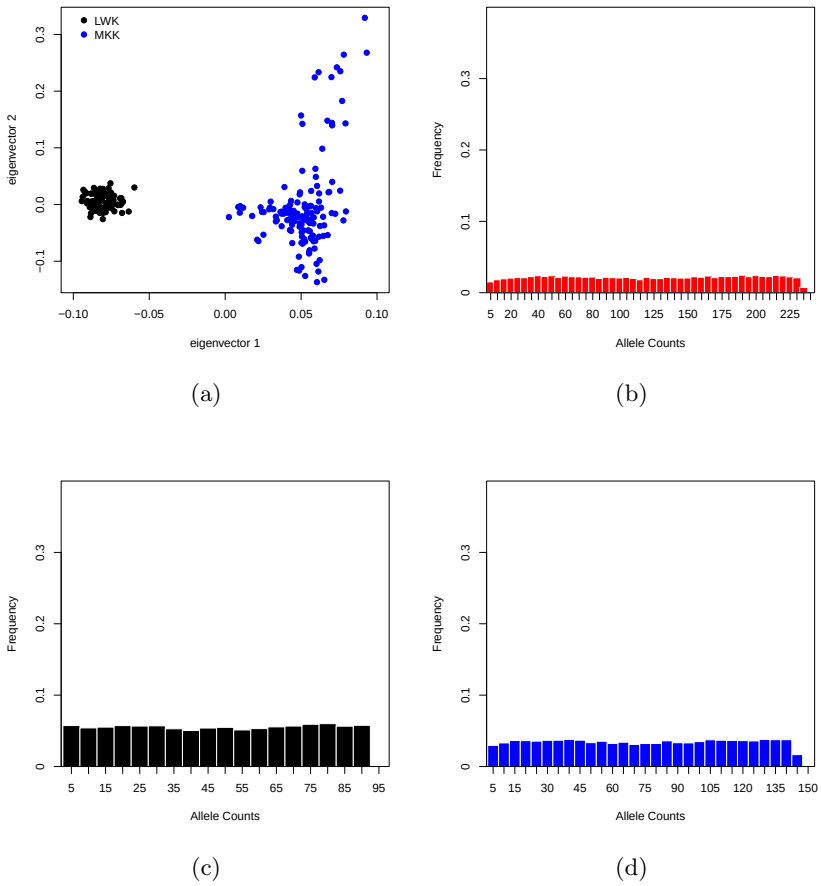


Figure 8.4: Plot of (a) first two principal components, (b) the allele frequency spectrum of SNPs from the combined samples and the allele frequency spectra of each of the two Kenyan populations ((c) MKK and (d) LWK).

ascertainment, which should be accounted for when testing hypotheses about migration versus isolation.

YRI and JPT F_{st} between populations YRI and JPT was estimated to be $\hat{F}_{st_1} = 0.171$ and $\hat{F}_{st_2} = 0.031$ between LWK and MKK. The International HapMap 3 Consortium (2010), supplementary information, provide intervals for each pairwise F_{st}. They reported these to be $(0.000, 0.031)$ between LWK and MKK and $(0.179, 0.206)$ between YRI and JPT. $\hat{F}_{st_2}$ lies just within the former interval, whereas $\hat{F}_{st_1}$ lies just outwith the interval, but the estimates are roughly consistent with the HapMap results.

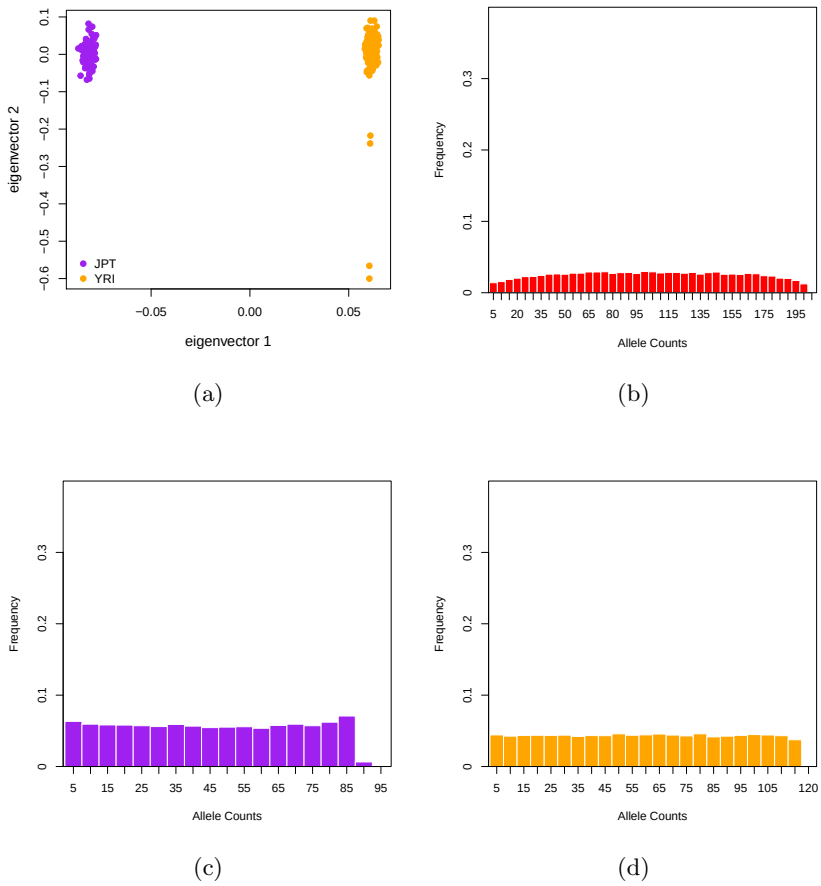


Figure 8.5: Plot of (a) first two principal components, (b) the allele frequency spectrum of SNPs from the combined YRI and JPT samples and the allele frequency spectra of each of the two populations separately ((c) JPT and (d) YRI).

8.1.3 Strategy for simulating data with ascertainment

In order to perform the hypothesis test about the demography of two HapMap populations, allele counts need to be simulated under the null model incorporating, at least to a first approximation, the ascertainment scheme present in the observed data. However, HapMap ascertained SNPs through several methods that prove difficult to imitate in detail.

Nielsen et al. (2004) devised ways of correcting for ascertainment for a range of schemes including of the "double-hit" ascertainment scheme similar to that in the HapMap project.

This ascertainment scheme has two steps. In the first step, in a population, a small ascertainment sample is used to find SNPs and in the second step, in those positions found to be variable, the process is repeated using another small ascertainment sample. This scheme is devised to find observed SNPs in two separate studies. They assumed that the ascertainment panel sizes for each of the two studies are known and the two ascertainment samples are disjoint but drawn from the same population. Lastly, they assumed that both samples are contained in the final sample. This scheme shows some similarities to the procedures described in the previous section.

Here, ascertained SNPs are simulated using a simplified scheme. Simulating data from two populations assuming that the populations diverged at some time from a common ancestor population, a small ascertainment panel, of size n_{asc} from each population is firstly genotyped and if the sample is variable, then the remainder of the sample is included in the final data. Figure 8.6 illustrates the process with two samples of haploid size 9 and $n_{asc} = 2$. Firstly, a genealogy is simulated with a total sample size of $2(n + n_{asc})$, with the green and orange dots corresponding to samples from populations one and two, respectively, and the blue dots showing the ascertainment sample. A Poisson number of mutations is randomly placed on the tree; in this case the red square shows a single mutation. The ascertainment sample is variable since one of the blue dots is affected by the mutation and so the remainder of the sample (the green and orange samples) form the final data set.

This simplified scheme does capture aspects of the actual methods used in that the ascertainment panel are not included in the final data although, as with Nielsen et al. (2004), it is assumed that the genotyped sample and the ascertainment sample are from the same populations. Otherwise, assumptions would needed to be made about the demography of more than two populations namely the populations whose demography is of interest, but also the populations from which that ascertainment sample are taken. There was further inconsistency with the sizes of the HapMap ascertainment panel ranging between 1 and 8 individuals. The scheme used here fixed an ascertainment sample of size two diploid

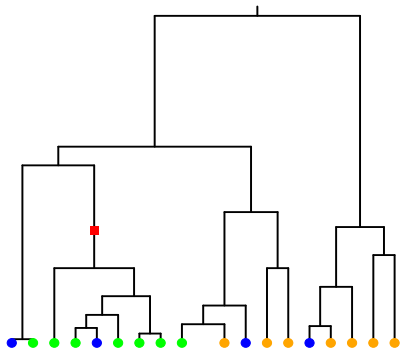
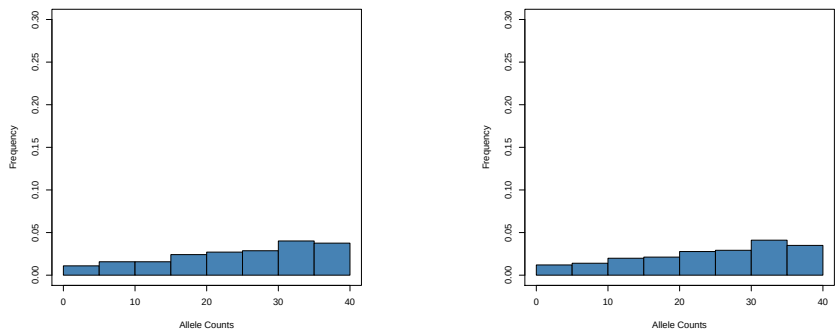


Figure 8.6: Example of coalescent tree with a total sample size of $2 \times 9 = 18$ and ascertainment size of $2 \times 2 = 4$.

individuals from each population.

Lastly, the simulation requires a specification of the population size of each of the four populations. The simulation procedure presented in section 3.4 assumed that the sample is from two subpopulations that diverged from a common ancestor, at some time in the past, and since have evolved independently and that each population, that is the ancestral population and the two subpopulations, are of the same population size N. This assumption is likely to be quite unrealistic in the human population. Park (2011) estimated the effective population size of each of the eleven HapMap populations in a range of ways and compared the results to previous estimates (where the effective population size is defined by Wakeley (2009) as the size of the Wright-Fisher population needed to produce the same level of genetic drift). In particular, Park estimated the effective population size of LWK and MKK to be 1502 and 1067, respectively, and produced estimates of 5101 and 3541 for populations YRI and JPT, respectively. Therefore, when modelling LWK and MKK, the diploid population size of each was set to 1500 and the diploid effective population size of JPT and YRI set to be 4000, when modelling the history of these populations.



(a) (b)

Figure 8.7: Allele frequency spectra of the 1500 SNPs used to test LWK and MKK (a) and the 1500 SNPs used to test YRI and JPT (b).

8.2 Hypothesis test

A sample of size 20 diploid individuals was sub-sampled from each of the four population samples of interest and 1500 SNPs sampled from chromosome 1. Considering only the 1500 SNPs for each pair of populations, figure 8.7 shows that, in both cases, these SNPs are influenced by ascertainment, showing similar patterns to those presented in figures 8.4 and 8.5. Test IV was used to compare the migration and isolation models of the two pairs of subpopulations (one pair LWK and MKK, and the other YRI and JPT), with ascertained data simulated according to section 8.1.3.

8.2.1 Results for LWK and MKK

The 1500 common SNPs between the two populations were randomly chosen from chromosome 1. Pritchard and Przeworski (2001) define a measure of pairwise linkage disequilibrium, R^2. This value was calculated between each pair of SNPs and the results given in figure 8.8. This heat map was produced using 'LDheatmap' package in R, written by Shin et al. (2006). The upper triangular heat map composes of R^2 values between SNPs. Values are coloured from light grey showing little correlation, blue showing an intermediate correlation and red showing high correlation. The right hand side column shows the

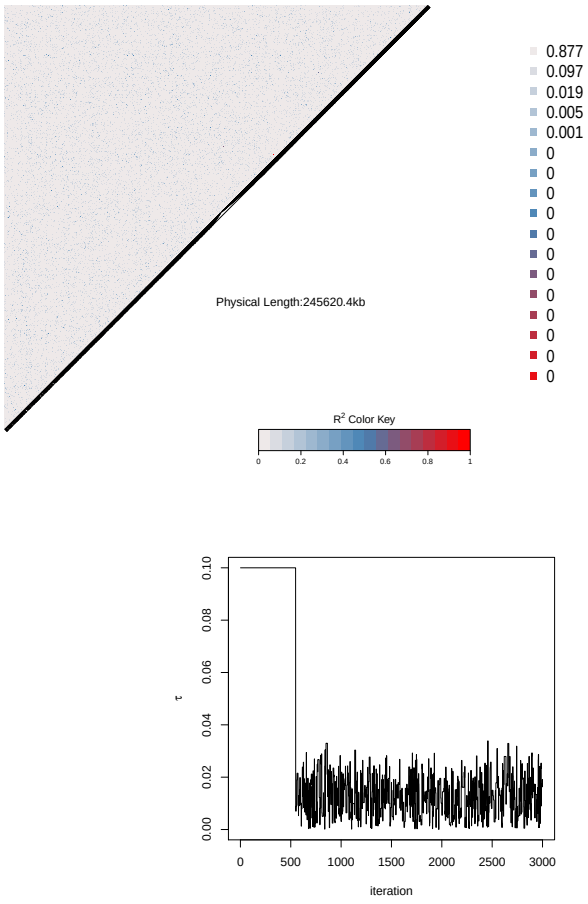


Figure 8.8: Linkage disequilibrium plot of the 1500 SNPs and proportion of pairs falling into each R^2 band (top). Trace plot for τ from ABC_MCMC algorithm for LWK and MKK (bottom).

number of pairs falling into each category. This figure contains low values of correlation, with colours ranging from light grey and blue. 88% of values fall into the lowest category corresponding to R^2 values less than 0.06. Therefore, it is assumed that SNPs used in this analysis are independent.

Testing whether these two populations diverged from a common ancestor firstly requires estimating a population divergence time τ using the ABC_MCMC algorithm with the results given in figure 8.8. Beginning the algorithm at an arbitrary τ value, the chain

appears to have converged within 1000 iterations. However, this is not to suggest that only 3000 iterations is an adequate run length and it may be advantageous to allow the chain to run for longer. The algorithm was repeated a second time, using a different initial τ, producing similar results. Taking the average value of the last one thousand draws gives $\hat{\tau} = 0.013$.

The distributions of the eight summary statistics under the isolation model are estimated and the results are given in figure 8.9. The red dot on each plot shows the observed value of the relevant summary statistic. The corresponding p-values are 0.475 (π), 0.356 (π_B), 0.515 (π_W), 0.475 (D_t), 0.139 (η_1), 0.545 (mean), 0.086 (variance) and 0.126 (η_{max}). Therefore, since each individual hypothesis is not rejected, the global hypothesis is not rejected suggesting that these data are consistent with an isolation model with $\hat{\tau} = 0.013$, therefore concluding these two populations diverged from a common ancestor around 2000 years ago (since $0.013 \times 2 \times N \times 25$ year $= 0.013 \times 2 \times 3000 \times 25 = 1950$.)

8.2.2 Results for YRI and JPT

A similar linkage disequilibrium analyses was used between populations YRI and JPT. The results are given in figure 8.10. 80% of the pairs od SNPs had an R^2 value less than 0.06, with the remaining pairs falling into the first few categories. It is assumed that the SNPs used are independent.

Estimating a population divergence time between these two populations, the ABC_MCMC algorithm was used and the results, showing only 1400 iterations, are presented in figure 8.10. After only a few iterations, the algorithm appeared to have converged to the required posterior distribution and beginning the algorithm from a different inital τ produced similar results. Taking the average value of the last 1000 draws gives $\hat{\tau} = 0.1$.

The distributions of the eight summary statistics under the isolation model are estimated and the results are given in figure 8.11, again with the red dot on each plot showing the observed value of the summary statistic. The corresponding p-values are < 0.001

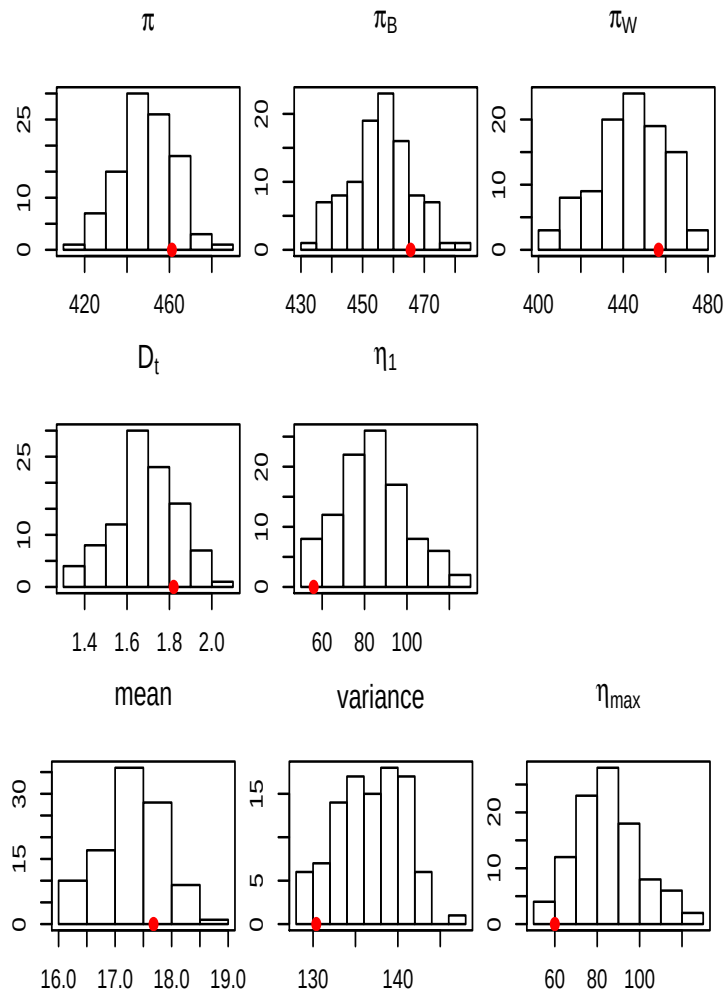


Figure 8.9: Simulated summary statistic distributions from populations LWK and MKK under the null hypothesis, with the observed value of the statistic shown as a red dot.

(π), < 0.001 (π_B), < 0.001 (π_W), < 0.001 (D_t), > 0.001 (η_1), < 0.001 (mean), 0.24 (variance) and < 0.001 (η_{max}). Therefore, since most of the hypotheses are rejected, with the exception of that based on the variance, the global hypothesis is rejected suggesting that these data are not consistent with an isolation model with $\hat{\tau} = 0.1$. In a hypothesis space consisting of just the isolation and migration models, we would then focus attention on the migration models.

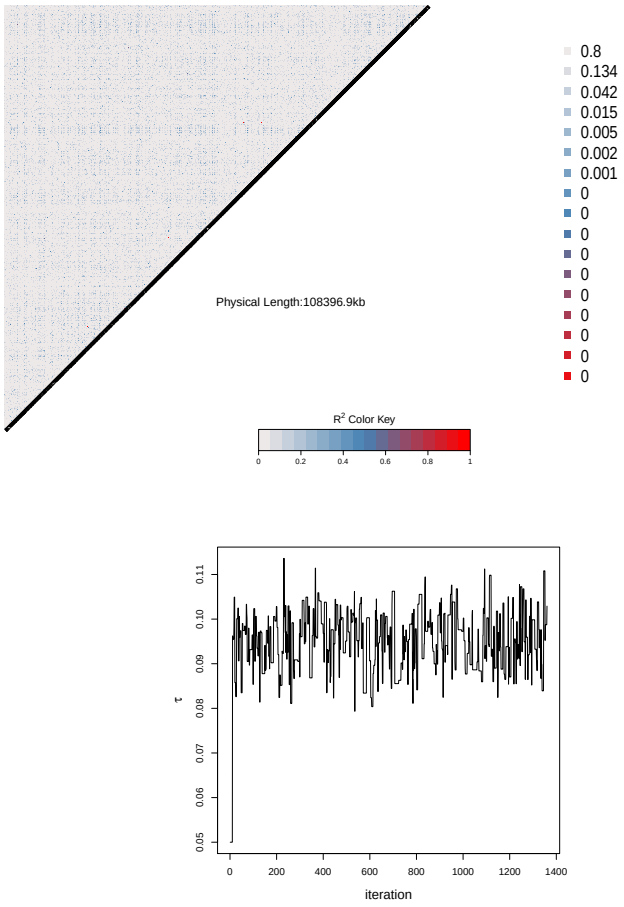


Figure 8.10: Linkage disequilibrium plot of the 1500 SNPs and proportion of pairs falling into each R^2 band (top). Trace plot for τ from ABC_MCMC algorithm for YRI and JPT (bottom).

8.2.3 Improvements

This chapter aimed to illustrate the use of the Test IV using data from the HapMap project. Comparing two Kenyan populations, the null hypothesis, stating that two populations are consistent with an isolation model, was accepted whereas comparing populations YRI and JPT, the isolation model was rejected. The observed data tested consisted of only 1500 SNPs from chromosome one and 20 diploid individuals from each population. A more thorough examination of the data would include more samples from each population

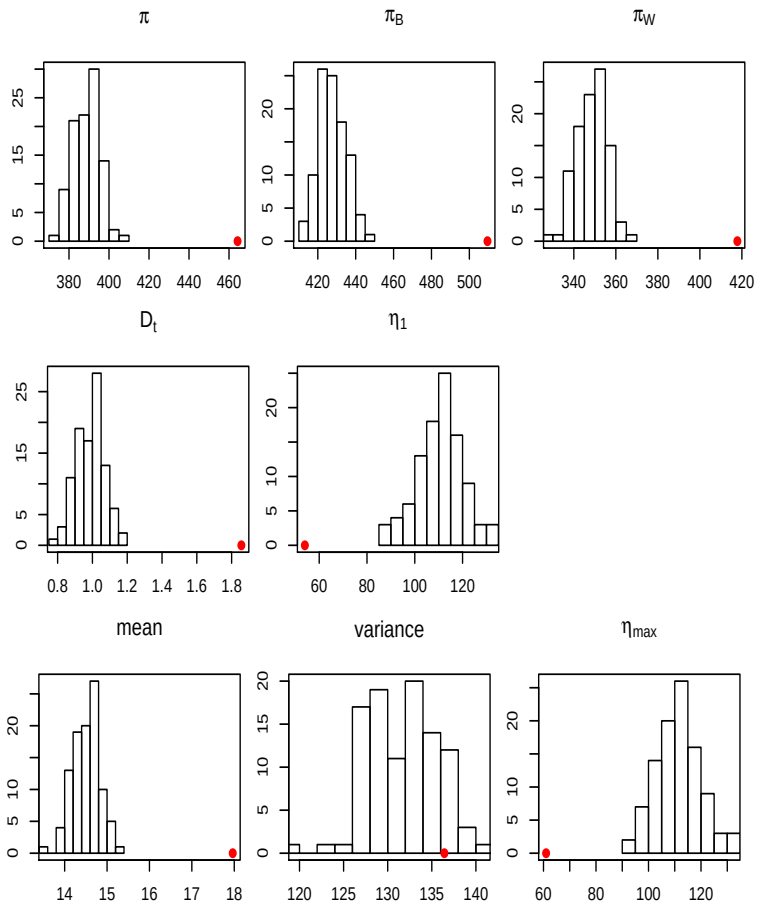


Figure 8.11: Simulated summary statistic distributions from populations YRI and JPT under the null hypothesis, with the observed value of the statistic shown as a red dot.

and SNPs from more than one chromosome. These results are counterintuitive since the Kenyan populations are geographically closer, compared to the other tested pair of populations, and so migration events are more likely than migration events between the Yoruba and Japanese populations. The purpose here was one of illustration rather than a deep exploration of the ancestry of these particular populations, however, by using a more substantial amount of the available data by expanding section 7.2.4 may produce more accurate results. This test also assumed the SNPs used were independent, an additional step may be to perform a similar test, accounting for the dependencies between SNPs.

Chapter 9

Discussion and conclusions

9.1 Summary

Barbujani and Bertorelle (2001) describe the demographic history of Europe based on genetic data and archeological records, dating primitive tools back to around 40,000 years ago. This type of analysis is incomplete in that it relies on the discovery and dating of such artifacts. Analysis of genes potentially can provide a more exhaustive explanation of modern human settlement and movements over time. The authors suggested that it may be possible to reconstruction the structure found in the the human population and they relate patterns of genetic diversity with documented historical events and archeological records. Although historical demographic characteristics impact genetic variation, the complex nature of diverse populations can often be modelled by many different evolutionary histories.

This thesis develops a hypothesis test which aims to distinguish between two demographic models in order to make inferences about historical events using SNP data. The migration model assuming the population is divided into two subpopulations that exchange migrants arbitrarily far back in the past and the isolation model assuming that two subpopulations diverged from a common ancestral population after which time they evolved independently.

The method employed tests the global null hypothesis

$$H_0: \text{ observed data are from an isolation model}$$

via a set of summary statistics.

Under H_0, the population divergence time τ of two subpopulations is estimated using F_{st}. Chapter 5 begins by assessing three possible estimators of F_{st} and then two different estimators of τ given F_{st}. Figure 5.4 provides interval estimates of τ and shows the most successful approach, out of those examined, is to use the F_{st} estimator of Reynolds et al. (1983) and to set

$$\hat{\tau} = \frac{\hat{F}_{st}}{1 - \hat{F}_{st}}.$$

Although this method was the most successful amongst those considered, it performed poorly for values of $\tau > 1$ and $\tau < 0.0005$. The explicit aim of this thesis was not to find the best estimator of such population parameters. However, inadequately estimating τ proved to inflate the type I error rate, shown in figure 6.7. In order to overcome this problem, section 6.3.1.1 aimed to find a threshold value δ defined such that if $\hat{\tau} < \delta$ then it is assumed the data are more consistent with an unstructured model and so $\hat{\tau}$ is set to 0. On the other hand, values of $\tau > 1$ correspond approximately to values of $F_{st} > 0.5$. Table 2.1 provides pairwise estimates of F_{st} for 19 HGDP populations in the range $(0, 0.35]$ and so only values of τ less that one are likely to be of relevance in human populations. That is not to say that larger τ values might not occur in modelling other species.

The hypothesis test adopts a parametric bootstrap approach and requires specification of a further two parameters, σ^2 and ϵ. Given $\hat{\tau}$, a draw $T \sim N(\hat{\tau}, \sigma^2)$ is made and data simulated under an isolation model with population divergence time T. F_{st} is estimated and if the absolute difference between the observed F_{st} and simulated F_{st} is less than δ, then the set of summary statistics is computed from the simulated data. The global hypothesis H_0 is accepted only if the set of null hypotheses $H_{0_1}, \ldots, H_{0_m}$ are each accepted. Each of the

m hypotheses are tested by comparing the observed value of each statistic to the simulated data correcting for multiple comparisons as describe by Hommel (1983). This procedure proved to be successful in controlling the type I error as shown in figure 6.14. Chapter 7 draws attention to some problematic areas in testing real SNP data sets, in particular, expanding the test to allow analysis of large data sets, where the SNPs' ascertainment needs to be modelled and these methods are illustrated using data from the International HapMap 3 Consortium (2010). The hypothesis test assumes that if data were ascertained then the ascertainment scheme is known and data can be simulated under this scheme. Chapter 8 illustrated the test on the two Kenyan populations and the Yoruba and Japanese populations from the HapMap project. HapMap data was ascertained through several complex schemes which are difficult to model in detail. Therefore, simulated SNPs were ascertained through a more simplified method, using an ascertainment sample of size 2 from each subpopulation in the sample. The null hypothesis was accepted in the case of the two Kenyan populations providing evidence that these two populations diverged from a common ancestor. On the other hand, the null hypothesis was rejected when comparing populations YRI and JPT.

9.2 Discussion

Chapter 6 tests the consistency of SNP data from two subpopulations with a model of isolation taking a frequentist approach to hypothesis testing. When interest lies in assessing the fit of data to two opposing models, or indeed more than two models, there are some pertinent issues in this analyses. The remaining sections evaluate the limitations of the analysis and present some potential areas for improvement and further work.

9.2.1 Limitations

Cox and Hinkley (1974) detail the limitations of significance tests. The foundation of the test is to compare observed data to data that might have arisen under a null model and

so examining how consistent the observed data are to the null model. Given a statistic T, the p-value is a measure of the consistency and Cox and Hinkley define it to be $Pr\{T > t_{obs}|H_0\}$, with t_{obs} the observed value of the statistic. A large p-value only indicates that the test statistic is unable to distinguish the 'true' model, in so far as this concept is meaningful, and the null model. It provides no evidence to reject the null model. However, there is no indication that the 'true' model and the null model are equivalent or that the null model is the best–fitting model. For example, the authors suggest that it is possible to obtain a large p-value when the observed data are inconsistent with the null hypothesis. Moreover, the p-value is not the probability that the data are from a specific model. Other model selection procedures, such as the one provided by the consideration of Bayes factors, can provide evidence in favour of a particular model over another. For example, the value of the Bayes factor, B_{01}, provides evidence against H_1 in favour of H_0 if $B_{01} > 1$. Therefore, the tests described in chapters 6 and 7 quantify the consistency of SNP data with an isolation model.

The set of summary statistics selected to perform the test can present a difficulty. In estimating model parameters, employing sufficient statistics provides a natural way of reducing the dimension of the data since they contain as much information about the parameter as the full data set. Chapter 5 discussed developing methods of finding an optimal set of statistics and ways of assessing the performance of the set in the context of estimating model parameters, using ABC where it may not be possible to find sufficient statistics. In this particular test, the set of statistics employed were $\pi, \pi_W, \pi_B, \eta_1, \eta_{max}$, D_t, the mean allele frequency and the variance in allele frequencies. These statistics were chosen since individually they appeared to be able to distinguish the models from data simulated under similar conditions, matching F_{st}, from both models. Figure 6.3 illustrates the distributions of the statistics from simulated data with a small migration rate between subpopulations and figure 6.4 shows the corresponding results with a larger migration rate. The test should be most successful when there is little migration between subpopulations, or equivalently a more ancient population divergence time.

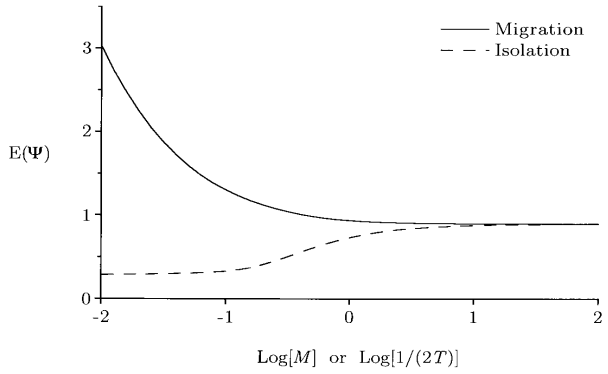


Figure 9.1: Figure 4 from Wakeley (1996).

Generally, the allele frequency spectrum is positively skewed since it is more common to find SNPs with a small (minor) allele frequency. Therefore instead of utilizing the mean and variance, one might consider the median and the interquartile range. In addition, the statistic Ψ introduced by Wakeley (1996), a function of the variance of the pairwise difference, is promising for distinguishing the two models. Figure 9.1 shows a reproduction of figure 4 from Wakeley (1996). Wakeley analyses a model with two subpopulations which either exchange migrants at rate M or diverged at time T in the past measuring time in N generations (that is, the population size of each subpopulation). This statistic shows differences from lower migration rates whereas as the migration rate increases, the two models show similar values of Ψ. He shows that that low values of Ψ provides evidence in favour of an isolation model.

9.2.2 Improvements

This section touches upon some improvements that may enhance the performance of the hypothesis test introduced in chapter 6. Although the test was shown to be powerful in distinguishing the two models of interest whilst controlling the type I error, there are a few areas were the test might be improved.

Section 3.2 provided the strategy for simulating SNP data in this thesis. Throughout, only biallelic SNP data have been considered and so the genealogical history of a sample

is simulated according to the method given by Hudson (1991). A single mutation is randomly added to a branch instead of adding a Poisson number of mutations with the mean depending on the scaled mutation rate and the length of the genealogy. This method of simulating data with a fixed number of mutations has cause for concern, as outlined by Wall and Hudson (2001) and Markovtsova et al. (2001). In particular, Wall and Hudson (2001) tested the consistency of the two methods of data simulation by comparing the value of a test statistic from data simulated from Hudson's method and data simulated using the fixed number of mutations method. Incorporating Hudson's method of simulation in the hypothesis test introduced in chapter 6 would involve firstly simulating a genealogy and then adding a Poisson number of mutations. Each simulated data set would include only the genealogies that contained exactly one mutation.

F_{st}–based estimators of population parameters m and τ are both quick and straightforward but can often be inaccurate. Section 5.1.2 discussed problems with F_{st}–based estimators and section 5.3.1 used an MCMC algorithm to estimate τ using F_{st}. The MCMC algorithm was shown to be more accurate in estimating τ (figure 5.9), but at a large computational cost, since each iteration involves simulating data under the proposed model and estimating F_{st}. For instance, given observed data from two subpopulations each of haploid sample size 10 and 1000 SNPs, (5.1) estimates τ practically instantaneously. On a 3.33GHz Intel Xeon processor, simulating data of this size from the isolation model with $\tau = 0.1$ takes on average 5.6 seconds. Therefore, 1000 simulations may take, on average, about 90 minutes. This thesis did not focus on efficient simulation of large SNP data sets. However, more efficient simulation will reduce the cost of the ABC_MCMC algorithm, for example, by replacing the simulation step by a compiled function in a language such as C or Fortran.

In cases where the original estimator fails, for example with ascertained data, the flexibility in ABC methods allows ascertainment to be included in the estimation of τ given the ascertainment scheme. Incorporating other summary statistics in the parameter estimation stage, instead of using only F_{st}, has the potential to improve the quality of this test. Then, the hypothesis test would require the specification of two sets of statistics, namely

a set to be used in parameter estimation and one to be used as test statistics.

As previously mentioned, this test relies on an adequate selection of summary statistics. Section 5.3.2 details some methods of selecting summary statistics for parameter estimation. The most intuitive procedure is to select statistics that show differences in the two models. Joyce and Marjoram (2008) show that using too many statistics adds noise to parameter estimation and so discuss identifying a minimal set of statistics without comprising the loss of information. In a hypothesis test, the aim is to find powerful test statistics in distinguishing the null and alternative models. Let $\mathcal{S} = \{\pi, \pi_W, \pi_B, \eta_1, \eta_{mean}, D_t, \text{mean}, \text{variance}\}$, then it may be of interest to find the minimal set $U \subseteq \mathcal{S}$, which preserves the power of the test whilst controlling the type I error. Assuming the parameters in the null model are adequately estimated, the hypothesis test acquires power from statistics that can reject the null model and, assuming there are statistics included in the test that can do this, the effects of including statistics that are not able to differentiate the two models will neither detract from nor contribute to the test.

Chapter 7 aimed to find a test that was able to cope with ascertained data. Two different directions for addressing this problem were either to correct for ascertainment or to incorporate ascertainment into the data simulation. Nielsen and Signorovitch (2003) describe a method of correcting the allele frequency spectrum given ascertained data and section 7.1.2.2 followed analogous steps to correct the within–subpopulation allele frequency spectrum. Although this method roughly reconstructs the true allele frequency spectrum, using the maximum likelihood estimates to calculate the set of summary statistics was unsuccessful. Adapting this procedure by finding a better estimate of the true allele frequencies given the ascertainment scheme within a subpopulation would help to more adequately estimate the summary statistics and so improve the performance of the hypothesis test in the presence of SNP ascertainment.

Section 7.2.4 briefly discussed incorporating larger data sets containing L SNPs with, for example, $L > 1,000,000$. The suggestion was to simulate data with a smaller number of SNPs, L_1, and scale each statistic that should be proportional, on average, to the number

of SNPs by $\frac{L_1}{L}$. This method may have been more appropriate in testing the HapMap data rather than sub–sampling the SNPs. The question of the successfulness of this proposition remains open.

Chapter 8 tests data from the HapMap project. SNPs were simulated with ascertainment, but, the ascertainment scheme used did not exactly replicate that used to ascertain SNPs in the HapMap project. Inaccurately simulating data to compare to the observed data may produced invalid results. However, the extend of the effects of these inaccuracies has not been considered in this thesis. In addition, it is clear that the size of the ascertainment panel effects the shape of the allele frequency spectrum, therefore changing the ascertainment sample size may also effect the results from this test.

9.2.3 Extensions

Chapter 5 looked at model selection via Bayes factors, and discussed the results of Robert et al. (2011) who identified problems with using non–sufficient statistics in computing the likelihood function. Bayes factors provide evidence in favour of one model compared to the other, whereas a conventional hypothesis test compares observed data to a null model and provides evidence in favour, or against, the null model but the p-value does not provide any degree of evidence to support the null model compared to the alternate model. In the context of this research, it is equally compelling to test whether the data are consistent with the migration model (as null), rather than the isolation model. Following a similar procedure as the one in chapter 5 to estimate τ, the scaled migration rate $M = N^T m$ can be estimated using

$$\hat{M} = \frac{1 - \hat{F}_{st}}{4\hat{F}_{st}},$$

with $\hat{F}_{st}$ estimated using equation (5.1). Also, M may be estimated using the ABC_MCMC algorithm. Test IV can be used by altering the bootstrapping steps to simulate from a migration model given $\hat{M}$, although the values of ϵ and σ^2 may need to be reassessed.

In addition, this test could easily be extended to incorporate models with more than two subpopulations, for example section 4.5 presents a model with four subpopulations. The migration model in this example includes three migrations rates m_1, m_2 and m_3, assuming that $m_{ij} = m_{ji}$ for $i, j = 1, 2, 3$ or 4. In the general case of P subpopulations one needs $\binom{P}{2}$ migration parameters. More complex demographic models can be contemplated at the expense of introducing more parameters, which would need to be estimated. For example, genetic variation can be affected by other demographic characteristics, such as population growth or a bottleneck. Figure 9.2, a depiction of the complex model described by Stoneking and Krause (2011), shows an ancestral population diverging into two subpopulations. After this time, both subpopulations expand with subpopulation 2 undergoing a bottleneck event. Migrants are also exchanged.

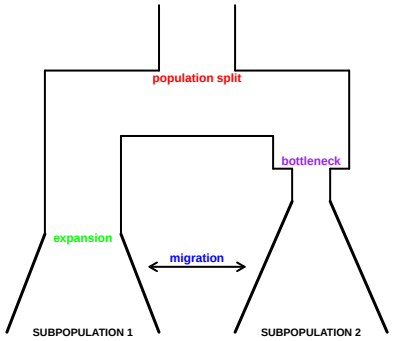


Figure 9.2: Complex demographic history scenario.

To test this model, compared to some alternative model, it is necessary to be able firstly to estimate the parameters and then to find and simulate a set of test statistics. The more complex the model becomes then the more parameters there are to estimate.

In conclusion, problematic areas in distinguishing demographic models using SNP data were highlighted. This thesis developed a method to differentiate between a model of isolation and one with migration via a frequentist hypothesis test. More generally, it

provides a framework for approaching model testing of demographic models for large genomic data sets, incorporating the effects of locus ascertainment.

Bibliography

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

Adams, J. (2008). Complex genomes: shotgun sequencing. *Nature Education 1*.

Albrechtsen, A., F. C. Nielsen, and R. Nielsen (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution 27*, 2534–2547.

Anderson, C. N. K., U. Ramakrishnan, Y. L. Chan, and E. A. Hadly (2005). Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics 21*, 1733–1734.

Astle, W. and D. J. Balding (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science 24*, 451–471.

Barbujani, G. and G. Bertorelle (2001). Genetics and the population history of Europe. *Proceedings of the National Academy of Sciences 98*, 22–25.

Barnes, C., S. Filippi, M. P. H. Stumpf, and T. Thorne (2011). Considerate approaches to achieving sufficiency for ABC model selection. *arXiv:2011arXiv1106.6281B*.

Bazin, E., K. J. Dawson, and M. A. Beaumont (2010). Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics 185*, 587–602.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian computation in population genetics. *Genetics 162*, 2025–2035.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology 13*, 827–836.

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 of the United States of America 98*, 4563–4568.

Boulesteix, A.-L. and K. Strimmer (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics 8*, 32–44.

Bretz, F., T. Hothorn, and P. Westfall (2011). *Multiple Comparisons Using R*. Chapman and Hall/CRC press, Boca Raton, FL.

Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonné-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza (2002). A human genome diversity cell line panel. *Science 296*, 261–262.

Cann, R. L., M. Stoneking, and A. C. Wilson (1987). Mitochondrial DNA and human evolution. *Nature 325*, 31–36.

Cavalli-Sforza, L., P. Menozzi, and A. Piazza (1993). Demic expansions and human evolution. *Science 259*, 639–646.

Cavalli-Sforza, L. L. (1969). Human diversity. *Genetics 2*, 405–416.

Cavalli-Sforza, L. L. and W. F. Bodmer (1971). *The Genetics of Human Populations*. W. H. Freeman and Company, San Francisco, CA.

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

Cavalli-Sforza, L., M. P. and A. Piazza (1993). *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Chadeau-Hyam, M., C. Hoggart, P. O'Reilly, J. Whittaker, M. De Iorio, and D. Balding (2008). Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics 9*, 364.

Chang, M. (2011). Multiple-hypothesis testing strategy. In *Modern Issues and Methods in Biostatistics*, Statistics for Biology and Health, pp. 1–30. Springer, New York.

Congdon, P. (2003). *Applied Bayesian Modelling*. Wiley, Chichester.

Cornuet, J.-M., F. Santos, M. A. Beaumont, C. P. Robert, J.-M. Marin, D. J. Balding, T. Guillemaud, and A. Estoup (2008). Inferring population history with DIY ABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics 24*, 2713–2719.

Cox, D. (2006). *Principles of Statistical Inference*. Cambridge University Press, London.

Cox, D. and D. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall, London.

Csilléry, K., O. François, and M. G. B. Blum (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution, in press*.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Depaulis, F. and M. Veuille (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution 15*, 1788–1790.

Donnelly, P. and S. Tavaré (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics 29*, 401–421.

Dunteman, G. H. (1989). *Principal Components Analysis.* Sage Publications, Newbury Park.

ENCODE Project Consortium (2007). Identification and analusis of functional elements in 1 percent of the human genome by the ENCODE pilot project. *Nature 447*, 799–815.

Excoffier, L. and G. Heckel (2006). Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics 7*, 745–758.

Excoffier, L., J. Novembre, and S. Schneider (2000). SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity 91*, 506–509.

Fearnhead, P. and D. Prangle (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesain computation. *Journal of the Royal Statistical Society B, in press*.

Feller, W. (1950). *An Introduction to Probability Theory and its Applications*, Volume 1. Wiley, New York.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection.* Clarandon, Oxford.

Gelman, A., J. B. Carlin, H. S. Stern, and R. B. Donald (2004). *Bayesian Data Analysis.* Chapman and Hall/CRC, Boca Raton, FL.

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

He, M., J. Gitschier, T. Zerjal, P. de Knijff, C. Tyler-Smith, and Y. Xue (2009). Geographical affinities of the HapMap samples. *PLoS ONE 4*, e4684.

Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution 27*, 905–920.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox (2005). Whole-genome patterns of common DNA variation in three human populations. *Science 307*, 1072–1079.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal 25*, 423–430.

Hommel, G., F. Bretz, and W. Maurer (2011). Multiple hypothesis testing based on ordered p values: a historical survey with applications to medical research. *Journal of Biopharmaceutical Statistics 21*, 595–609.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology 7*, 1–44.

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.

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature 467*, 52–58.

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

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature 473*, 1299–1320.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 millionSNPs. *Nature 449*, 851–862.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics 29*, 295–327.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, New York.

Joyce, P. (1998). Partition structures and sufficient statistics. *Journal of Applied Probability 35*, 622–632.

Joyce, P. and P. Marjoram (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology 7*.

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

Kimura, M. and J. F. Crow (1964). The number of alleles that can be maintained in a finite population. *Genetics 49*, 725–738.

Kimura, M. and T. Ohta (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences of the United States of America 75*, 2868–2872.

Lange, K. (1999). *Numerical Analysis for Statisticians*. Springer, New York.

Laval, G. and L. Excoffier (2004). SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics 20*, 2485–2487.

Li, Y., N. Vinckenbosch, G. Tain, E. Huerta-Sanchez, T. Jiang, H. Jiang, and A. Albrechtsen (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genetics 42*, 969–972.

Lopes, J. S., D. Balding, and M. A. Beaumont (2009). PopABC: a program to infer historical demographic parameters. *Bioinformatics 25*, 2747–2749.

Markovtsova, L., P. Marjoram, and S. Tavaré (2001). On a test of Depaulis and Veuille. *Molecular Biology and Evolution 18*, 1132–1133.

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

Nachman, M. W. and S. L. Crowell (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics 156*, 297–304.

Nei, M. (1995). Genetic support for the out-of-Africa theory of human evolution. *Proceedings of the National Academy of Sciences of the United States of America 92*, 6720–6722.

Nielsen, R. (2004). Population genetic analysis of ascertained SNP data. *Human Genomics 1*, 218–224.

Nielsen, R., M. J. Hubisz, and A. G. Clark (2004). Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics 168*, 2373–2382.

Nielsen, R., J. L. Mountain, J. P. Huelsenbeck, and M. Slatkin (1998). Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution 52*, 669–677.

Nielsen, R. and J. Signorovitch (2003). Correcting for ascertainment biases when analysing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology 63*, 245–255.

Nielsen, R. and J. Wakeley (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics 158*, 885–896.

Nordborg, M. (2007). Coalescent theory. In Handbook of Statistical Genetics, pp. 843–877. John Wiley and Sons, Chichester.

Novembre, J. (2008). Genes mirror geography within Europe. *Nature 456*, 98–101.

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

Nunes, M. A. and D. J. Balding (2010). On optimal selection of summary statistics for Approximate Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology 9, 34*.

Park, L. (2011). Effective population size of current human populations. *Genetics Research 93*, 105–114.

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

Pritchard, J. and M. Przeworski (2001). Linkage disequilibrium in humans: Models and data. *The American Journal of Human Genetics 69*, 1–14.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution 16*, 1791–1798.

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

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ramírez-Soriano, A. and R. Nielsen (2009). Correcting estimators of θ and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics 181*, 701–710.

Relethford, J. (2008). Genetic evidence and the modern human origins debate. *Heredity*, 555–563.

Reynolds, J., B. S. Weir, and C. C. Cockerham (1983). Estimation of the coancertry coefficient: basis for a short-term genetic distance. *Genetics 105*, 767–779.

Robert, C., J. Cornuet, J. Marin, and N. Pillai (2011, February). Lack of confidence in ABC model choice. *ArXiv:1102.4432v4*.

Robert, C., J.-M. Marin, and N. S. Pillai (2011, January). Why approximate Bayesian computational (ABC) methods cannot handle model choice problems. *ArXiv:1101.5091v2*.

Rosenberg, N. (2004). Distruct: a program for the graphical display of population structure. *Molecular Biology Notes 4*, 137–138.

Ross, S. M. (1997). *Simulation*. Academic Press, San Diego, CA.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution 4*, 406–425.

Sherry, S. T., M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research 29*, 308–311.

Shin, J.-H., S. Blay, and B. McNeney (2006). LDheatmap:an r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software 16*.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 413–429.

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

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution 47*, 264–279.

Stephens, M. (2007). Inference under the coalescent. In Handbook of Statistical Genetics, pp. 878–908. John Wiley and Sons, Chichester.

Stoneking, M. and J. Krause (2011). Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics 12*, 603–614.

Studier, J. A. and K. J. Keppler (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution 5*, 729–731.

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

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics 123*, 585–595.

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

Varadhan, R. (2011). *alabama: Constrained nonlinear optimization*. R package version 2011.9-1.

Wakeley, J. (1996). Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical Population Biology 49*, 369–386.

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Roberts and Company Publishers, Greenwood Village, CO.

Wakeley, J. and J. Hey (1997). Estimating ancestral population parameters. *Genetics 145*, 847–855.

Wall, J. D. and R. R. Hudson (2001). Coalescent simulations and statistical tests of neutrality. *Molecular Biology and Evolution 18*, 1134–1135.

Wang, D. G., J.-B. Fan, C.-J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science 280*, 1077–1082.

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

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. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data* (2nd ed.). Sinauer Associates, Sunderland, MA.

Whitlock, M. and D. McCauley (1998). Indirect measures of gene flow and migration: $f_{st} \neq 1/(4nm + 1)$. *Heredity 92*, 117–125.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen, and C. D. Bustamante (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences of the United States of America 102*, 7882–7887.

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

Wright, S. (1969). *Evolution and the Genetics of Populations, Vol. 2: The Theory of Gene Frequencies*. University of Chicago Press, Chicago IL.