Interpretable Models of Genetic Drift Applied Especially to Human Populations

Alasdair McIntosh

This thesis is submitted in
fulfilment of the requirements
of the Degree of
Doctor of Philosophy

School of Mathematics & Statistics
College of Science and Engineering
University of Glasgow

September 2017

Abstract

This thesis aims to develop and implement population genetic models that are directly interpretable in terms of events such as population fission and admixture. Two competing methods of approximating the Wright–Fisher model of genetic drift are critically examined, one due to Balding and Nichols and another to Nicholson and colleagues. The model of population structure consisting of all present-day subpopulations arising from a common ancestral population at a single fission event (first described by Nicholson et al.) is reimplemented and applied to single-nucleotide polymorphism data from the HapMap project. This Bayesian hierarchical model is then elaborated to allow general phylogenetic representations of the genetic heritage of present-day subpopulations and the performance of this model is assessed on simulated and HapMap data. The drift model of Balding and Nichols is found to be problematic for use in this context as the need for allele fixation to be modelled becomes apparent. The model is then further developed to allow the inclusion of admixture events. This new model is, again, demonstrated using HapMap data and its performance compared to that of the TreeMix model of Pickrell and Pritchard, which is also critically evaluated.

Acknowledgements

First, I would like to thank my family for their patience and understanding throughout my studies. Thank you to the R foundation, Lyx.com and CodeLite.org for providing their software for free and also to the International HapMap Project for their data. I am also very grateful to the EPSRC for providing funding for this project. As well as the University of Glasgow, I also need to thank the University of the West of Scotland and Glasgow City Council for providing study facilities in the later stages of this project. A very big thank you to my outstanding supervisor, Vincent Macaulay, for his support, encouragement and advice throughout this project. Finally, I would like to thank those who have, over the years, derided the idea that I could gather the skills to even aspire to attempt a project such as this, that I should accept things as they are and not have ideas above my station. The thought of proving you all wrong was just the motivation I needed to get me through the difficult times on this project.

Declaration

I have prepared this thesis myself; no section of it has been submitted previously as part of any application for a degree. I carried out the work reported in it, except where otherwise stated.

Contents

List of Tables

List of Figures

Chapter 1

Background

"Why do these people look so different from us?" This question, or some variant of it, is one that almost all parents have had posed to them by their small children at some point and can become a test of parental tact and diplomacy if the question is posed too loudly in a public space. Nevertheless, the questions of how people whose ancestry is from different parts of the world came to be there, why people from particular parts of the world appear more similar to each other than they do to people from other parts, and what that can tell us about humans in the past forms the scientific field of biological anthropology.

1.1 Deoxyribonucleic Acid

One place to look for clues to the answer to these questions is from people's DNA (Deoxyribonucleic Acid), a molecule that resides in every living cell and determines much about the growth, development and even susceptibility to and ability to recover from disease of the person to whom it belongs. DNA is a molecule that is typically formed from two biopolymer strands in the shape of a double helix (figure 1.1).

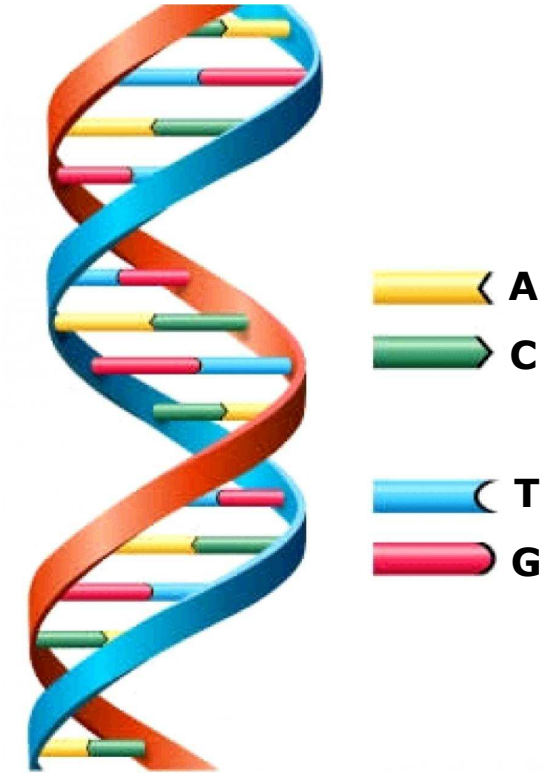

Figure 1.1: Double Helix of DNA
A schematic representation of a short section of a strand of DNA showing the double helix and how the Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) nucleotides are contained in pairs along its length. Image reproduced from clipart-library.com.

Each strand contains a sequence of nucleotides along its length. These nucleotides naturally occur in four varieties, Adenine (A), Cytosine (C), Thymine (T) or Guanine (G). The strands in the pair contain nucleotides that correspond with (or complimentary to) each other such that where one has A, the other has T and where one has C, the other has G. These form a so-called base pair. In humans and other eukaryotic organisms (animals, plants, fungae and a few others such as algae) their DNA is organised into a number of chromosomes. Humans are a diploid species which means that they have two sets of chromosomes, one set from each parent. Haploid species only have a single set. Humans ordinarily have 23 pairs of such chromosomes, 22 autosomes and a pair of sex chromosomes, an X and a Y for males and a pair of Xs for females.

The human genome contains about 3,000 million base pairs in total (Human Genome Project, 2003). In comparison, the *Escherichia coli* bacterium's genome contains about 5 million base pairs (Blattner et al., 1997). The fruit fly, *Drosophila melanogaster*, that features in so much biological research, as a model organism, has about 175 million base pairs in its genome (Ellis et al., 2014). It might be tempting to think that the number of base pairs in an organism's genome is related to its complexity or size. This is not the case. Onion genomes have about 16,000 million base pairs, over 5 times that of humans (Palazzo and Gregory, 2014), whilst the genome of the freshwater amoeba, *Polychaos dubium* was reported as having a massive 670,000 million base pairs in its genome (Friz, 1968). This figure is, however, subject to confirmation using more recent techniques.

1.1.1 Single Nucleotide Polymorphisms (SNPs)

Of these 3,000 million base pairs in the human genome, it is estimated that all but about 10 million are the same for almost all humans so that, for example, if one human has adenine at a particular locus in the genome, almost all humans will have adenine at that locus. The remaining 10 million can contain different

nucleotides in different people. These are called Single Nucleotide Polymorphisms (SNPs). SNPs are loci (i.e. specific positions on the genome) where more than one nucleotide variant has been identified among humans and at least two of these variants have a frequency above a very small minimum threshold. They occur at 1 in 300 base pairs on average (Making SNPs Make Sense, 2017). SNPs can occur in coding and non-coding regions of the genome. A coding region is one where the DNA sequence can be used to produce a protein. Each group of three nucleotides in a coding region correspond to a particular amino acid in the protein's chain (Hartl and Clark, 1997). A SNP in such a region can lead to a different protein being produced. While the vast majority of the genome is non-coding, that does not mean it does not always have a useful role. For example, some non-coding regions help facilitate the transcription of nearby coding DNA and non-coding DNA near the end of chromosomes (telomeres) help to provide a buffer zone that protects coding DNA from damage and degradation (Mandal, 2014). SNPs are less common in coding regions because the changes in protein structures to which a changed nucleotide can lead can have a negative impact on the chances of the resulting human surviving or reproducing to pass the SNP variant on to a future generation. On the other hand, those changes in non-coding regions are less likely to have reproductive implications and are thus more likely to survive into the next generation (Barreiro et al., 2008).

1.1.2 Linkage and Independence

Focussing on one strand of the double helix, since the other is determined by it, while the variant that appears at a SNP on one chromosome is independent of one that appears at a SNP on another chromosome, it is not entirely independent of that which appears at another SNP on the same chromosome. The reason for this is called linkage. As has been mentioned, humans have 23 pairs of chromosomes in each cell. However, in the cells involved in reproduction, the gametes, ova in

females and sperm in males, there are only one set of 23 chromosomes. These are produced by a process called meiosis. During that process each pair of chromosomes is separated but during separation they are sometimes cut at corresponding positions and recombined (Figure 1.2).

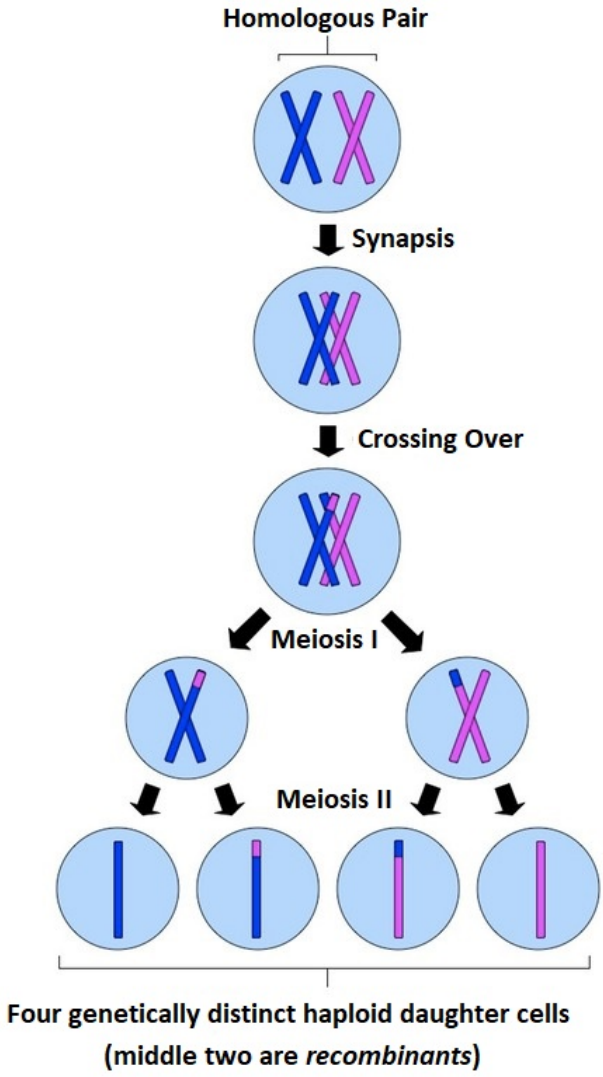

Figure 1.2: Meiosis
The two versions of the same chromosome that an individual has is called a homologous pair. First both versions of the chromosome are duplicated. At the Synapsis stage these sets of duplicates pair up. The pairs are held together at points on their length called the chiasma. Often the genetic material from one version of the chromosome swaps over to the other version and vice versa at these points. This is called crossing-over. The cell then divides once and the chromosome pairs seperate at the chiasma leaving one modified duplicate pair of a chromosome in each cell. The cells then divide again so that there are four cells which have only one version of each chromosome in each of the four cells. Chromosomes that have been modified by crossing over are called recombinants. This image was reproduced from BioNinja. (BioNinja, 2017)

These produce a new pair of chromosomes that contain the genetic information from the beginning of one of the original pair of chromosomes and from the end of the other. This is called recombination. More rarely, such cuts can happen two or more times. Each of the new pair becomes part of a different gamete. While a variant at a SNP could find itself in the gamete with a variant that appears at another SNP on either of the two copies of another chromosome, it will be more likely to appear with the variant that appears at another SNP on the same copy of the chromosome that it is on. It will be even more likely to appear alongside it the nearer it is to the first SNP on the same chromosome. The closer together they are, the less likely it is that a recombination event occurs between them during the recombination process. So these variants are more likely to appear together in the next and subsequent generations. Loci on a chromosome that are close enough to each other that the proportions of each variant (allele) they have at each SNP are not independent at the population level, are said to be in linkage disequilibrium.

1.1.3 Mutation

But where do SNPs come from? How do they arise? It is estimated that every time human DNA is passed from one generation to the next it results in about 60 new mutations (Conrad et al., 2011). Mutations can occur naturally as copying errors when DNA is duplicated for cell division and can take several forms such as a sequence of nucleotides being repeated, nucleotides being inserted or deleted or a single nucleotide changing into another (a substitution). This would suggest that an average base pair had a chance of about 1 in 50 million of having a novel mutation occur in a single generation. Lipson et al. (2015) report a rate of about 1.6 mutations per 100 million bases per generation which is a slightly lower rate.

1.2 Shared Ancestry

Any person alive has two biological parents (apart from a very small number recently born using a new technique where mitochondrial DNA comes from a third parent (Hamzelou, 2016)). These in turn will have two parents each and so on back through the generations, there being more and more people that the present day person is descended from as each generation in a diagram such as figure 1.3 is added.

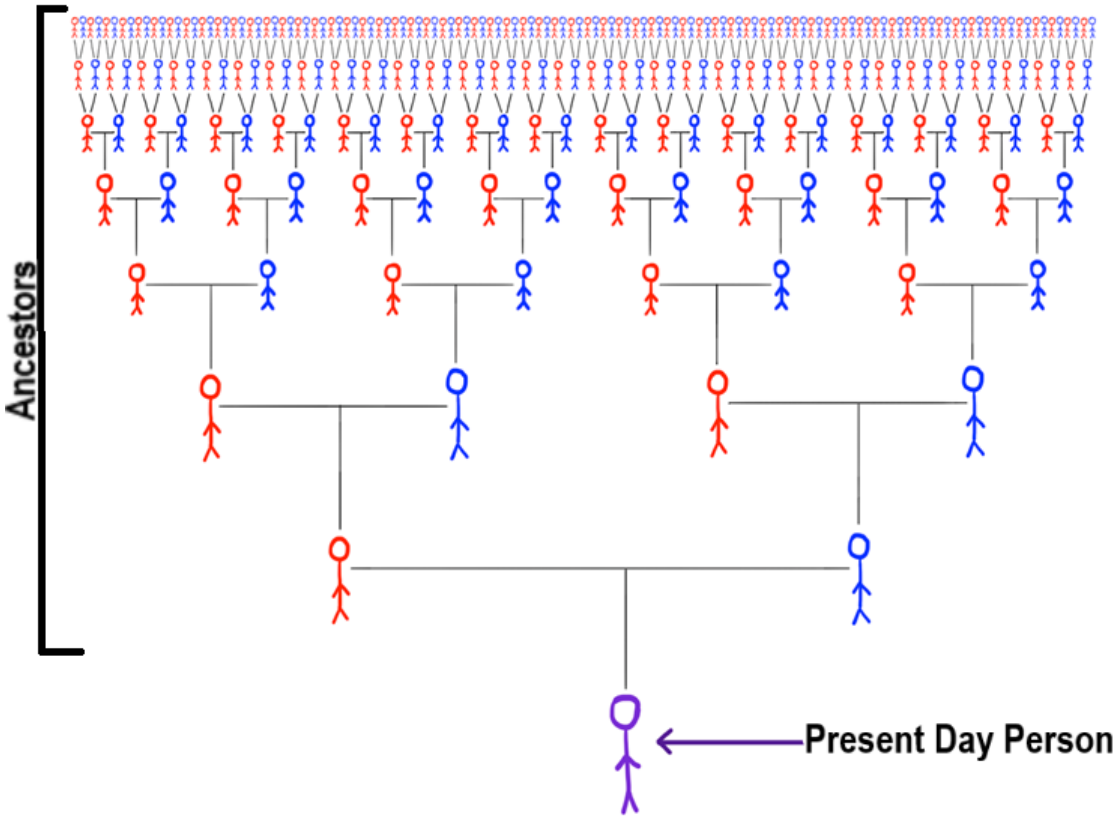

Figure 1.3: Ancestors
The present-day person has two biological parents who each have two biological parents who each have two biological parents and so on. Although the diagram shows the number of ancestors doubling with each generation, this is not generally the case. People who are knowingly or unknowingly recently related can pair up to produce children reducing the number of ancestors in earlier generations. For example, if the present-day person's parents were cousins, they would only have 6 great grandparents rather than 8. This figure is based on an image in waitbutwhy.com. (Urban, 2014).

For any two people presently alive, such a pair of diagrams can be constructed

adding generations until eventually, there will be a person or persons who will appear on both diagrams. This is their most recent common ancestor. However, each part of the genome has been inherited along different paths through their family tree, and so for any two individuals, different parts of their genome will have different most recent common ancestors. Going further back, there will be further individuals common to both trees who represent other common ancestors. As has been noted above, the Y chromosome in humans occurs only in males and can only be inherited patrilinearly. It rarely undergoes recombination with the X chromosome and then only at its extreme ends; the rest of it can only change through mutation. These mutations are then passed to subsequent generations. These can be used to establish the way that present day men are related patrilinearly. If the assumption is made that the same mutation is so unlikely to occur more than once at the same locus in the chromosome that the possibility can be discounted, people with the same Y chromosome sequence can be assumed to have a common patrilineal ancestor in whose development the mutation originally occurred. They in turn can be assumed to have a common ancestor with those who carry a Y chromosome that is the same except for that mutation. By following this process repeatedly, a hierarchy of common ancestors can be built up until a shared common ancestor Y chromosome is arrived at for all living males. The person who is the most recent common patrilineal ancestor of all living human males in this way is termed Y chromosome Adam or Y-MRCA. Poznik et al. (2013) estimates that this individual would have lived between 156,000 and 120,000 years ago.

Mitochondrial DNA is a small amount of DNA that exists outside a cell nucleus and is not part of the 23 pairs of chromosomes that reside within a cell nucleus. There is mitochondrial DNA in the female reproduction cell (ovum) when it fuses with the male reproductive cell (sperm) to form a zygote. The mitochondrial DNA in the sperm cell is almost never passed on (and is perhaps actively destroyed) leaving only that from the ovum in the resulting embryo cells. In fact, Pyle et al. (2015) could find no evidence of male mitochondrial DNA being passed

on. Mitochondrial DNA is therefore inherited matrilinearly. A similar process of analysing the mutations in mitochondrial DNA that exist in modern-day humans can be followed as that described above for the Y chromosome to build up a hierarchy of matrilineal most recent common ancestors. The most recent matrilineal common ancestor of all living people is termed mitochondrial Eve or mt-MRCA. Poznik et al. (2013) report that they would have lived between 148,000 and 99,000 years ago.

1.3 Natural Selection and Genetic Drift

Despite all humanity having shared ancestry and having the overwhelming proportion of the human genome in common, people who have ancestry in particular parts of the world clearly share physical characteristics that are not shared with people with ancestry in other parts of the world. Natural selection in relation to local environment in prehistoric times could explain some of these characteristics (Smithsonian, 2017). For example, in a sunny part of the world, lighter skin pigmentation is a disadvantage. It sunburns easily and has a higher cancer risk but also too much UV light leads to a degradation in folate and folic acid. Folate has a role in preventing some birth defects (Borradale and Kimlin, 2012). In a part of the world where dull and overcast weather is more common, the lighter skin pigmentation is an advantage because it can make better use of the limited solar ultra-violet radiation to produce vitamin D. In the modern world, however, with more varied diet and better access to sun-block and vitamin supplements these particular differences provide no biological advantage either way. Nevertheless, not all differences can be explained by historical local advantages. To understand how such differences could arise by random chance, the concept of genetic drift needs to be introduced.

Imagine a population consisting of n individuals. They have two versions of each chromosome each and so the population has $2n$ versions of each chromosome in

total. Suppose there is a locus that is a SNP and there are two variants at that SNP, C and T. Further suppose that there are $2n\alpha$ $\left(\alpha \in \left\{0, \frac{1}{2n}, \frac{2}{2n}, ..., 1\right\}\right)$ versions of the chromosome with a C at the locus, and hence $2n(1-\alpha)$ with a T in this population. Putting the question of which sex each of the individuals are to one side, reproduction in relation to this locus will be assumed to happen randomly ("random mating"). This assumes that the locus does not affect the probability of reproduction and is not in linkage disequilibrium with another SNP that does affect the probability of reproduction. Each chromosome in the next generation will have a probability of α of having a C at the locus and a probability of $1-\alpha$ of having a T. If the overall population size remains n, then the number of chromosomes with C at the locus is random and can be modelled as Binomial$(2n, \alpha)$. This has expected value $2n\alpha$, the same as the number observed in the first generation and variance $2n\alpha(1-\alpha)$. In this population of constant size, the number of chromosomes with C can rise or fall over time from one generation to the next (figure 1.4) until either the C or T variants (alleles) completely disappears. When this happens the remaining allele is said to be fixed. That is because, unless there is a mutation, the next generation and all subsequent generations will only have that allele because there is no individual in the previous generation from whom to inherit the other allele. This random process where the proportion of chromosomes with a C at the locus can rise or fall over time, not driven by any force like natural selection, is called genetic drift. The model just described is the Wright-Fisher model from Wright (1931) and Fisher (1930).

Next, imagine a population with $n_1 + n_2$ individuals in it and $2(n_1 + n_2)\alpha$ versions of the chromosome with a C at the locus. If the population stayed intact the next generation would have a number of chromosomes with a C at the locus modelled by Binomial$(2[n_1 + n_2], \alpha)$. However, suppose it splits into two groups, one with n_1 individuals and one with n_2 individuals. The second group goes off to live in isolation, perhaps on an island, and they never meet again. The one with n_1 individuals has $2n_1\beta$ versions of the chromosome with a C at the locus and the

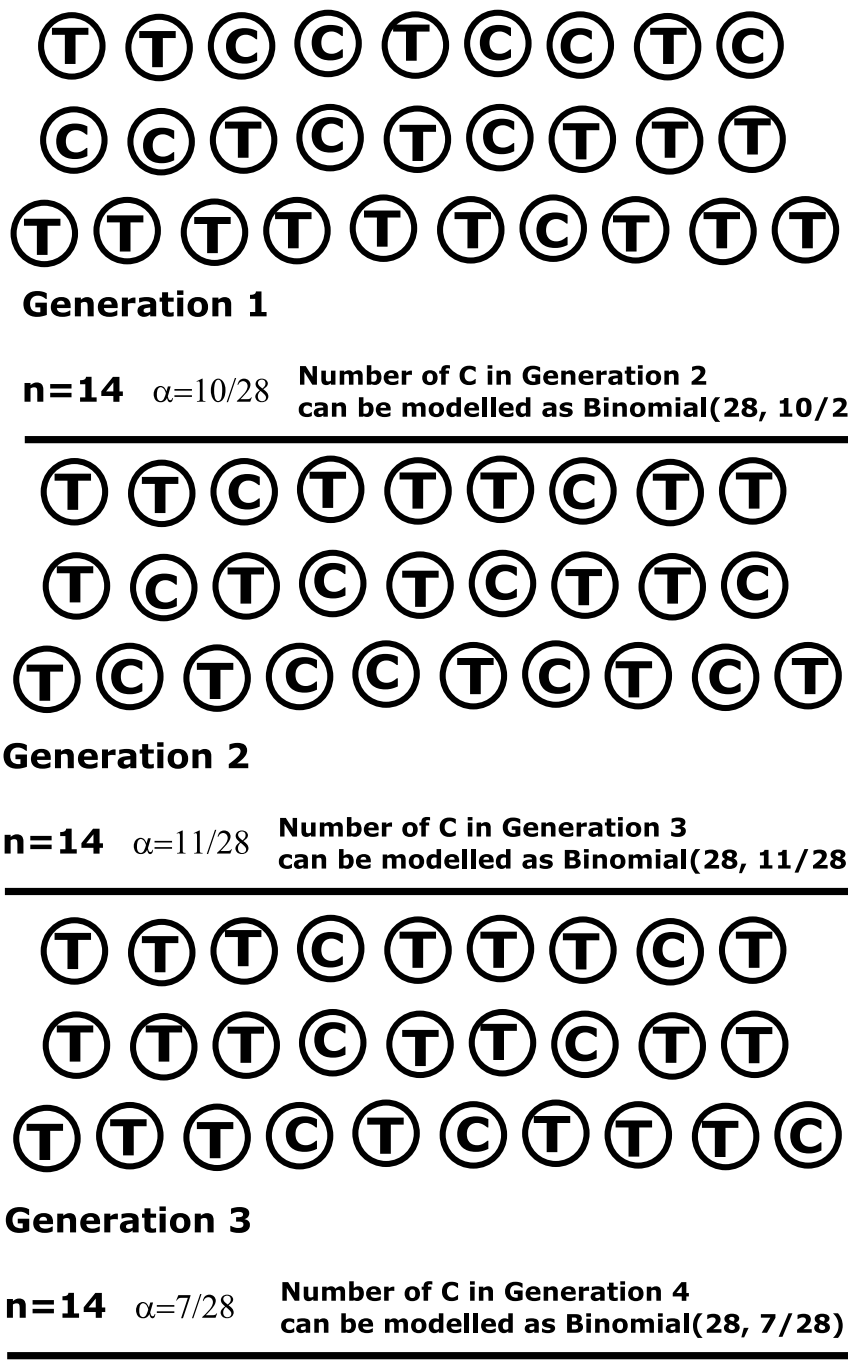

Figure 1.4: Genetic Drift Over 3 Generations

Genetic drift over 3 generations of size n=14 using the Wright-Fisher model. The proportions of each allele C or T can fluctuate from generation to generation in a process called genetic drift. Over a sufficiently large number of generations one of the two variants will die out. Assuming no mutation, once a variant dies out, it can never reappear in subsequent generations.

group with n_2 individuals has $2n_2\gamma$. There are two ways that the two populations could end up having quite different proportions of the allele C at the locus over time. First, β and γ could be quite different so that the two groups have different proportions and be quite different from each other at the outset. The second is that even if β and γ are very similar and both about the same as α, the number of chromosomes with a C at the locus for the next generation for the first group will be modelled as a draw from Binomial$(2n_1, \alpha)$ and for the second as Binomial$(2n_2, \alpha)$. The proportions of the allele in that second generation are likely to be slightly different. Over subsequent generations, the proportion of chromosomes with a C at the locus will vary for the two populations independently of each other, so that as time goes on they can become more different. That is to say, they will experience genetic drift independently in different ways. Eventually, the allele could even become fixed in different states. One group could end up with only the C variant at the locus and the other with only T. If this is extended to cover the proportions of alleles at other SNPs and some of these contribute to some physical characteristics which are neither advantageous or disadvantageous, the two populations can start to look less alike over time. These processes can start to provide an explanation why some people who have ancestry in particular parts of the world look different to those from other parts of the world. However, the vast majority of differences between them across their genome will be unrelated to visible characteristics. A great many of the differences may have no functional effect but some will have less obvious effects such as contributing risk factors to particular disease conditions.

1.4 Models of Population Genetics

1.4.1 Population Trees

The situation described in the preceding paragraph could be represented as a population tree (Figure 1.5) such as those described by Cavalli-Sforza et al. (1994). The common ancestral population is at the root of the tree and the two derived subpopulations are at the leaves. In this simple example there is a single branching or bifurcation. A more complex relationship between a number of present-day subpopulations will also have a single common ancestor population. Such a relationship will involve a larger number of bifurcations. If it is assumed that there is negligible contact between populations after they have branched from each other, then a tree with s subpopulations will have $s - 1$ such bifurcations in it such as the one in figure 1.6 (reproduced from Cavalli-Sforza et al. (1994)). Models with population trees of this type will be considered in more detail in chapter 4.

A variant of bifurcating population trees arises if two or more populations split off at about the same time resulting in a tree with multifurcations rather than just bifurcations. An example of this is shown in figure 1.7. In an extreme case all the present-day subpopulations could be assumed to split from the shared ancestral population at about the same time, leading to a single multifurcation such as that in figure 1.8. Models with this type of tree will be considered further in chapter 3.

1.4.2 Admixture

These models assume that after each bifurcation or multifurcation, the resulting subpopulations and their descendants never have sufficient contact with each other to interbreed to any meaningful extent again. There are several models that allow this assumption to be relaxed in different ways. Suppose that in the above example, after one of the populations has left to go to the island, many generations pass

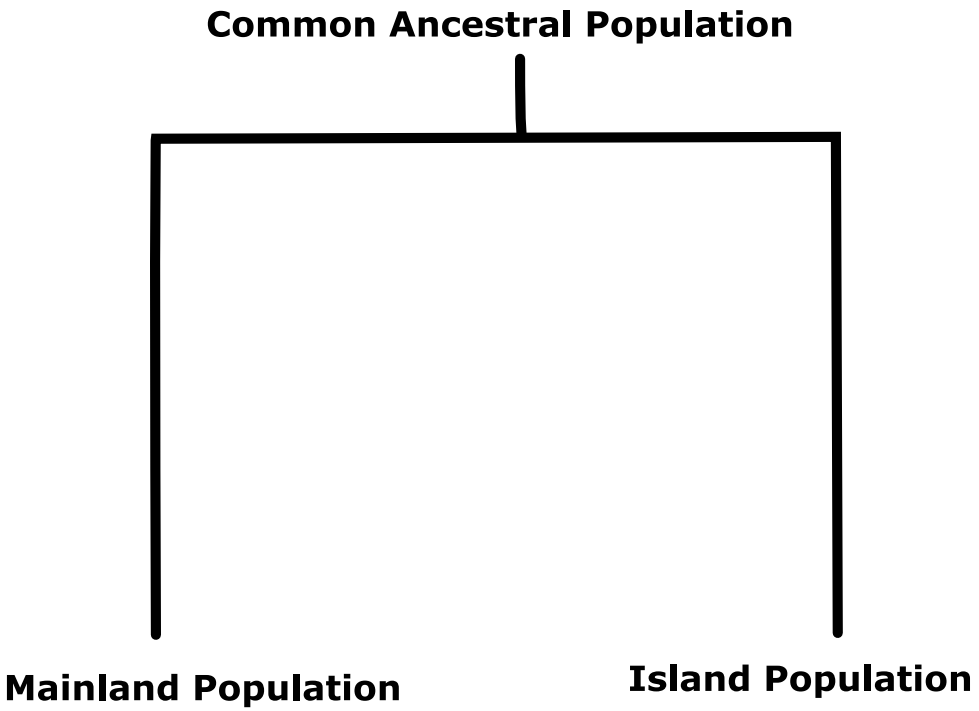

Figure 1.5: The Simplest Population Tree
In this simple example of a population tree, a common ancestral population splits into one subpopulation that remains on the mainland and another subpopulation that moves to an island. The fork or bifurcation represents the split and lines can be thought of as representing periods of genetic drift. In this model the two resulting subpopulations have no further reproductive contact with each other.

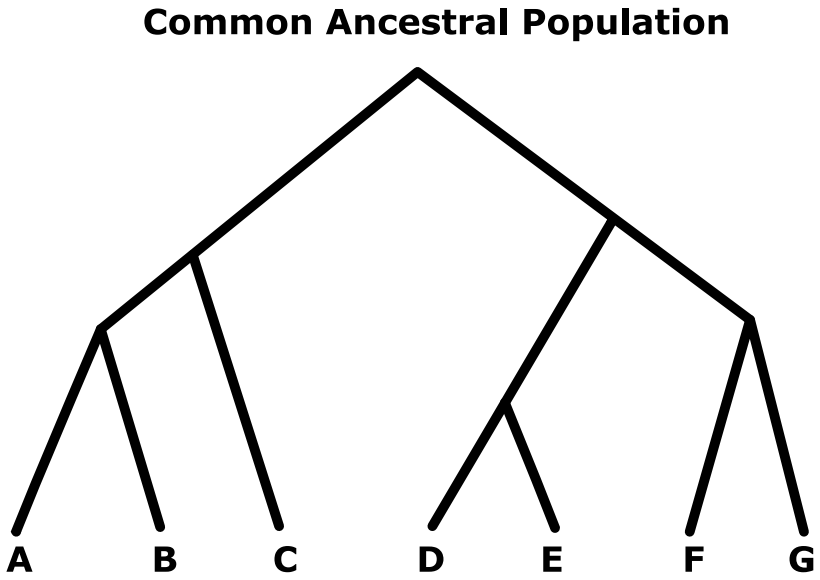

Figure 1.6: Example Population Tree for 7 Subpopulations
An example population tree with 7 subpopulations descended from a common ancestral population. With 7 subpopulations lettered A to G, there will be 6 bifurcations including the one at the root, one less than the number of subpopulations.

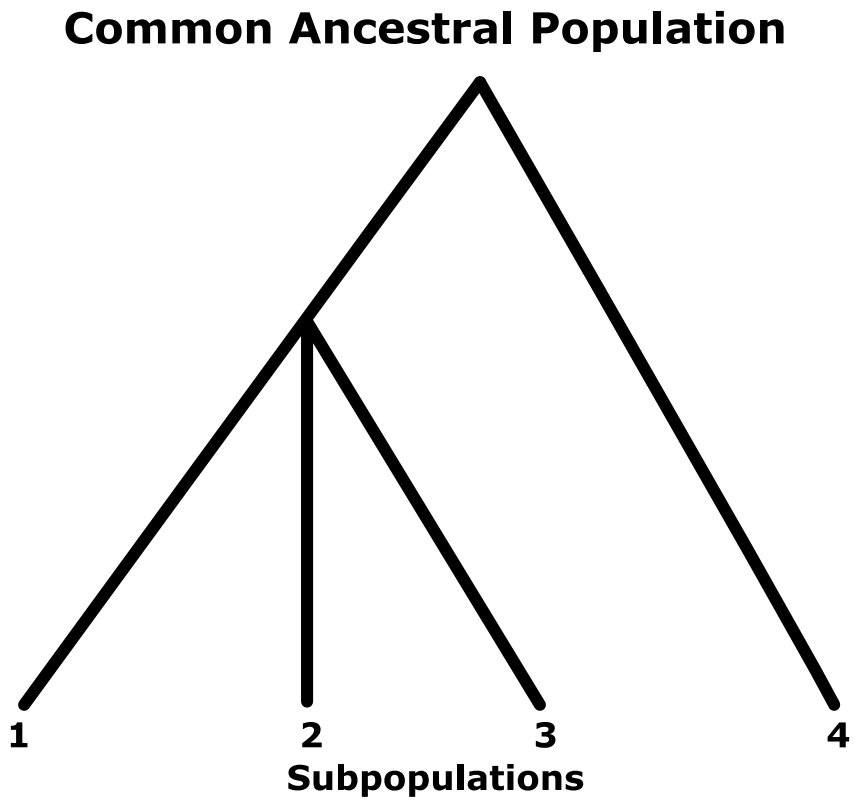

Figure 1.7: A Population Tree With a Multifurcation
In this example of a population tree, a common ancestral population splits into one subpopulation (subpopulation 4) and another which further undergoes a multifurcation resulting in subpopulations 1 to 3.

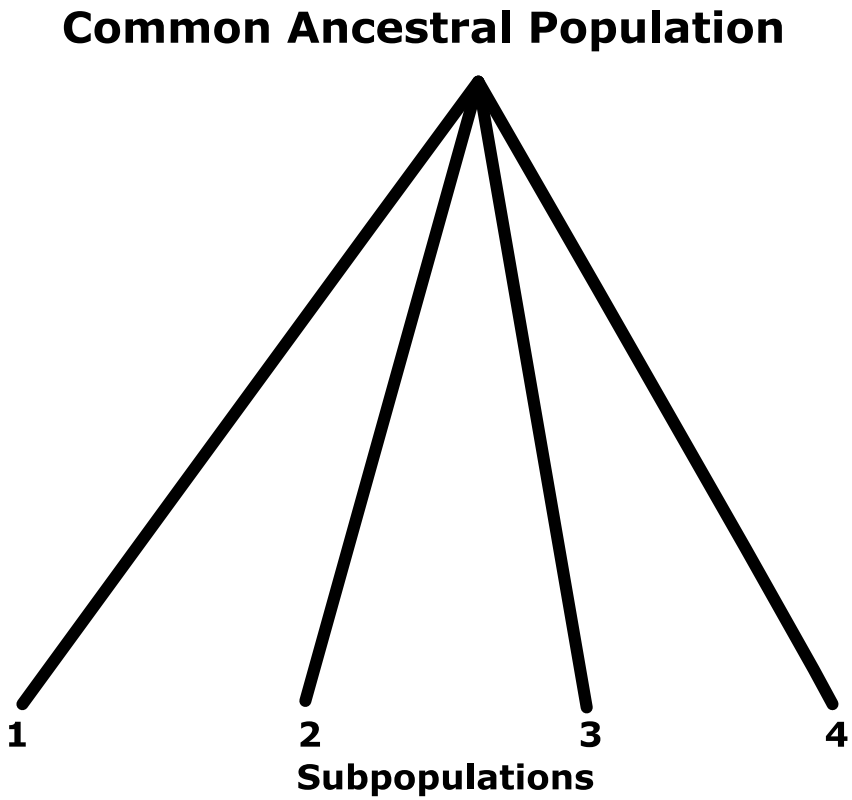

Figure 1.8: A Population Tree With Only a Multifurcation
In this example of a population tree, a common ancestral population undergoes a single multifurcation resulting in subpopulations 1 to 4.

before a second island is discovered. The population on the mainland splits again with some moving to this second island. Sometime after, the population of the first island become aware of this second island and some of these also choose to leave and move to the second island. The populations on the mainland and first island then continue to experience genetic drift independently as before but those who have moved to the second island meet and eventually interbreed. Thus a new third population is created on the second island that has some of its genetic material from the mainland population and some from the first island's population. This kind of population is called an admixed population. Rather than being represented by a tree, it is more naturally represented as a network. The story above could result in the network shown in figure 1.9. This can be modelled as a one-off event as by Wang (2003) as depicted in the figure or as a continuous inflow to the second island as by Roberts and Hiorns (1962). Admixture will be considered in more depth in chapter 5.

1.4.3 Many Islands Model

Hartl and Clark (1997) review yet another type of model, in which the populations become isolated from each other as before but the assumption that they have negligable reproductive contact with each other and experience genetic drift independently is relaxed in a different way. Here, a more significant number of individuals are assumed to move between the populations at each generation (figure 1.10). If the populations are sufficiently large to make genetic drift negligible compared to the effects of migration between the islands, the allele frequencies converge over time to a common frequency. Otherwise, the populations become differentiated and experience genetic drift differently but not entirely independently of each other. The allele frequencies in the resulting subpopulations are not as different as if they had been entirely isolated. For example, in a number of subpopulations that experience drift independently and separately, a variant at a

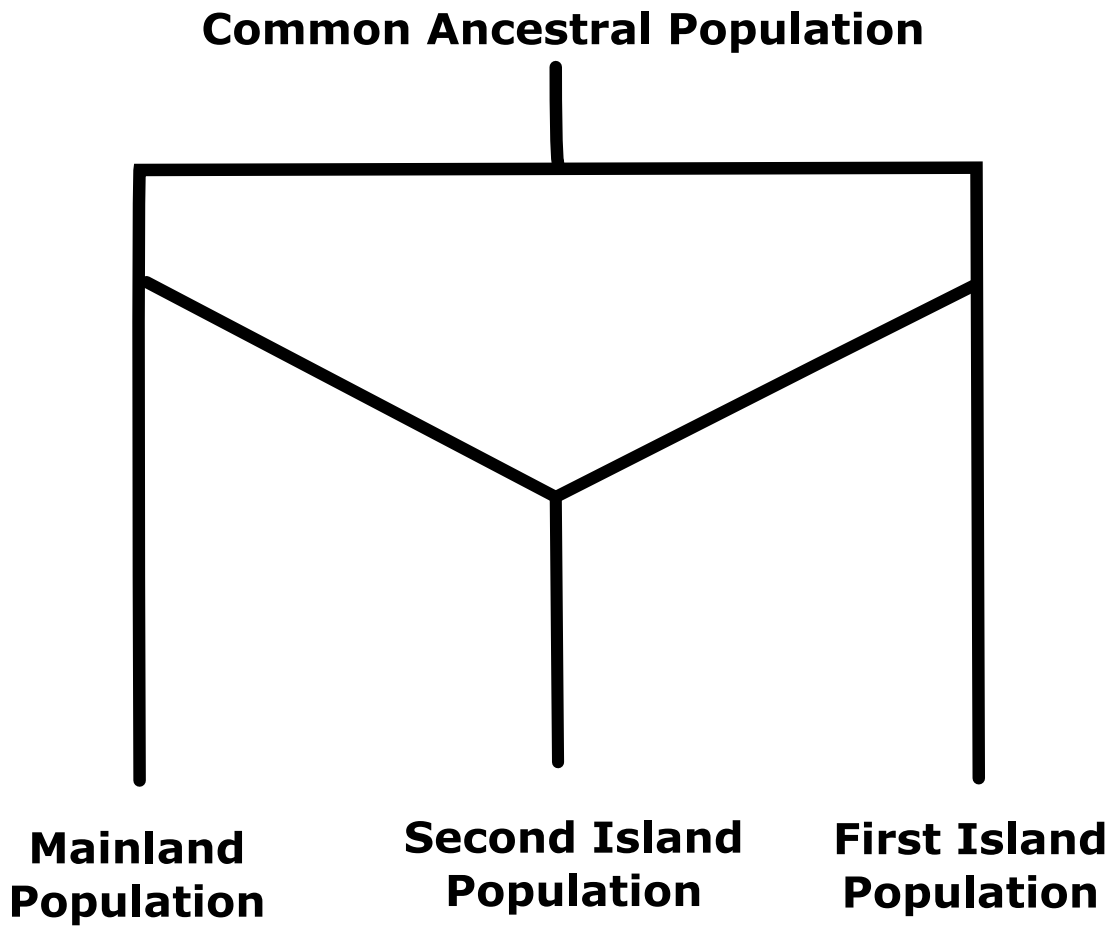

Figure 1.9: A Simple Admixture Network
In this example of admixture, a common ancestral population splits into one subpopulation that remains on the mainland and another subpopulation that moves to a first island. The only reproductive contact between the two groups then occurs when some of each of these two populations discover and move to a second island where they meet each other and form a new third subpopulation together which has no further contact with the populations remaining on the mainland or the first island.

locus can become fixed or die out in one or some of the subpopulations but not others. Once the variant has become fixed or died out in that subpopulation then it will remain in that state for all time if it is isolated and no mutation is assumed. However, if a number of individuals can move between subpopulations, a variant that is missing from one of the subpopulations that still exists in another can be reintroduced by immigration from other subpopulations. If the islands form a linear chain (figure 1.10 left) and individuals can only move into a neighbouring island subpopulation but no others, this can lead over time to gradients of allele frequency either increasing or decreasing from the first to the last island in the chain. Much more complicated variants of this island model are possible such as those depicted in figure 1 of Evanno et al. (2005).

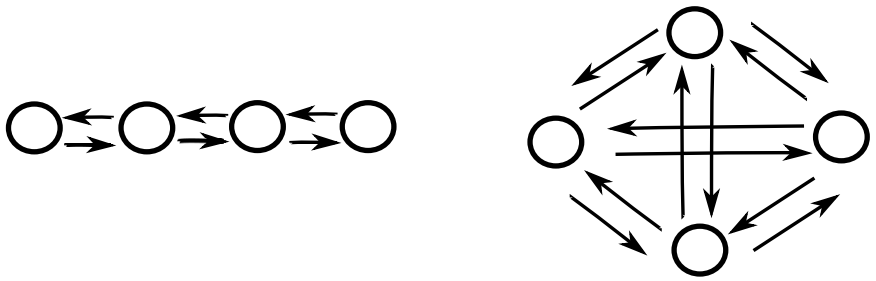

Figure 1.10: Two Examples of the Many Islands Model
In this figure, the four circles in each of these two examples represent subpopulations that have arisen from a bifurcation process such as that depicted in figures 1.5 and 1.9. However, now instead of having no further reproductive contact, they exchange individuals with other subpopulations each generation. In the right hand example, any subpopulation can exchange individuals with any other. In the left hand example, the subpopulations can only exchange individuals with a neighbouring subpopulation and no others.

Any of these previous models describe different types of population structure. Population structure arises from hidden relatedness. To think of population structure another way, although all pairs of individuals will have a most recent common ancestor, two individuals from the same geographic location are more likely to have had a most recent common ancestor who lived more recently than two individuals from more distant geographical locations. Those who share a common ancestor that lived more recently tend to have more of their genome in common. Scaling this up to the level of local populations, if sufficiently representative samples of DNA were taken from groups of individuals in a number of different locations, the

proportions of each allele observed at each locus on the genome for those populations in locations with more recent contact with each other will tend to be less different over the genome than those from populations that have been isolated for a long time.

1.5 Some Existing Techniques for Analysing Population Structure

AMOVA (Analysis of Molecular Variance) is a statistical technique, similar to Analysis of Variance (ANOVA), that is used to detect population structure and is credited to Excoffier et al. (1992). It looks for the amount of variation between individual's chromosomes within samples from local populations and between such samples or, at higher strata, between groups of local samples. The difference between two chromosomes is the Euclidean distance, where in its most basic form, one is counted for every independent locus at which the two chromosomes differ and the square root of the total taken as the distance. If there is no population structure then almost all the variation will be within samples. If there is population structure in the samples, then significant amounts of the variation will be observed between samples. Significance can be tested by permutation. There is an implementation of AMOVA in the software ARLEQUIN (Schneider et al., 2000).

PCA (Principal Component Analysis) has been around since Pearson (1901). Price et al. (2006) describe how it can be used in the context of population genetics. In effect what PCA does is project a dataset with d_2 dimensions into an d_1 dimensional space (where $d_1 < d_2$). The main steps are

1) Subtract the mean of each of the d_2 dimensions of the dataset from each datapoint, effectively translating the dataset so that it is centred on the origin.

2) A $d_2 \times d_2$ covariance (or correlation) matrix is calculated for the resulting dataset.

3) The eigenvalues and unit eigenvectors are calculated for this covariance matrix. This produces d_2 eigenvalues $\mathfrak{e}_1,...,\mathfrak{e}_{d_2}$ which can be ordered by size, greatest first.

4) The eigenvectors corresponding to the first d_1 eigenvalues in decreasing order are the principal components. The eigenvectors are orthogonal and are linear combinations of the original d_2 dimensions or variables.

The size of an eigenvalue is proportional to the proportion of total variance of the data projected onto its corresponding eigenvector. The d_1 largest of these account for the most variance possible within the dataset using only d_1 dimensions. This can be useful to approximate and visualise huge datasets. When d_1= 2 or 3 the dataset can be represented visually in 2D or 3D scatterplots. Even when $d_1 > 3$, plots can be produced of the data using pairs of these principal components as axes. These visualisations can be used to identify clusters in genetic data that can correspond to subpopulations whose members are more closely related to each other than they are to members of other clusters and thus to signal the presence of population structure in the data. The software EIGENSTRAT (Price, 2017) described by Price et al. (2006) was developed for this purpose.

The methods used in the package STRUCTURE (Pritchard, 2017) are due to Pritchard et al. (2000). Instead of using some measure of distance to describe the differences between the chromosomes in the sample and forming clusters that way, it works by assuming that Hardy-Weinberg equilibrium exists within each cluster and that there is no linkage disequilibrium between loci. Recall that for humans, each individual has two versions of each chromosome. In Hardy-Weinberg equilibrium (HWE), the probability of observing an allele at a locus on the second version chromosome of the individual is independent of the allele observed at that locus on the first version of that chromosome. These assumptions about linkage and HWE imply that each allele at each locus for each individual is an independent

draw from a probability distribution given the subpopulations (cluster) of origin of each individual and the allele frequency for each subpopulation at each locus. The method seeks to find the allocation of individuals to subpopulations (clusters) that most closely satisfies these assumptions. It does this using Bayesian methods and MCMC. Denote $\mathcal{X}_{\zeta i}$ as the genotype of the individual ζ at the ith locus, $\mathcal{Z}_\zeta$ as the subpopulation that individual ζ belongs to and $\pi_{\eta ij}$ as the frequency of allele η in subpopulation j at locus i. Then $Pr(\mathcal{Z}, \pi|\mathcal{X}) \propto Pr(\mathcal{Z})Pr(\pi)Pr(\mathcal{X}|\mathcal{Z}, \pi)$ by Bayes' Theorem. $Pr(\mathcal{Z})$ and $Pr(\pi)$ are priors. Pritchard et al. (2000) suggest a discrete uniform and Dirichlet priors, respectively, for these. The steps of the MCMC process are then to choose a set of allele frequencies given the data and the subpopulation that the individuals belong to, from $\pi|\mathcal{X}, \mathcal{Z}$ and then to choose a subpopulation for each individual given the data and the set of allele frequencies, from $\mathcal{Z}|\mathcal{X}, \pi$. This results in samples from the posterior distributions for $\mathcal{Z}$ and, as a by-product, for π. One of the interesting features of STRUCTURE is that it can be used to help decide how many clusters or subpopulations, $\mathcal{K}$, there should be. It produces estimates of the probability of the data for different values of $\mathcal{K}$, $Pr(\mathcal{X}|\mathcal{K})$, the marginal likelihood. The original paper cautions that while these do seem to produce plausible results in practice, it should only be used as a guide. Evanno et al. (2005) remarks that from their simulations using STRUCTURE, finding the maximum of the modulus of the second order (with respect to $\mathcal{K}$) of the rate of change of the likelihood of $\mathcal{K}$ was a better predictor of the true value of $\mathcal{K}$ than the marginal likelihood itself. Falush et al. (2016) describe how similar output from STRUCTURE can arise from quite different population histories. Additional information is needed to distinguish between them.

There have also been attempts to model genetic drift to describe the relationship between clusters or subpopulations more directly. One such approach is developed by Nicholson et al. (2002) which assumes all the present-day subpopulations arose from their common ancestral population following a single multifurcation. The genetic drift experienced by each subpopulation is modelled using a modified Nor-

mal (Gaussian) distribution. If the allele frequency at locus i is π_i in the ancestral population and the genetic drift for subpopulation j is quantified as c_j then the subpopulation's allele frequency is modelled as a $N(\pi_i, \pi_i[1-\pi_i]c_j)$ distribution with the modification that all resulting values above 1 are taken to be 1 and all resulting values below 0 are taken to be 0. Markov Chain Monte Carlo techniques are used to obtain posterior distributions for all the ancestral allele frequencies π_i, present-day subpopulation allele frequencies α_{ij} and genetic drifts for each subpopulation c_j. The bulk of this thesis will be concerned with developing this model to cover much more complex population tree structures and admixtures. As such, this model will be described in much more depth in chapter 3 and its advantages and disadvantages discussed in chapters 3 and 4.

1.6 Review of Admixture Models

There is a considerable body of previous work on statistical models of admixture, a new model for which will form the meat of chapter 5. One of the earliest examples is by Bernstein (1931). Although DNA had been discovered in cells as early as 1869, the role of DNA in transmitting inherited traits would not be confirmed until the early 1950s. Nevertheless, models of pairs of genes on chromosomes transmitting heritable traits were already being developed. These models were able to explain the proportions of phenotypes observed to be inherited from parents by their offspring. Despite the fact that Gregor Mendel's work on genetics, describing such a model to explain the results of his plant experiments, had been published much earlier (Mendel, 1866), it was only becoming fully accepted as having much wider implications by 1931. The phenotype that Bernstein's paper is concerned with is that of blood groups. (A phenotype is an observable characteristic that results wholly or in part from a particular genotype.) Karl Landsteiner had pioneered work on blood groups in the early 20th century and had been awarded the Nobel prize for medicine in 1930. The theory behind blood groups was also

still at an early stage. The Rhesus types, for example, would not be discovered until later in the decade. The model described by Bernstein (1931) uses the four blood groups, A, B, AB and O. The alleles for O are recessive while those for A and B are codominant. Bernstein takes the proportion of the A, B and O phenotypes in a subpopulation, $\mathfrak{a}$, $\mathfrak{b}$ and $\mathfrak{o}$ respectively and calculates $\mathfrak{p} = 1 - \sqrt{\mathfrak{o} + \mathfrak{b}}$, $\mathfrak{q} = 1 - \sqrt{\mathfrak{o} + \mathfrak{a}}$ and $\mathfrak{r} = \sqrt{\mathfrak{o}}$. The values $\mathfrak{p}$,$\mathfrak{q}$ and $\mathfrak{r}$ will sum to a value close to 1. Let the error be $\mathfrak{D} = 1 - (\mathfrak{p} + \mathfrak{q} + \mathfrak{r})$. To obtain three quantities that sum closer to 1, he then computes $\dot{\mathfrak{p}} = \left(1 - \sqrt{\mathfrak{o} + \mathfrak{b}}\right)\left(1 + \frac{\mathfrak{D}}{2}\right)$, $\dot{\mathfrak{p}} = \left(1 - \sqrt{\mathfrak{o} + \mathfrak{b}}\right)\left(1 + \frac{\mathfrak{D}}{2}\right)$, $\dot{\mathfrak{r}} = \left(\sqrt{\mathfrak{o}} + \frac{\mathfrak{D}}{2}\right)\left(1 + \frac{\mathfrak{D}}{2}\right)$. The error is then much smaller, $\frac{\mathfrak{D}^2}{4}$.

In the paper, an equilateral triangle of height 1 is drawn where each side represents one of $\dot{\mathfrak{p}}$, $\dot{\mathfrak{q}}$ and $\dot{\mathfrak{r}}$. A point for a subpopulation is plotted at a distance $\dot{\mathfrak{p}}$ perpendicular to the $\mathfrak{p}$ side, $\dot{\mathfrak{q}}$ perpendicular to the $\mathfrak{q}$ side and $\dot{\mathfrak{r}}$ perpendicular to the $\mathfrak{r}$ side. For two such populations, two points can be drawn as shown in figure 1.11. An admixture of these two populations, it is argued, will have a point that will lie along a line connecting these two points. The position of the point on the line will be proportional to the proportions of the two populations represented in the admixture. If it is an admixture created from 25% of population 1 with 75% of population 2, then the point will be three quarters of the way from the point for population 1 to the point for population 2 as shown in figure 1.11. More generally, this is just saying that $\dot{\mathfrak{p}}_{\mathfrak{M}} = w\dot{\mathfrak{p}}_1 + (1 - w)\dot{\mathfrak{p}}_2$, where $\dot{\mathfrak{p}}_{\mathfrak{M}}$ is the value of $\dot{\mathfrak{p}}$ for the admixed population. $\dot{\mathfrak{p}}_1$ and $\dot{\mathfrak{p}}_2$ are the $\dot{\mathfrak{p}}$ for subpopulations 1 and 2, respectively, and w is the admixture parameter, the proportion of subpopulation 1's contribution to the admixture. This type of expression for an admixture will appear again in later models. The Streng triangle idea relies on the theorem that for any point in or on an equilateral triangle, the sum of the shortest distances from that point to each of the three sides is equal to the height (the shortest distance from a corner to the opposite side) of the triangle. The remainder of Bernstein's 1931 paper discusses the use of the model on data for a number of populations to infer historical population flows. Given that it must do so without a modern

understanding of the underlying genetic basis for the blood groups, no concept of genetic drift and with relatively small datasets which are confined to this one phenotype, it is a good simple early model. Nevertheless, the paper remains one of historical interest.

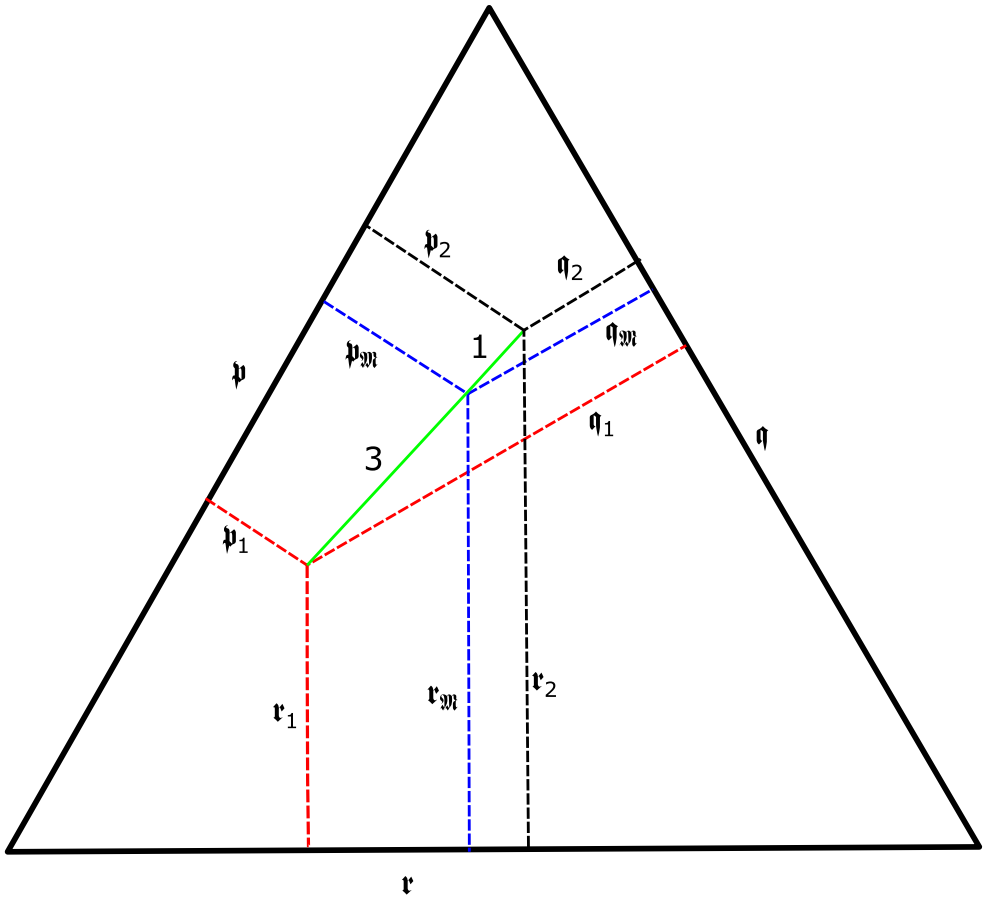

Figure 1.11: Streng Triangle

A similar diagram to that shown in Bernstein (1931) where it is referred to as a Streng triangle. The formulae for $\dot{\mathfrak{p}}_1$,$\dot{\mathfrak{q}}_1$ and $\dot{\mathfrak{r}}_1$ are calculated from the proportions of blood groups A B and O in subpopulation 1, those for $\dot{\mathfrak{p}}_2$,$\dot{\mathfrak{q}}_2$ and $\dot{\mathfrak{r}}_2$ are similarly calculated for subpopulation 2. These give rise to two points within the triangle, one for each subpopulation. These are the points with distance $\mathfrak{p}$ measured perpendicularly from the side labelled $\mathfrak{p}$, $\mathfrak{q}$ measured perpendicularly from the side labelled $\mathfrak{q}$ and $\mathfrak{r}$ measured perpendicularly from the side labelled $\mathfrak{r}$. An admixture of these two subpopulations is then expected to have its point on the line (shown in green) joining these two points. Its position along the line would be related to the proportions of the subpopulations represented in the admixture. If there were a 3:1 ratio of subpopulation 2 to subpopulation 1 in the admixture (i.e., 25% of subpopulation 1 and 75% of subpopulation 2) then the point on the triangle for the admixed population would be three quarters of the way along the line from the point for subpopulation 1 to subpopulation 2. The values, $\dot{\mathfrak{p}}_\mathfrak{M}$, $\dot{\mathfrak{q}}_\mathfrak{M}$ and $\dot{\mathfrak{r}}_\mathfrak{M}$, expected for the admixed subpopulation can then be measured and read off.

By the 1960s, the structure and role of DNA in genetic inheritance was known.

An example of one of the pieces of work on admixture in the aftermath of this major breakthrough is by Roberts and Hiorns (1962), which presents a deterministic dynamic model for the allele frequency in an admixed population where the admixture is not treated as a single event but as a continuous flow of people from the parent populations into the admixed population. Later, in Roberts and Hiorns (1965) they develop a least squares approach for estimating the proportionate contribution to the admixed population from each parent population. The model is relatively simple. If $\mathbf{Q}$ is a matrix of allele frequencies with a column for each locus and a row for each parental subpopulation, $\mathbf{q}$ is a column vector of allele frequencies for the admixed population and $\mathbf{w}$ is a column vector of contribution proportions for each parental population to the admixed population, then their model is

$$\mathbf{w}^T = \mathbf{q}^\mathbf{T}\mathbf{Q}^T \left[\mathbf{Q}\mathbf{Q}^T\right]^{-1}. \tag{1.1}$$

However, they omit to mention that the resulting vector, $\mathbf{w}$, has elements that do not necessarily sum to one and that they perform an additional step of scaling the vector so that the elements do sum to one. It is also not mentioned that the elements of $\mathbf{w}$ are not necessarily positive. Indeed, in the helpful worked example they give in the paper of African, Indian and Portuguese subpopulations contributing to an admixture of Nordestinos in São Paulo, Brazil, changing a single allele frequency in the admixed population would have led to a negative element in $\mathbf{w}$. So this model does not necessarily produce useful results but is another interesting early attempt to quantify the contribution of parental subpopulations to an admixed subpopulation.

Thompson (1973) introduces genetic drift into a model of admixture. It takes account of genetic drift since the admixture event. It parametrises the measure of drift in terms of the number of generations, t and effective population size, N. A normal distribution model of genetic drift is used with the mean of the present-day allele frequency, representing the earlier allele frequency and the variance used is $\frac{t}{8N}$. It also models the sampling variance as being $\frac{1}{8n}$, where n is the

sample size. It also features the admixture event being modelled as $\alpha_{admix} = w\alpha_1 + (1-w)\,\alpha_2$, where α_{admix} is the admixed subpopulation's allele frequency, α_1 and α_2 are the allele frequencies in the two parent subpopulations, and w, the proportion that parental subpopulation 1 contributes to the admixture, is the admixture parameter. It is interesting that $\alpha_{admix} = w\alpha_1 + (1-w)\,\alpha_2$ is the same as implied by the Streng triangle of Bernstein (1931).

Chikhi et al. (2001) use a likelihood MCMC approach to model drift since a single admixture as well as the admixture coefficient. The underlying model, is still similar to that of Thompson (1973). They were limited by the technology available at the time. Their simulated data sets contained data on only 20 loci, leading to posterior distributions for the admixture parameter that were very wide. When applied to a human data set of Jamaicans, it was found that the European admixture component had a 95% credible interval of 1.9% to 14.1%. Wang (2003) develops this model further and takes a maximum likelihood approach. The drift of the two parent subpopulations of the admixed subpopulation since their common ancestral population is explicitly incorporated into the model. The approach still only models a single admixture event and, as will be shown in chapter 5 further episodes of drift can be incorporated. Wang uses his own modified version of the diffusion approximation of Crow and Kimura (1970) to model genetic drift because the diffusion approximation itself is too computationally intensive. Choisy et al. (2004) compare the MCMC approaches with the others available at the time and finds that they perform better in situations where the parental populations of the admixture are not greatly differentiated from each other and where the admixture proportions are far from being 50:50, particularly when effective population sizes are low. Nevertheless, Choisy et al. (2004) do not consider the extra time that MCMC methods take to be worthwhile.

Excoffier et al. (2005) takes an interesting alternative approach using Approximate Bayesian Computation (ABC). Like Wang (2003), they model a single admixture and the drift before and after the admixture. Their model also allows for mu-

tations. Instead of using MCMC, it uses another approach. In the first step, parameter values are drawn from their priors. A dataset is simulated according to those priors and summary statistics of that dataset are calculated. This is repeated a large number of times. The second step is to calculate the summary statistics for the actual dataset. In the third step the large set of summary statistics are compared to that of the real dataset. A metric, such as Euclidean distance between each of the million simulated summary statistics and the real dataset's summary statistics, is calculated. In the fourth step, a small proportion of the simulations with the lowest such metrics are retained. The rest are discarded. The fifth step is to use local and weighted linear regression on the retained simulations to estimate the parameter values that generated the real dataset.

Patterson et al. (2006) take an entirely different approach, using Principal Components Analysis (PCA) to estimate admixture parameters. It makes no attempt to model genetic drift or the admixture process directly but does have the advantage of being fast and being able to be used on an admixed population when there are many more than two parental populations. They point out the similarities that this PCA approach has to clustering approaches. These approaches are a useful and relatively fast alternative when only the admixture proportions in the present day subpopulations are required and the genetic drift processes before and after the admixtures are not of interest. This work led to the development of the EIGENSTRAT package.

The paper of Alexander et al. (2009) is interesting in that it also was motivated by an attempt to control for population structure in association studies. Like Patterson et al. (2006) it also does not model genetic drift or the admixture process directly but does seek to estimate the proportions that a number of parental populations contribute to the genomes of individuals within an admixed population. This work led to the development of the package ADMIXTURE (Alexander et al., 2017). It takes a likelihood-based approach and then uses an EM algorithm to maximise the likelihood. However, they wanted their program to run faster than a

pure EM approach and after using EM to reach the neighbourhood of a maximum, they switch to a block relaxation algorithm to complete the process faster. They report that their program has a running time of a similar order to EIGENSTRAT.

It was mentioned earlier that Roberts and Hiorns (1962) modelled admixture as a continuous process rather than a one-off event. Models of admixture as a process in time rather than an event are still developed. The paper of Verdu and Rosenberg (2011) is one such example. It treats time as a discrete quantity measured in generations and develops a stochastic model for the distribution of allele frequencies in an admixed population with an initial contribution from two parental populations and then different contributions flowing into it from the two parental populations in each generation. It does not take genetic drift within the parental populations into consideration. The larger the number of generations, and the smaller the sizes of these parental populations, the more that will be a problem. It is, nonetheless, easily generalisable to more than two parental populations.

Frichot et al. (2014) use a least-squares method. Again, genetic drift is not modelled, just the admixture proportions. The method makes no assumption that Hardy-Weinberg Equilibrium has been reached. They argue that this makes this approach better than Alexander et al. (2009)'s approach when there are reasons to doubt that assumption. They also report that in practical tests their algorithm was faster than that of Alexander et al. (2009) particularly so as the number of parent populations increases.

1.7 Applications of Population Structure Models

Historically, a common motivation for the development of these models is for use in anthropology; the reconstructing of unrecorded human history to describe the spread of humans across the planet. They are also applicable to other species. There are, however many other practical applications of these population genetic

models. One is in forensics using methods such as those described by Kayser and de Knijff (2011). In forensics, markers from a DNA sample from a crime scene are compared to those from a suspect. There may be a match. If so, the chance of an equal (or better) match from a random member of the population is assessed. The problem that can arise when there is population structure is that a DNA profile that is uncommon in the general population may be more common in a subpopulation associated with ethnicity or location or both. If ethnicity or location played any role in the choice of suspect, comparing their profile wrongly, to that of the general population rather than the subpopulation could lead to an overstatement of the probability that the crime scene sample belongs to the suspect.

Another application is in controlling for population structure in Genome Wide Association Studies (GWAS). The objective of a GWAS is to discover which locations in the human genome are associated and potentially causal for particular phenotypes. Often the phenotype of concern is susceptibility to a particular disease. To do this allele frequencies in samples with and without that biological quality are analysed. A problem arises in GWAS in that if population structure exists in the samples and is not taken account of then it can lead to an elevated rate of false associations between loci on the genome and biological qualities. Balding (2006) describes this problem and many of the methods that have been used to take account of it.

Chapter 2 describes the generic methods that will be used throughout the rest of the thesis.

Chapter 2

Methods

A number of generic statistical methods that will be used in the following chapters will be discussed here. The chapter will start off discussing Bayesian Inference and Markov Chain Monte Carlo methods before looking at Gibbs Sampling and Metropolis-Hastings sampling. An adaptive algorithm used within Metropolis-Hastings sampling to help ensure it performs well will then be discussed. Rejection sampling will then be considered and a customised version developed for this project will be described. The chapter will then move on from sampling to describe the Neighbour Joining Algorithm of Saitou and Nei. Gelman's R statistic which is used as evidence that a model has not converged properly is described. Watanabe Aikeke's Information Criterion, a method for choosing between potential models by balancing how well they represent the data agains their complexity, is introduced. Finally, Post Predictive Checking, a method for determining how well a model represents the important aspects of the data is then described.

2.1 Bayesian Inference and Markov-Chain Monte Carlo (MCMC)

The Bayesian approach is the main approach to analysing data that will be used during most of this thesis. It is desirable to know what the probability distribution of a set of parameters, $\boldsymbol{\theta}$, in a probabilistic model, is given the available data, $\mathbf{D}$, that is $p(\boldsymbol{\theta}|\mathbf{D})$. One way of approaching this is to use Bayes' Rule (Gelman et al., 2013):

$$p(\boldsymbol{\theta}|\mathbf{D}) = \frac{p(\boldsymbol{\theta})\, p(\boldsymbol{D}|\boldsymbol{\theta})}{p(\mathbf{D})}. \tag{2.1}$$

Here p(.) might be probability densities, mass functions or a combination of both. In the simplest cases, the left hand side can be found analytically. The distribution $p(\boldsymbol{\theta})$ is known as the prior distribution. It encodes beliefs about the likely and unlikely values of the parameters before the data at hand have been examined. $p(\boldsymbol{D}|\boldsymbol{\theta})$, the probability of the data being observed conditional on the parameters encapsulates the mechanics and distributional assumptions of the model being used and is proportional to the likelihood function viewed as a function of θ. Both of these terms involve making probabilistic assumptions about the model parameters and the relationship of data to them respectively. In practice, since well-known distributions are often chosen for these terms, expressions can usually be derived for these, albeit often very complicated ones, particularly in the case of hierarchical models where distributions involve some parameters which in turn depend on distributions involving other parameters in a tree-like way. The denominator, $p(\mathbf{D})$, sometimes called the marginal likelihood or the evidence can in practice be the most problematic. Sometimes, it can be found using the Law of Total Probability, that is summing $p(\boldsymbol{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ over all possible values of the parameters, $\boldsymbol{\theta}$. In practice, this is often impractical or impossible. However $p(\mathbf{D})$ is constant with respect to $\boldsymbol{\theta}$. It is therefore common to use a version of Bayes' rule that leads to an unnormalised expression for the distribution $p(\boldsymbol{\theta}|\mathbf{D})$, the posterior

distribution:

$$p(\boldsymbol{\theta}|\mathbf{D}) \propto p(\boldsymbol{\theta})\, p(\boldsymbol{D}|\boldsymbol{\theta})\,. \tag{2.2}$$

One means of making inference is to explore $p(\boldsymbol{\theta}|\mathbf{D})$ by repeatedly sampling from $p(\boldsymbol{\theta})\, p(\boldsymbol{D}|\boldsymbol{\theta})$. However, it is often not possible or practical to sample directly from such a posterior distribution. Again, this is particularly the case in complicated hierarchical models. One idea that has been developed to sample from the posterior distribution in a less direct way is to exploit Markov Chain Monte Carlo (MCMC) methods. These methods depend on producing a Markov chain of simulated values of $\boldsymbol{\theta}$ (i.e., one where the next simulation only depends on the last), the limiting distribution of which is constrained to be the posterior distribution. As the chain approaches equilibrium, successive draws are taken from distributions that become better approximations to the posterior distribution. The idea is that these successive draws eventually become close enough to being representative of the posterior distribution that a series of them can be used to approximately describe the properties of that posterior distribution that are of interest. Note however that the draws are correlated, since they are taken from a Markov chain.

2.2 Gibbs Sampling

Gibbs sampling, (e.g., Gelman et al., 2013), is an MCMC procedure that allows sampling from the posterior distribution. It does this in a way such that over a sufficient number of iterations or draws, the distribution being drawn from becomes a better approximation to the posterior distribution and the set of successive draws become more representative of a (correlated) sample from the posterior distribution. It starts by partitioning the set of parameters, $\boldsymbol{\theta}$ into a number, ϑ, of subsets. In each iteration, each of the ϑ subsets of parameters is drawn from its conditional distribution given the values of the others subsets. This requires finding an expression for the distribution of the ιth subset condi-

tional on the others up to proportionality. Using Bayes rule as before, this is $p(\boldsymbol{\theta}_\iota|\mathbf{D}, \boldsymbol{\theta}_{-\iota}) \propto p(\boldsymbol{\theta}_\iota)\, p(\boldsymbol{D}, \boldsymbol{\theta}_{-\iota}|\boldsymbol{\theta}_\iota) = p(\boldsymbol{\theta}_\iota, \boldsymbol{D}, \boldsymbol{\theta}_{-\iota}) = p(\boldsymbol{\theta}, \boldsymbol{D}) = p(\boldsymbol{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$. This is known as the full conditional for $\boldsymbol{\theta}_\iota$. This is easiest to do when the subsets each contain just one parameter. There are, however, occasions when it is desirable to group parameters together and draw them from a joint full conditional distribution, such as when the values of the parameters are highly correlated in the posterior distribution, since doing so can lead to the chain exploring the joint posterior distribution more efficiently. It sometimes happens that the form of these conditionals are such that they can be sampled from directly, e.g., when they are a known standard distribution. Nonetheless, in many practical cases the forms of these conditionals are more complicated and another means of sampling from them needs to be used.

2.3 Metropolis-Hastings Sampling Within Gibbs

Since it is not always possible to sample directly from the full conditional distributions, a number of alternative methods have been devised to perform this step within the Gibbs sampling framework. One of the most remarkable and useful of these is the Metropolis-Hastings algorithm (Hastings, 1970) which is also described in more detail by Gelman et al. (2013). First an approximate set of starting values, $\boldsymbol{\theta}^0$ are assigned to the parameters $\boldsymbol{\theta}$. The superscript 0 here is refers to the state of $\boldsymbol{\theta}$ at iteration $t = 0$. These can be rough guesses or any other estimates to provide a starting point for the process. At each iteration, i, a probability distribution is used to draw from to generate a proposed new value for $\boldsymbol{\theta}_\mathrm{i}$, $\boldsymbol{\theta}^*_\mathrm{i}$. The proposed new value of the parameter can depend on the value of the parameter from the previous iteration in some way. So, for example, if $\boldsymbol{\theta}_\mathrm{i}$ was just a single parameter, θ_i and the proposal distribution chosen was a normal distribution, it would be usual to choose that normal distribution to have mean θ_i^{t-1}, the value that the parameter had after the last iteration, and some variance σ^2 which can be chosen arbitrarily. To take

account of bias in non-symmetric distributions, a ratio is calculated. If $\mathfrak{g}\left(\theta_i^*|\theta_i^{t-1}\right)$ represents the probability density at the proposal, θ_{i}^*, of the proposal distribution with its parameters dependent on $\theta_{\mathrm{i}}^{t-1}$, then $\mathfrak{g}\left(\theta_{\mathrm{i}}^{t-1}|\theta_{\mathrm{i}}^*\right)$ represents the probability density at $\theta_{\mathrm{i}}^{t-1}$ of the proposal distribution with its parameters dependent on θ_{i}^*, that is the reverse of the proposed change to the value of the parameter. The ratio

$$\mathcal{Q} = \frac{\mathfrak{g}\left(\theta_{\mathrm{i}}^{t-1}|\theta_{\mathrm{i}}^*\right)}{\mathfrak{g}\left(\theta_{\mathrm{i}}^*|\theta_{\mathrm{i}}^{t-1}\right)} \tag{2.3}$$

is then determined. With a symmetric proposal distribution such as a normal distribution this ratio will always be 1 and this step can be omitted.

The next step is to use the expressions that have been found for the full conditionals to calculate both $p(\theta_{\mathrm{i}}^*|\boldsymbol{D}, \boldsymbol{\theta}_{-\mathrm{i}})$ and $p(\theta_{\mathrm{i}}^{t-1}|\boldsymbol{D}, \boldsymbol{\theta}_{-\mathrm{i}})$. The values of the other parameters used to calculate this probability density (or probability in the discrete case) are either those at iteration $t-1$, for a parameter that has not yet been updated at this iteration, or at iteration t, if it has, i.e., it is the most recent value for the parameter. The ratio

$$\rho = \frac{p(\theta_{\mathrm{i}}^*|\boldsymbol{D}, \boldsymbol{\theta}_{-\mathrm{i}})}{p(\theta_{\mathrm{i}}^{t-1}|\boldsymbol{D}, \boldsymbol{\theta}_{-\mathrm{i}})} \tag{2.4}$$

is calculated. In the next step an acceptance probability, γ is calculated where $\gamma = \min(\rho\mathcal{Q}, 1)$. With probability γ, the proposal is accepted and so θ_{i}^* is assigned as the value of θ_{i}^t, otherwise θ_{i}^t retains its value from the previous iteration, $\theta_{\mathrm{i}}^{t-1}$.

One of the attractions of this algorithm is that it is usually straightforward to turn into computer code. In practice, because of the low numerical values of the probability densities that are often involved and the distortions that can occur when computers have to represent very small positive values in digital floating point arithmetic, it is very often practically easier and more accurate if most of the Metropolis-Hastings algorithm calculations are carried out using logs. It is nevertheless, remarkable that in the long run, given a sufficiently large number of

iterations, this algorithm does produce a representative sample from the posterior distribution. (See Chib and Greenberg (1995) for details of why this works).

2.4 Adaptive Metropolis-Hastings Algorithms

In the previous subsection, it was mentioned that the variance of the proposal distribution in the Metropolis-Hastings algorithm within a Gibbs sampling framework could be chosen in an arbitrary way. However, some choices of variance lead to a larger number of iterations being needed in the Markov Chain before the resulting distribution can be said to be representative of the target posterior distribution than others. If the choice of variance is too large, the proposed new value for the parameter, θ_{i}^{*}, will tend to be further from the value it took at the last iteration, $\theta_{\mathrm{i}}^{t-1}$. This typically leads to lower probabilities of acceptance. This results in the parameter keeping its value from the last iteration more often, and it can end up doing that for many iterations at a time. If the parameter does not change value often enough it will take more iterations for the Markov Chain to produce a series of values for the parameters that will be representative of the posterior distribution being sampled, a scenario described as "poor mixing". Conversely, if the choice of variance is too small, the proposed θ_{i}^{*} will tend to be closer to $\theta_{\mathrm{i}}^{t-1}$ and while this will lead to a higher acceptance probability and prevent the Markov Chain sticking in the same way, the moves will be small and it will take more iterations for the Markov Chain to explore the full range of values and combinations of values which the posterior distribution covers, again "poor mixing".

There is therefore a "sweet spot", an optimum choice for the variance of the proposal distribution. The problem is that there is no easy way of knowing where it will be before starting the MCMC process. It can be found approximately by trial and error but where there are many such parameters this haphazard approach is often simply impractical.

Roberts and Rosenthal (2009) proposed a surprisingly simple adaptive MCMC algorithm aimed at solving this problem. Instead of simply guessing what the value of the proposal variance should be, they would let a computer algorithm learn about where the optimal value approximately is over a number of iterations at the beginning of the Markov Chain. First, an initial guess is made at the best proposal variance, $\sigma^2 = \exp(2ls_{\text{i}})$ where ls_{i} can take a chosen value. If there is no information about what a good value for ls_{i} would be, then setting $ls_{\text{i}} = 0$ is as reasonable a starting point as any.

Roberts and Rosenthal (2009) state that, in one dimension, the optimal acceptance rate is 0.44 (Gelman et al., 1996). So when θ_i represents a single parameter, the optimal choice of proposal variance will lead to an acceptance rate of about 0.44. Roberts and Rosenthal (2009) are clear that such a rule about the acceptance rate is only based on approximations and has not been rigorously proven. Nevertheless, even if only very approximately true, the algorithm resulting from targeting an acceptance rate of 0.44 will still produce better results than guessing the proposal variances.

After a particular number of iterations of the MCMC process, such as a batch of 100, the acceptance rate over these iterations can be calculated and the value of ls_{i} for the parameter can be adjusted accordingly. If the acceptance rate was less than 0.44 over the κth such batch, then ls_i can be decreased by $\frac{1}{\kappa}$ for the next batch. Similarly, if the acceptance rate was more than 0.44 over the κth such batch, then ls_{i} can be increased by $\frac{1}{\kappa}$ for the next batch. If this process is continued for a sufficiently large number of batches, the value of ls_{i} will tend towards an approximately optimal value. After a sufficient number of batches, the value of ls_{i} and therefore the proposal variance can be held at the approximately optimal value that has been found for the rest of the Gibbs sampling process.

It should be noted that the sample of iterations taken from the Markov Chain as an approximation to the posterior distribution should be taken after the proposal

variances are being held constant. This is because the adaptive part of the process violates the Markov property. Since the decision to change the value of ls_i after each batch depends on the last whole batch of iterations, the value of the next state of the Markov chain depends on what happened over the whole previous batch of iterations and not simply the present state of the Markov chain. Nevertheless, it is usual practice to discard a number of the early iterations of the Markov Chain anyway as "burn in", because, as noted in the previous sections, the series of states of the Markov chain only become draws from the target posterior distribution after a sufficient number of iterations have elapsed. If the number and size of the batches in the adaptive part of the process are chosen such that the iterations over which the adaptive process takes place end before the end of the "burn in" period of iterations then no difference is made to the total number of iterations needed before sampling from the approximate posterior distribution can begin.

If the adaption process were continued beyond the burn-in period there would be blocks of chain states with different variances in their proposal distributions. Within each block, the states would represent draws from the target distribution. However, there would be different speeds of mixing between each block. When the variance is nearest to the optimum value there will be faster mixing, while when it is further away, there will be slower mixing. There might not appear to be any particular problem on examination of the draws from the target distribution. It might appear that the whole space of the target distribution has been explored. However, the states that were visited when the mixing was at its slowest would be over-represented compared with other states and so the chain as a whole would not properly represent the target distribution.

2.5 Rejection Sampling

Metropolis-Hastings is not the only approach to drawing from a full conditional distribution within the Gibbs sampling framework. Another approach is rejection

sampling. In its simplest form, rejection sampling from a full conditional distribution consists of a number of steps. This can be done if the parameter for which it is the full conditional has support on a finite interval, $[\nu, \xi]$ say.

2.5.1 Simple Rejection Sampling

Simple rejection sampling is described in various sources such as Casella et al. (2004). Step 1 is to find the maximum value of the full conditional distribution or choose a value that is guaranteed to be above its maximum. There are various ways to do this. Finding the points where the derivative is 0 and using the second derivative to check whether the points are maxima rather than minima is one way. The derivative may not always be easy to find. A more brute-force approach is to choose a number, ϱ of points at regular intervals, $\nu, \nu + \frac{\xi-\nu}{\varrho-1}, \nu + \frac{2(\xi-\nu)}{\varrho-1}, \ldots, \nu + \frac{(\varrho-2)(\xi-\nu)}{\varrho-1}, \xi$, evaluate the full conditional at these points, and determine which of these values, λ, produced the highest value for the full conditional. It is then assumed that the maximum lies somewhere in the interval $\left[\lambda - \frac{\xi-\nu}{\varrho-1}, \lambda + \frac{\xi-\nu}{\varrho-1}\right]$. A second more accurate search is now done by selecting a new set of points at regular intervals between these two points. The process can be repeated, narrowing the search each time until the maximum is found to sufficient accuracy. It is important on the first search to choose a sufficiently large ϱ otherwise a narrow peak could fall between two points and be overlooked.

Once the maximum or some greater value, $\mathfrak{m}$, is found by one method or another from step 1, step 2 is to sample a value $\mathfrak{x}$ from Uniform(ν, ξ). Step 3 is to sample a value $\mathfrak{y}$ from Uniform$(0, \mathfrak{m})$. These two steps sample a random point uniformly within $[\nu, \xi] \times [0, \mathfrak{m}]$. Step 4 finds $\mathfrak{z}$, the value of the full conditional evaluated at $\mathfrak{x}$. At step 5, if $\mathfrak{y} > \mathfrak{z}$ then return to step 2 and choose another value for $\mathfrak{x}$, otherwise $\mathfrak{x}$ becomes the value drawn from the full conditional. These last two steps determine whether the point drawn uniformly from the $(\xi - \nu) \times \mathfrak{m}$ rectangle is above or below the full conditional. If it is below the line, it is accepted. If it is above,

it is rejected and a new point is drawn and tested. It is this rejection process that gives the algorithm its name. It can be seen that one of the advantages of rejection sampling is that it does not need the area under the full conditional function curve to integrate to 1. It can be used on unnormalised distributions without any additional difficulty.

2.5.2 The "Shawlands" Rejection Sampling Algorithm

Often, simple rejection sampling will be sufficient to be used without modification. However, there are some situations where some modification could improve efficiency. One such situation is where the full conditional is suspected to form a single very sharp peak (figure 2.1). Simple rejection sampling could take a long time in this situation. An $\mathfrak{x}$ is far more likely to be sampled that is not at or near such a peak and will almost certainly be rejected. This could result in a great many rejections occuring before an $\mathfrak{x}$ is selected at or near the peak. In practical terms, a computer program using this algorithm could appear to be doing nothing, potentially for many hours, until an $\mathfrak{x}$ is accepted.

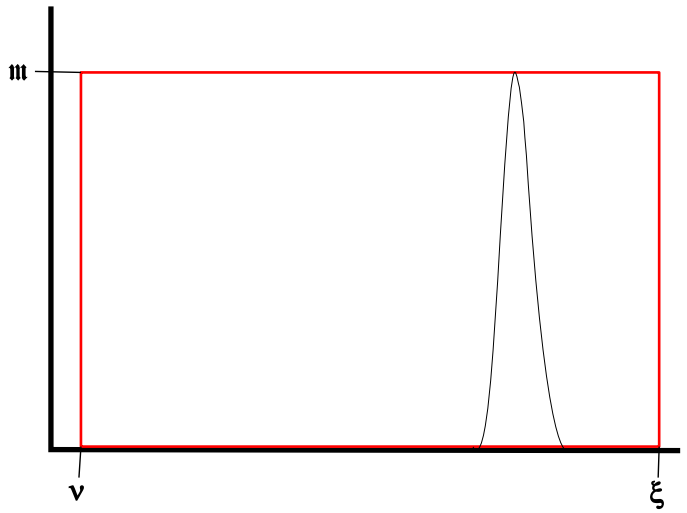

Figure 2.1: Simple Rejection Sampling
In simple rejection sampling a random point is chosen within the red box. If it is below the curve of the full conditional distribution, such as the one shown in black, it is accepted. If not, it is rejected and another point selected. Where the full conditional forms a sharp peak, such as in this situation, it could take many rejections before a point is selected. The problem would be worse if the peak was even sharper than that shown.

This problem can be alleviated by adopting a modified version of rejection sampling. It employs the ϱ points used to find an approximate maximum in step 1 of the description of the simple version of the algorithm to slice the full conditional distribution up into $\varrho - 1$ slices. Step 1 chooses a number, ϱ of points at regular intervals, $\nu, \nu + \frac{\xi-\nu}{\varrho-1}, \nu + \frac{2(\xi-\nu)}{\varrho-1}, \ldots, \nu + \frac{(\varrho-2)(\xi-\nu)}{\varrho-1}, \xi$, evaluates the full conditional at these points, and, as before, finds the value λ, producing the highest value for the full conditional. Again, it is then assumed that the maximum lies somewhere in the interval $\left[\lambda - \frac{\xi-\nu}{\varrho-1}, \lambda + \frac{\xi-\nu}{\varrho-1}\right]$. A second more accurate search is now done to find the universal maximum by selecting new set of points at regular intervals within this interval. Third and fourth searches or the bisection method can again be used for increased accuracy. (If the full conditional is suspected of having several local maxima, second searches of this sort can also be done where the full conditional evaluated at one of the ϱ points, λ_j is greater than it is for the two points adjacent to it, $\lambda_j - \frac{\xi-\nu}{\varrho-1}$ and $\lambda_j + \frac{\xi-\nu}{\varrho-1}$ to find more accurate values of these local maxima.)

Step 2 slices the full conditional up into $\varrho - 1$ slices each of width $\frac{\xi-\nu}{\varrho-1}$. The

information from step 1 is used to find the maximum within each slice. Where a slice does not have a local maximum, found by the second more accurate search in step 1, the maximum for the jth slice $\mathfrak{m}_\mathfrak{j}$ is assumed to be the higher of the two values of the full conditional found at the slice's boundaries because the full conditional is assumed to be strictly increasing or decreasing over the interval of that slice. Otherwise $\mathfrak{m}_\mathfrak{j}$ is the maximum value of the full conditional that was found within the slice during the more accurate search for a maximum at step 1.

In step 3, the area of each slice is calculated $A_\mathfrak{j} = \frac{\xi-\nu}{\varrho-1}\mathfrak{m}_\mathfrak{j}$. Step 4 calculates the total area enclosed by all the slices, $A_{tot} = \sum_{\mathfrak{j}=1}^{\varrho-1} A_\mathfrak{j}$ and the proportion of the total area that each slice accounts for, $\frac{A_\mathfrak{j}}{A_{tot}}$. These proportions sum to one and so can be taken to be probabilities in a discrete probability distribution. Step 4 uses that discrete probability distribution to randomly select a slice with probability proportional to its area.

The next steps are similar to performing simple rejection sampling within the selected slice. Suppose that the jth slice has been selected at step 4. It is bounded by $\lambda_\mathfrak{j}$ and $\lambda_\mathfrak{j} + \frac{\xi-\nu}{\varrho-1}$ on the x axis and by 0 and $\mathfrak{m}_j$ on the probability axis. Step 5 samples a value, $\mathfrak{x}$, from $\text{Uniform}\left(\lambda_\mathfrak{j}, \lambda_\mathfrak{j} + \frac{\xi-\nu}{\varrho-1}\right)$. Step 6 samples a value $\mathfrak{y}$ from $\text{Uniform}(0, \mathfrak{m}_\mathfrak{j})$, so steps 5 and 6 sample a random point uniformly within the slice. Just as in the simple case, step 7 finds $\mathfrak{z}$, the value of the full conditional evaluated at $\mathfrak{x}$. However, at step 8, if $\mathfrak{y} > \mathfrak{z}$ then the algorithm rejects $\mathfrak{x}$, returns to step 4 and uses the discrete probability distribution to select a slice again. Otherwise $\mathfrak{x}$ becomes the value drawn from the full conditional.

In effect what this does is approximate the full conditional function with blocks of width $\frac{\xi-\nu}{\varrho-1}$ before doing rejection sampling. This modification makes it much more likely that the first $\mathfrak{x}$ chosen will not be rejected and that fewer rejections will be needed before an $\mathfrak{x}$ is selected compared with simple rejection sampling. This is particularly useful where the full conditional function is anticipated to contain a sharp peak. However, there are more steps involved, so where such a sharp peak

is not likely, it will be less efficient. There is also a trade off to be made in the choice of ϱ. The larger ϱ is, the more numerical calculations are involved but the closer the blocks can approximate the full conditional and so the thinner the peak it can deal with without a lot of rejections (figure 2.2). Whether the choice of a larger ϱ makes the algorithm faster or slower depends on how sharp the peak is in the full conditional. There is a judgement to be made based on the user's belief of how likely that situation is to arise.

This scheme was devised as a solution to a problem that arose during work on this thesis. During simple rejection sampling, the computer occasionally appeared to freeze or slow down dramatically on that task and yet was performing other tasks normally. On further investigation it was found that the problem was that the sampler was attempting to sample from a function with a single very sharp peak. The idea of approximating the area under a function with rectangles is far from being a new one and can trace its history all the way back to Leibniz's idea for calculus (Leibniz, 1684) who in turn drew inspiriation from Cavalieri's idea for approximating the area under a curve by adding up the lengths of evenly spaced parallel lines drawn below it (Cavalieri, 1635). Here it has been employed to avoid the problem of having a large number of rejected values for $\mathfrak{x}$ and so a lot of wasted processor time. While it would be surprising if this idea has not been used before, nothing exactly the same has been uncovered by a search. This may be because it is mainly useful in the specific situation of sampling from a function with a single very sharp peak. It has since been suggested that this quick-fix may be an accidental innovation and is in need of a name. The idea for this scheme suggested itself during a walk through the Shawlands area of Glasgow where, as with many other areas of Glasgow, there were a cluster of tower blocks dominating the skyline. The tower blocks looked like the rectangles in figure 2.2, providing the inspiration for the solution to the problem. It is therefore suggested that the sampling scheme described in this subsection could be named after the area of Glasgow where the idea occurred.

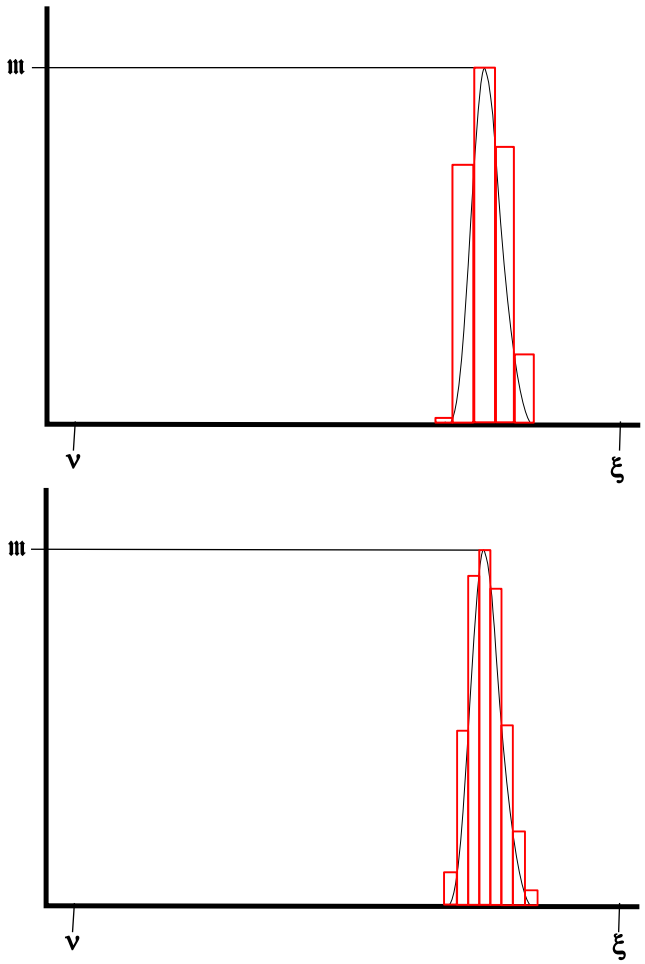

Figure 2.2: "Shawlands" Rejection Sampling
In the "Shawlands" rejection sampling scheme described in this section, one of ϱ blocks of width $\frac{\xi-\nu}{\varrho-1}$ and height equal to the maximum value that the full condtional takes over their width, is chosen with a probability proportional to its area. A point is then randomly chosen within that block. If it is below the density of the full conditional distribution, such as the one shown in black, it is accepted. If not, it is rejected and the process restarts by randomly choosing a block again. In situations where the full conditional forms a sharp peak, such as in this situation, this process could be faster than simple rejection sampling because fewer rejections would be expected before a point is accepted. The blocks form an approximation to the area under the full conditional distribution. The lower diagram has a larger value of ϱ, the number of blocks. In that case, the blocks form a closer approximation to the area under the curve, fewer rejections would be expected before a point is accepted reducing the expected time the algorithm takes to run. However the increased number of blocks increases the number of calculations that need to be done which increases the expected running time.

To show how much time can be saved by this scheme, a number of tests were carried out using a beta distribution, with both parameters greater than 1, as the distribution being sampled from by rejection sampling. Obviously, there are far more efficient ways to sample from a beta distribution but it is being used here

as an example of a function with a single sharp peak. The peak becomes sharper, the larger the parameters are. To show relative differences in running times, the sampling method described in both this and the previous subsection were implemented in R and the times taken to successfully draw a sample of 100,000 values are recorded. Six pairs of parameters and three choices of ϱ are tested. As seen in table 2.1 , where the fastest times taken for each distribution are shown in bold, where the distribution has a wide peak, such as in Beta(2,3), the extra calculations required by this scheme are not worthwhile and simple rejection sampling is faster. However for sharper peaks, such as for Beta(20,30) and beyond, the Shawlands method is faster and the difference in speed becomes more appreciable as the distribution becomes more sharply peaked. In addition, the best value of ϱ becomes larger for sharper peaks, as expected.

Table 2.1: Rejection Sampler Times Taken (in seconds) to Draw a Sample of Size 100,000

	Rejection Sampling Method			
	Simple	Shawlands		
distribution		$\varrho = 10$	$\varrho = 100$	$\varrho = 1000$
Beta(2,3)	**1.83**	2.26	2.23	3.67
Beta(20,30)	5.01	2.75	**2.28**	3.65
Beta(200,300)	15.04	5.90	**2.50**	7.41
Beta(2000,3000)	46.83	17.32	**3.17**	7.00
Beta(20000,30000)	146.58	53.74	6.54	**5.13**
Beta(200000,300000)	462.10	169.76	19.52	**5.34**

For some extra insight into why the Shawlands method is so fast in this case, the top histogram in figure 2.3 summarizes the number of attempts needed to make a successful sampling for the sample of 100,000 for the Beta(200000,300000) distribution for the Shawland sampler with $\varrho = 1000$. In well over 80% of cases the sample was made first time. It rarely took the sampling method more than 6 attempts to achieve a point under the distribution curve. In contrast, for simple rejection sampling the equivalent histogram is shown at the bottom of figure 2.3. It found a point below the distribution curve at the first attempt less than 0.4% of the time . It was common for it to take more than 1000 attempts to find a point

below the curve and could take it many thousands of attempts.

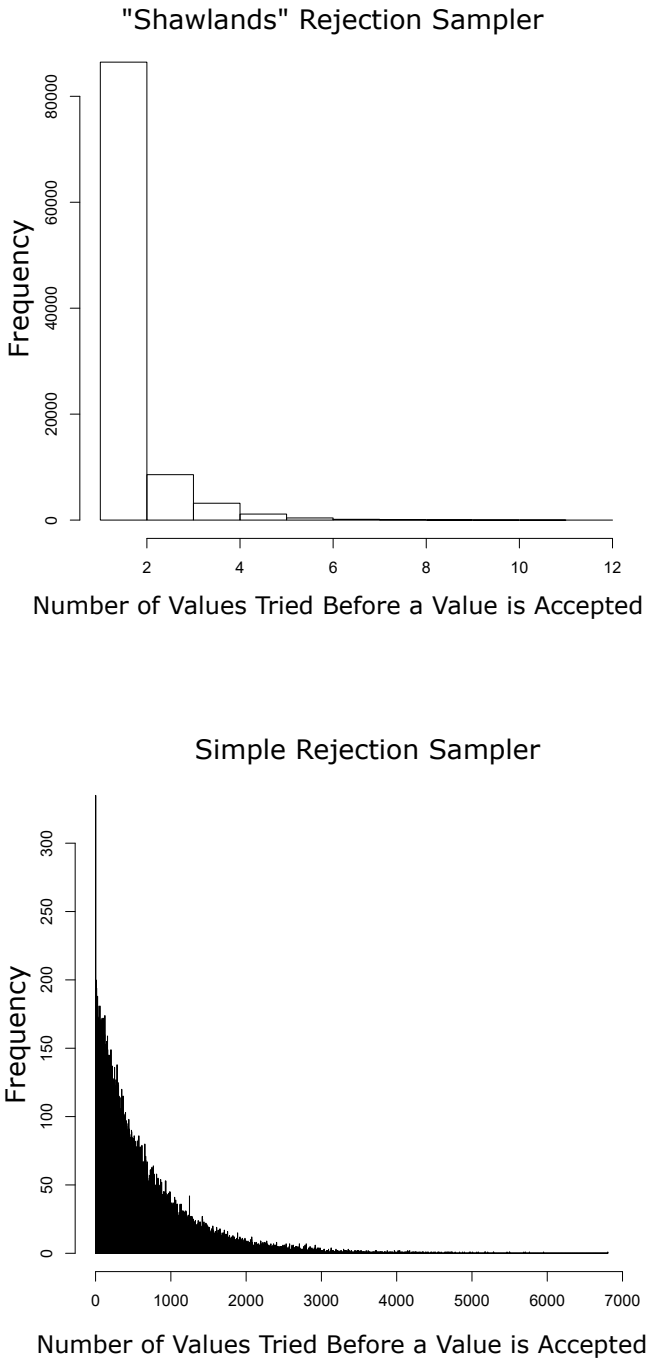

Figure 2.3: Histograms for the Number of Attempts Needed Before a Valid Rejection Sample for Beta(200000,300000) for (top) Shawlands Rejection Sampling and (bottom) Simple Rejection Sampling

This is all very well in the case where many of the calculations needed for Shawlands rejection sampling are needed anyway to find the maximum for simple rejec-

tion sampling. The Shawlands method then doesn't need many extra calculations over simple rejection sampling. But what about the case where the maximum can be found analytically? In this case, during a practical situation such as Gibbs sampling, the Shawlands method would need to evaluate the function at many points before each draw since in a Gibbs sampler each draw could be from a different distribution whereas simple rejection sampling would not need to do this if an analytical maximum were known. Could there be cases where Shawlands rejection sampling is faster even in this situation? This was tested by requiring the Shawlands sampler to repeat the calculations for the ϱ points on the distribution again before each of the 100,000 samples from the distribution were taken while allowing the simple rejection sampler to use the known maximum without needing any such calculations. In this case, the only thing slowing down the simple rejection sampler is the number of attempts it needs to make before selecting a point below the function curve. Even in this situation, which is particularly disadvantageous for the Shawlands sampling method, there are situations where it is still faster if very sharply peaked distributions are being regularly encountered as table 2.2 shows.

Table 2.2: Rejection Sampler Times Taken to Draw from 100,000 Distributions

	Rejection Sampling Method			
	Simple	Shawlands		
distribution	(analytical maximum)	$\varrho = 10$	$\varrho = 100$	$\varrho = 1000$
Beta(200000,300000)	**464.91**	474.76	475.94	1812.61
Beta(2000000,3000000)	1459.16	834.94	**493.33**	1802.26

Here the simple rejection sampler takes a similar amount of time for the Beta(200000,300000) distribution even when the maximum is known analytically. This is because the time it takes is dominated by the amount of time it takes to sample a point under the curve, which by nature is random. The extra calculations needed for each sample for the Shawlands sampler results in the whole process taking a similar amount of time to the simple rejection sampler for $\varrho = 10$ and $\varrho = 100$. However, if the distributions are even more extremely sharply peaked such as for Beta(2000000,3000000), the Shawlands sampler becomes quicker again

even with having to make so many extra calculations before each sample. Here, of the three values of ϱ tested $\varrho = 100$ was fastest but there will be an optimum value of ϱ somewhere between 10 and 1000.

In conclusion, the Shawlands sampler is useful in specific situations where a very sharply-peaked distribution is being sampled from, particularly if the maximum of the distribution function cannot be found analytically or in a reliable way by some fast method.

2.6 Saitou and Nei's Neighbour Joining Algorithm

In the first chapter, population trees were introduced to describe the relationship between present-day subpopulations and their common ancestors. But which tree structure out of the many possible structures should be chosen to represent the genetic relationships between the subpopulations?

If the tree is a bifurcating one then one way to reconstruct the tree is to use the Neighbour Joining (NJ) method described by Saitou and Nei (1987) and Studier and Keppler (1988).

The algorithm can be summarised in six steps.

Step 1 Make a distance matrix where the entries in the $\mathcal{A}$th row and $\mathcal{B}$th column represent some measure of distance between subpopulations $\mathcal{A}$ and $\mathcal{B}$. This will be a symmetric matrix with 0s along its main diagonal. In this thesis the estimated distances were obtained by making pairwise F_{ST} estimates from the data using the equation

$$F_{ST} = \frac{1}{2L}\sum_{i=1}^{L}\frac{\left(\hat{\alpha}_{i,\mathcal{A}} - \hat{\alpha}_{i,\mathcal{B}}\right)^2}{\overline{\alpha}_{i,\mathcal{AB}}\left(1 - \overline{\alpha}_{i,\mathcal{AB}}\right)} \tag{2.5}$$

where $\hat{\alpha}_{i,j} = \frac{x_{i,j}}{n_{i,j}}$ for locus i and $j = \mathcal{A}$ or $\mathcal{B}$, representing the two subpopulations in the pairwise estimate and $\overline{\alpha}_{i,\mathcal{AB}} = \frac{x_{i,\mathcal{A}}+x_{i,\mathcal{B}}}{n_{i,\mathcal{A}}+n_{i,\mathcal{B}}}$. $x_{i,j}$ is the allele counts for one

of two variants for subpopulation j at locus i. $n_{i,j}$ is the sample size (twice the number of subjects for a diploid species) for subpopulation j at locus i.

This produces a $\mathcal{N} \times \mathcal{N}$ symmetric distance matrix

$$\begin{pmatrix} 0 & \mathcal{D}_{12} & \mathcal{D}_{13} & \cdots & \mathcal{D}_{1\mathcal{N}} \\ \mathcal{D}_{21} & 0 & \mathcal{D}_{23} & \cdots & \mathcal{D}_{2\mathcal{N}} \\ \mathcal{D}_{31} & \mathcal{D}_{32} & 0 & \cdots & \mathcal{D}_{3\mathcal{N}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{D}_{\mathcal{N}1} & \mathcal{D}_{\mathcal{N}2} & \mathcal{D}_{\mathcal{N}3} & \cdots & 0 \end{pmatrix}, \tag{2.6}$$

where there are N subpopulations and $\mathcal{D}_{\mathcal{AB}} = \mathcal{D}_{\mathcal{BA}}$ is the F_{ST} calculated from (2.5) for subpopulations $\mathcal{A}$ and $\mathcal{B}$.

Step 2 For each unjoined subpopulation, $\mathcal{A}$, out of the $\mathfrak{t}$ remaining, compute

$$\mathfrak{u}_{\mathcal{A}} = \sum_{\mathcal{B}=1}^{\mathfrak{t}} \frac{\mathcal{D}_{\mathcal{AB}}}{\mathfrak{t}-2}. \tag{2.7}$$

Step 3 Choose unjoined subpopulations, $\mathcal{A}$ and $\mathcal{B}$ for which $\mathcal{D}_{\mathcal{AB}} - \mathfrak{u}_{\mathcal{A}} - \mathfrak{u}_{\mathcal{B}}$ is the smallest.

Step 4 Subpopulations $\mathcal{A}$ and $\mathcal{B}$ are neighbours, so draw branches joining these subpopulations to a new common node which represents the common ancestral population from which they are both descended.

A branch length can be calculated, if needed, from subpopulation $\mathcal{A}$ to the new node and is

$$\mathfrak{v}_{\mathcal{A}} = \frac{\mathcal{D}_{\mathcal{AB}} + \mathfrak{u}_{\mathcal{A}} - \mathfrak{u}_{\mathcal{B}}}{2}, \tag{2.8}$$

and from subpopulation $\mathcal{B}$ to the new node is

$$\mathfrak{v}_{\mathcal{B}} = \frac{\mathcal{D}_{\mathcal{AB}} + \mathfrak{u}_{\mathcal{B}} - \mathfrak{u}_{\mathcal{A}}}{2}. \tag{2.9}$$

Step 5 The new node represents the ancestral population of subpopulations $\mathcal{A}$ and $\mathcal{B}$. The entries for subpopulations $\mathcal{A}$ and $\mathcal{B}$ will be removed from the distance matrix and replaced with one row and column for the new ancestral population. The distance entries in the matrix from this new node $\mathcal{S}$ to each of the other remaining subpopulations is calculated from

$$\mathcal{D}_{\mathcal{ST}} = \frac{\mathcal{D}_{\mathcal{AT}} + \mathcal{D}_{\mathcal{BT}} - \mathcal{D}_{\mathcal{AB}}}{2}, \tag{2.10}$$

for subpopulation $\mathcal{T}$.

Step 6 If the distance matrix is now a 3×3 matrix then stop, otherwise return to step one, treating ancestral subpopulations in the same way as the original subpopulations. This will find the next two subpopulations to join together at a new node and so on.

2.7 Gelman's R Statistic

One way to test for lack of convergence in an MCMC sampler is to use Gelman's R statistic. This is described in Chapter 8 of Gilks and Richardson (1996) and in Gelman et al. (2013). It consists of running $\mathfrak{d}$ parallel chains of $\mathfrak{n}$ iterations each so that, in this context, $\mathfrak{I} = 1, ..., \mathfrak{d}$ and $\mathfrak{J} = 1, ..., \mathfrak{n}$ so that $\mathfrak{I}$ labels the chain and $\mathfrak{J}$ labels the iteration in each chain. Ideally, each of the $\mathfrak{d}$ chains has a different initial state. Let $\mathfrak{B}$ be the between-chain variance,

$$\mathfrak{B} = \frac{\mathfrak{n}}{\mathfrak{d} - 1} \sum_{\mathfrak{I}=1}^{\mathfrak{d}} \left(\bar{\mathfrak{Y}}_{\mathfrak{I}} - \bar{\mathfrak{Y}} \right)^2, \tag{2.11}$$

where

$$\bar{\mathfrak{Y}}_{\mathfrak{I}} = \frac{1}{\mathfrak{n}} \sum_{\mathfrak{J}=1}^{\mathfrak{n}} \mathfrak{Y}_{\mathfrak{I}\mathfrak{J}} \tag{2.12}$$

is the mean of the $\mathfrak{I}$th chain, and

$$\bar{\mathfrak{Y}} = \frac{1}{\mathfrak{d}} \sum_{\mathfrak{I}=1}^{\mathfrak{d}} \bar{\mathfrak{Y}}_{\mathfrak{I}} \tag{2.13}$$

is the grand mean over all $\mathfrak{d}$ chains. The within-chain variance is

$$\mathfrak{W} = \frac{1}{\mathfrak{d}} \sum_{\mathfrak{I}=1}^{\mathfrak{d}} \mathfrak{s}_{\mathfrak{I}}^2, \tag{2.14}$$

where

$$\mathfrak{s}_{\mathfrak{I}}^2 = \frac{1}{\mathfrak{n}-1} \sum_{\mathfrak{J}=1}^{\mathfrak{n}} \left(\mathfrak{Y}_{\mathfrak{I}\mathfrak{J}} - \bar{\mathfrak{Y}}_{\mathfrak{I}}\right)^2 \tag{2.15}$$

is the within-chain variance for the ith chain. Gelman's R statistic is then

$$\hat{\mathfrak{R}} = \sqrt{\frac{\mathfrak{n}-1}{\mathfrak{n}} + \frac{\mathfrak{B}}{\mathfrak{n}\mathfrak{W}}}. \tag{2.16}$$

It can be seen that $\hat{\mathfrak{R}}$ is determined by the ratio of $\mathfrak{B}$, the between-chain variance and $\mathfrak{W}$, the within-chain variance. If the sequence has converged, then these two measures of variance should be about equal because the $\mathfrak{d}$ chains should be indistinguishable from each other. As a result, their ratio should be near 1. According to Gilks and Richardson (1996), if $\hat{\mathfrak{R}}$ is above 1.1 – 1.2 then the statistic provides evidence that the sequence has not converged. Unfortunately, there is no way of proving conclusively that the Markov chain process has converged; the best that can be done is to say that there is no evidence that it has not converged.

2.8 WAIC

When there are two or more candidate models of the data, the question of which of these models is "best" arises. It is advantageous for a model to describe the data as well as possible. However relying solely on such a criterion would give an inherent advantage to more complex models. Models with more estimated

parameters will have an automatic advantage in describing the data. Indeed a model with a sufficient number of parameters could fit the data exactly. Many of these additional parameters may not contribute usefully to an explanation of the actual process that gave rise to the data and may just be describing noise. Typically such models then have poor predictive properties. Additionally, Occam's Razor, holds that the simplest explanation for an event is the most likely to be true (Collins, 2017). While it can be counter-argued that the real world processes that gave rise to the data are, in reality, highly complex, by including such unhelpful spurious additional complexity, the main features of the process that are of interest become obscured. For this reason, models with fewer parameters that are almost as good at describing the data are preferred to more complex ones so there must be some penalty for models with greater numbers of parameters. So a measure of how good a model is should incorporate terms that measure how well the model describes the data and that also penalise complexity. This subsection draws on pages 166 to 178 of Gelman et al. (2013).

Ideally, in a Bayesian context it would be useful to know how likely each model was given the data, $p(M|D)$, where M is the model and D is the data. Using Bayes' rule this becomes

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}, \tag{2.17}$$

where $p(M)$ represents the prior belief in the truth of the model. Penny et al. (2006) state that in Bayesian model selection, the model is chosen which has the highest probability $p(M|D)$. Where there are two possible models, $M = 1$ and $M = 2$, equation 2.17 becomes

$$p(M = 1|D) = \frac{p(D|M = 1)p(M = 1)}{p(D|M = 1)p(M = 1) + p(D|M = 2)p(M = 2)}, \tag{2.18}$$

for $M = 1$ and

$$p(M = 2|D) = \frac{p(D|M = 2)p(M = 2)}{p(D|M = 1)p(M = 1) + p(D|M = 2)p(M = 2)}, \tag{2.19}$$

for $M = 2$.

In this case, if the prior odds ratio is defined as $\frac{p(M=1)}{p(M=2)}$, and the posterior odds ratio as $\frac{p(M=1|D)}{p(M=2|D)}$, then these can be related by a Bayes factor, $\mathfrak{F}$,

$$\frac{p(M = 1|D)}{p(M = 2|D)} = \mathfrak{F}\frac{p(M = 1)}{p(M = 2)}. \tag{2.20}$$

where

$$\mathfrak{F} = \frac{p(D|M = 1)}{p(D|M = 2)} \tag{2.21}$$

In practice, however, it can be very difficult to calculate $p(D|M)$, particularly for complex models. A number of other methods have been devised to aid model selection based around the model parameters θ and their estimates $\hat{\theta}$.

A measure that is commonly used is the Akaike Information Criterion (AIC) (Akaike, 1973). This has a relatively simple formula which shows how the ideas in information criteria measures work. In these criteria, D represents the data and θ represents the parameters. AIC is defined as

$$\text{AIC} = -2\ln\left(p(D|\hat{\theta})\right) + 2\mathcal{C}. \tag{2.22}$$

In this formula, the first term depends on the probability (density) of the data given the estimated parameters. In the case of AIC, the latter are maximum likelihood estimates. The log of the probability of the data given the parameters or log-likelihood is also known as the log predictive density. Models for the data with a high probability have a low value for this first term and those with a lower probability have a greater value. In the second term $\mathcal{C}$ represents the number of parameters, so the second term gives a higher score to models with more param-

eters. Since the model with the lower AIC score is preferred, this represents a penalty for having a more complex model.

AIC is not appropriate in the case of the Bayesian hierarchical models that are dealt with in most of this thesis. Simply penalising by the number of parameters is not appropriate. For example, in hierarchical models it is sometimes possible to integrate out some of the intermediate parameters. The overall model remains the same, yet AIC would penalise the model with the parameters integrated out, and hence with fewer parameters as a result, less harshly than the equivalent model with those parameters remaining. Additionally, parameters with more informative priors have less freedom to change to fit the data and so contribute less to the overfitting problem. Until the last few years, the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) has been a commonly used information criterion for Bayesian hierarchical models. DIC is defined by

$$\mathrm{DIC} = -2\ln\left(p(D|\hat{\theta})\right) + 2\mathcal{V}_{\mathrm{DIC}}. \tag{2.23}$$

Here the parameter estimates are not the maximum likelihood estimates, but the mean of the posterior distribution of each parameter. $\mathcal{V}_{\mathrm{DIC}}$ is called the effective number of parameters and is analogous to $\mathcal{C}$ in (2.22). $\mathcal{V}_{\mathrm{DIC}}$ is calculated from

$$\mathcal{V}_{\mathrm{DIC}} = 2\left[\ln\left(p\left[D|\hat{\theta}\right]\right) - \frac{1}{\mathcal{Y}}\sum_{\mathcal{W}=1}^{\mathcal{Y}}\ln\left(p\left[D|\theta_{\mathcal{W}}\right]\right)\right], \tag{2.24}$$

where the posterior distribution of the parameters has been approximated by $\mathcal{Y}$ draws from it such as would be obtained from an MCMC sampler, after discarding burn-in. $\theta_{\mathcal{W}}$ is the state of the parameter set at iteration $\mathcal{W}$ of the process. The second term is (an estimate of) the posterior expectation of the log-likelihood. It can be seen that if the parameters are not free to move far, such as might be the case when they have a very informative prior, then the two terms in $\mathcal{V}_{DIC}$ will be close to each other and the effective number of parameters small. If they are free to cover a wider range, the second term will include iterations with a $\theta_{\mathcal{W}}$ that leads

to relatively small likelihoods and hence to larger values of $\mathcal{V}_{DIC}$, representing a larger penalty in the DIC for a greater number of effective parameters.

Another popular information criterion is the Bayesian Information Criterion (BIC). The formula for BIC is

$$\text{BIC} = -2 \ln \left(p(D|\hat{\theta}) \right) + \mathcal{C} \ln \Xi, \tag{2.25}$$

where Ξ is the number of pieces of independent data. The relationship to AIC is obvious. The first term is identical and rewards accuracy. The second term penalises the number of parameters, the penalty increasing with the size of the data set. Where $\Xi > e^2$, this penalty will be greater than that in AIC. Gelman et al. (2013) does not consider it useful as a predictor of model performance. However, sufficient others do find it useful and keep it in common use. It does have the attraction of being related to marginal likelihood, $p\left(D|M\right)$, under certain assumptions such as large Ξ , $\Xi \gg \mathcal{C}$ and the priors $p\left(\theta|M\right)$ being relatively linear near $\hat{\theta}$,

$$p\left(D|M\right) \approx \exp \left[-\frac{BIC}{2} + O\left(\Xi^0\right) \right], \tag{2.26}$$

where $O\left(\Xi^0\right)$ are terms of order Ξ^0.

More recently, the Watanabe Akaike Information Criterion (WAIC) has emerged (Watanabe, 2010). It makes greater use of the posterior distribution of θ. One of the problems with DIC is its use of point estimates of θ based on the posterior distribution. In the case of a multimodal or a unimodal but highly skewed posterior distribution, situations can arise where the posterior mean value is not very typical of the posterior distribution as a whole. It could sit near a deep minimum between two modes of a posterior distribution or in the tail of a very skewed distribution. The use of point estimates, as in DIC, is not in keeping with the spirit of the Bayesian approach.

WAIC is defined by

$$\text{WAIC} = -2 \sum_{\psi=1}^{\Xi} \ln \left[\frac{1}{\mathcal{Y}} \sum_{\mathcal{W}=1}^{\mathcal{Y}} p\left(D_\psi | \theta_\mathcal{W}\right) \right] + 2\mathcal{V}_{\text{WAIC}}. \tag{2.27}$$

Bearing in mind that likelihood, $p(D|\theta) = \prod_{\psi=1}^{\Xi} p(D_\psi|\theta)$, when there are Ξ pieces of independent data, the log-likelihood is $\ln\left(p\left[D|\theta\right]\right) = \sum_{\psi=1}^{\Xi} \ln\left(p\left[D_\psi|\theta\right]\right)$. Then $\frac{1}{\mathcal{Y}} \sum_{\mathcal{W}=1}^{\mathcal{Y}} p\left(D_\psi|\theta_\mathcal{W}\right)$ is an estimate of the posterior mean of $p\left(D_\psi|\theta\right)$, replacing $p(D|\hat{\theta})$ in DIC. Similarly, $\mathcal{V}_{\text{WAIC}}$ takes an expectation over the posterior distribution of θ rather than uses a point estimate. In effect, $\mathcal{V}_{\text{WAIC}}$ represents an estimate of the effective number of parameters in just the same way as $\mathcal{V}_{\text{DIC}}$ does:

$$\mathcal{V}_{\text{WAIC}} = 2 \sum_{\psi=1}^{\Xi} \left[\ln\left(\frac{1}{\mathcal{Y}} \sum_{\mathcal{W}=1}^{\mathcal{Y}} p\left(D_\psi|\theta_\mathcal{W}\right) \right) - \frac{1}{\mathcal{Y}} \sum_{\mathcal{W}=1}^{\mathcal{Y}} \ln\left(p\left[D_\psi|\theta_\mathcal{W}\right]\right) \right]. \tag{2.28}$$

This formula may superficially look more complicated but is simpler to use in practice. The draws from the posterior distribution of θ are readily available as a result of the MCMC sampling process. Even with large datasets and complex models, WAIC can be computed quite readily. As with the other information criteria, models with lower values of WAIC are preferred.

WAIC, is particularly well suited to being used to compare Bayesian hierarchical models. WAIC is also known as the Widely Applicable Information Criterion but Watanabe's name has become attached to it because of their published work on the subject e.g., Watanabe (2010). Alternatives, such as K-fold cross-validation would be computationally more time consuming according to Vehtari and Gelman (2014).

So should the candidate model with the lowest WAIC always be selected? Mechanistically choosing the model with the lowest WAIC without referring back to the real-world problem or process that the model is intended to represent could

lead to a model being selected that describes the data well but does not make sense in relation to the real-world process. Inevitably some human judgement is required to ensure the selected model makes sense in the context of the process it is intended to represent. In a Bayesian context, this is reflecting the fact that different models are a-priori more or less probable. WAIC only selects a model which balances describing the data well with complexity. It makes no judgement about whether the model is sensible or believable or not. It may be that, out of one or more models that have similar WAIC, there is a cogent argument to be made for selecting one that does not have the smallest WAIC if the parameters included or posterior parameter distributions are easier to explain in relation to the aspect of the real world that it was intended to describe.

2.9 Post Predictive Checking

Another tool that can be used to assist in considering how well a model represents the data is post predictive checking. Here pages 143-159 of Gelman et al. (2013) give more detail. The idea of post predictive checking is relatively simple but very effective. A model and its parameter estimates can be used to generate a simulated data set. In a Bayesian context, this set of parameter estimates can be obtained from one draw from their joint posterior distribution. This can be repeated, making a new draw from the posterior distribution each time, to generate a large number of such simulated data sets. The idea of post predictive checking is that if the real world data set were shuffled in among these simulated data sets and if the model was a good description of that data, then the real world data set would not look unusual compared to the simulated data sets. But exactly how should the real world data set be compared to the simulated data sets? To make the comparison, some quality of the data set needs to be expressed as a single number. The choice of that quality depends on what aspects of the data set it is considered important to capture in the model. It could be anything. It could be one of the traditional

measures of location or spread of the data or anything else that is of interest. Once a measure has been chosen, that measure can be calculated for each of the simulated data sets and for the real world data set. The value for the real world data set can then be compared to the values for the simulated data sets to see if it is unusual in any way. "Unusual" could mean in the tails of the distribution of simulated values, and this can be captured by a tail probability (analogous to a p-value). If D is the real data, D_{sim} is a simulated data set, and $T(D)$ is the function used to calculate the measure of the quality of the data set that is of interest, the posterior predictive p-value is:

$$p = Pr\big(T(D_{sim}) \geq T(D)|D\big). \tag{2.29}$$

With a sufficiently large number of simulated data sets, this can be estimated by calculating the proportion of the simulated data sets for which $T(D_{sim}) \geq T(D)$. Both large (close to 1) and small (close to 0) values of p are usually of interest in this context because they both indicate that the real world data is out of place (in the left and right tails, respectively) among the simulated data sets with respect to the quality being tested. The criterion for judging closeness to 0 or 1 is arbitrary and depends on how important it is to the experimenter that the model represent the aspect of the data being measured by the function T.

More than one such quality may be of interest requiring several such T functions to be evaluated and their associated predictive p-values obtained. Multiple-testing considerations are not important here. It is true that if a lot of such p-values are obtained, some will be extreme by chance. However, the experimenter will consider it more important that the model represent some qualities well and less important that it represents other qualities less well. If no model can represent all the tested qualities well, the experimenter can choose a model that represents the most important ones well.

In the applications that will appear later in this thesis, it will be desirable for the

model to represent the relatedness between present-day subpopulations well. A natural choice to quantify relatedness between two subpopulations is to calculate Wright's pairwise F_{ST} (Wright, 1951).

Chapter 3

Models for Quantifying Genetic Drift

This chapter will develop a model of genetic drift involving all present-day subpopulations arising from a single multifurcation event from one common ancestral population. Three models of genetic drift will be considered. The Wright–Fisher Model which describes genetic drift from one generation to the next will be introduced. Two approaches to approximating it over a larger number of generations will then be considered. The first was developed by Balding and Nichols (Balding and Nichols, 1995) which is based on a beta distribution and another developed by Nicholson and others (Nicholson et al., 2002) is based on a modified Normal distribution. These will then be compared in the context of the simple single multifurcation model.

3.1 The Wright–Fisher Model

3.1.1 Drift of Rare Alleles in the Wright–Fisher Model

The goal of modelling genetic drift is a probability model that captures the salient features of the Wright–Fisher model. This model was first described by Fisher (1930), who was aware of earlier work by Wright that was not published until 1931 (Wright, 1931). This model assumes that the number of instances of a particular allele at a locus at generation $t+1$, a_{t+1} is taken by randomly drawing n times with replacement from the pool of alleles in generation t, so that the distribution of the allele at generation $t+1$ is Binomial with parameter a_t/n, the proportion of the allele at generation t, in a (constant) population of size n, which is twice the number of individuals in a diploid species like humans:

$$Pr(a_{t+1} = x) = \binom{n}{x} \left(\frac{a_t}{n}\right)^x \left(1 - \frac{a_t}{n}\right)^{n-x}. \tag{3.1}$$

The model has some interesting properties. It allows for an allele to become fixed for all time. If at some generation t, either $a_t = 0$ or $a_t = n$, then $a_{t+K} = 0$ or $a_{t+K} = n$, respectively, for all positive K. This makes sense because if we assume a model with no mutation and an allele is not present in the population at generation t, then no individual in a subsequent generation can inherit it. Similarly, if it is the only allele present at a locus at generation t, then all individuals in all subsequent generations must inherit it. Of course, if mutation is common enough, this would be a poor model.

Another property of the model is that if the proportion of an allele, $\alpha_t = a_t/n$, is known at generation t but not at a subsequent time, then the expected proportion of the allele in a subsequent generation is the same as that last known proportion regardless of how many generations into the future the expectation is taken. This

remains true even as $K \to \infty$. That is,

$$\mathrm{E}\left(\alpha_{t+K}|\alpha_t\right) = \alpha_t \text{ for all } K > 0. \tag{3.2}$$

It is sometimes stated wrongly (e.g., by Hartl and Clark 1997) that the proportion of an allele in the Wright–Fisher model is just as likely to increase as decrease from one generation to the next regardless of how common or rare that allele is. For any finite population this is not true; e.g., for all $\alpha_t < 0.5$, there is a higher probability of a decrease than an increase (and a lower probability when $\alpha_t > 0.5$).

For practical reasons, it is not possible to use the Wright–Fisher model directly for inference. Under the model, $Pr\left(\alpha_{t+K}|\alpha_t\right)$ has too complicated a dependence on K. The need here is to model drift over a potentially very large and unknown number of generations so it is impractical to use the Wright–Fisher model. The reason for using the Balding–Nichols drift model or any other model is to approximate the behaviour of the Wright–Fisher model over a large number of generations. From the above discussion, it is desirable for such a model to have similar properties to the Wright–Fisher model. That is,

- the distributions of the proportions of alleles should be similar to those that the Wright–Fisher model would produce over a large number of generations;
- it should allow for an allele to become fixed;
- the expected proportion of an allele under drift should be its last known proportion;
- while it is therefore desirable for the mean of the proportion of an allele after a period of genetic drift to be the last known proportion, it is not necessary for its median to be that last known proportion.

In order to visualise the distributions of proportions of an allele under drift in the Wright–Fisher model, it is helpful to look at a few simulations of that model over

a large number of generations (figure 3.1). The pattern observed here is that, with increasing numbers of generations, the distribution of the proportion of an allele spreads out and eventually collects at the points 0 and 1.

3.2 The Balding–Nichols Model

3.2.1 Drift of Rare Alleles in the Balding–Nichols Model

A beta distribution model of genetic drift was suggested by Balding and Nichols (1995). In that paper it appears in its more general multivariate form as a Dirichlet distribution but in the case of only two variants at a locus, it simplifies to the beta distribution $\alpha_{t+K}|\alpha_t \sim \text{Beta}\left(\frac{\alpha_t(1-c)}{c}, \frac{(1-\alpha_t)(1-c)}{c}\right)$, which has a mean of the starting proportion of the allele α_t and a variance of $c\alpha_t\,(1-\alpha_t)$, where c represents a measure of genetic drift into which, for example, numbers of generations and fluctuations in population size have been abstracted. This distribution has the property that the expected future proportion is the present proportion α_t of the allele. The beta distribution is convenient to work with and can lead to models where the proportion of the allele at the end of the period of drift can be integrated out to produce a Beta-Binomial model for the allele counts. However, it has drawbacks. It does not allow an allele to become fixed, that is it does not allow the proportion of the allele to reach 0 or 1, although it does allow proportions very close to 0 or 1. In practice, however, it can produce proportions that are within machine precision of 0 or 1. In the more complex models that will be considered in subsequent chapters, the proportion of the allele α_{t+K} resulting from one period of drift becomes the starting point for a subsequent period of drift. In those cases, this machine precision issue makes it necessary to prevent either parameter of the beta distribution for the subsequent period of drift becoming 0 (for which the beta distribution is undefined), so artificial barriers just above 0 and just below 1 have to be imposed in that situation.

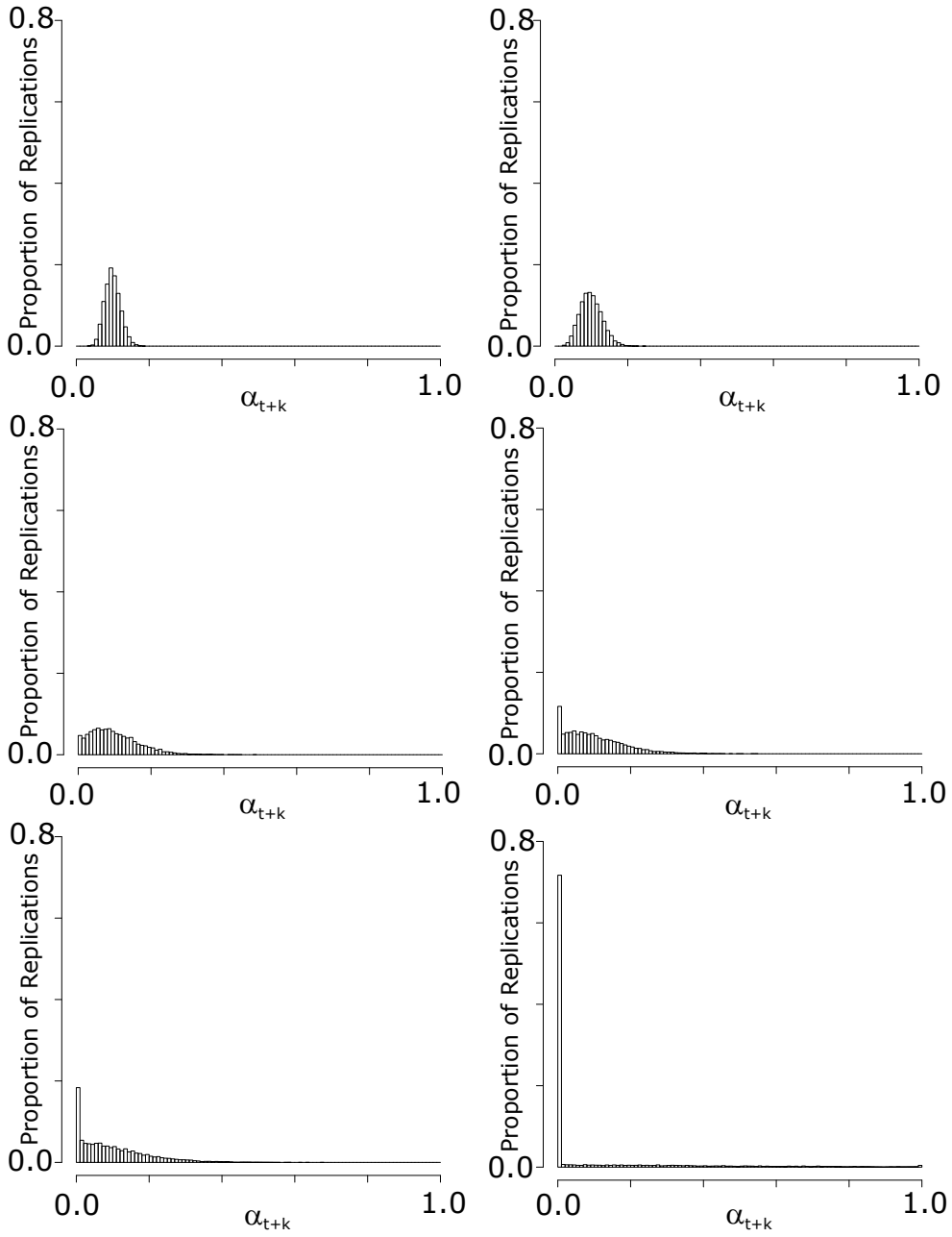

Figure 3.1: Wright–Fisher Model: Distribution of α_{t+k} for Increasing Genetic Drift
This illustrates how the distribution of $\alpha_{t+K}|\alpha_t$ develops with increasing K, the number of additional generations for an initial value of $\alpha_t = 0.1$ and a population size of 1000 for the Wright–Fisher Model. The graphs show simulations of 10,000 replications with values of K of 5,10,51,78,105 and 692 generations. In particular, note how with increasing k, the distribution becomes skewed, the mode shifts left and the probability density collects at atoms at first 0 and eventually 1.

The other drawback is that the shape of the distribution of the proportions of an initially rare allele with increasing drift look somewhat different in shape to those

for the Wright–Fisher model. In particular, the Balding–Nichols distribution is skewed and does not develop the characteristic atom at 0 (or 1), as can be seen by comparing figure 3.1 with figure 3.2. While it does have the property of the expected future proportion of an allele being its present one, the skewness makes it much more likely the next proportion of the allele will be lower than its current one, compared with the Wright–Fisher model, so more likely to be closer to (although never actually reaching) 0. This means that over successive periods of drift, a rare allele will, far more likely than not, go on becoming rarer (more often than the Wright–Fisher model would predict) without ever completely dying out (which the Wright–Fisher model allows). In data simulated under the Balding–Nichols model, a large amount of drift is highly likely to lead to only a small change in the proportion of a rare allele and that change is very likely to be in the direction that makes it rarer.

3.2.2 Implementation of the Balding–Nichols Model

The idea is to apply a variant of the single multifurcation model described by Nicholson et al. (2002) (hereafter called the Nicholson–Donnelly model) to obtain a measure of genetic drift for each subpopulation. This measure of genetic drift is conceptually similar to the F_{ST} measure described by Wright (1951) which is widely used elsewhere, but differs in that it is specific to each subpopulation.

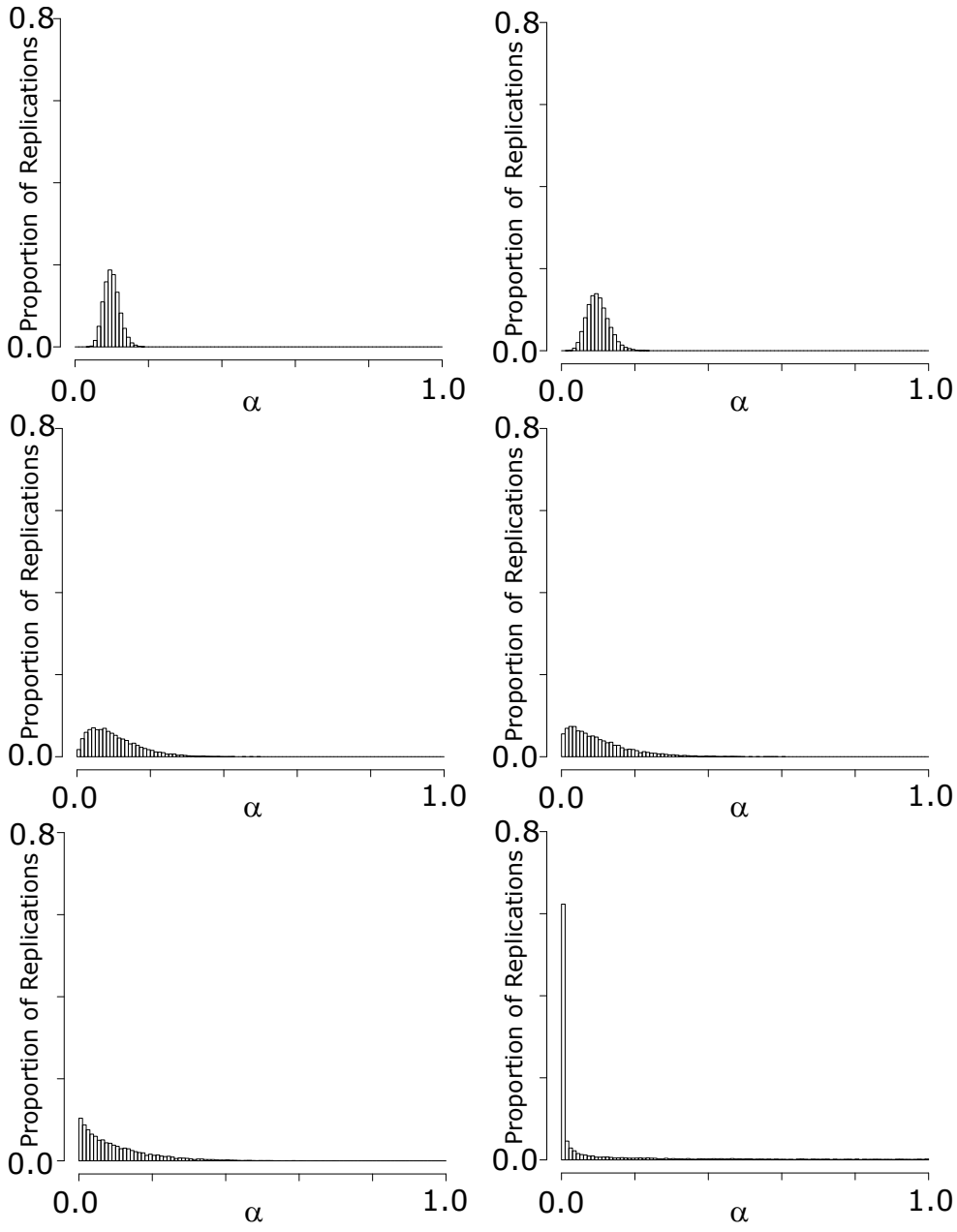

Figure 3.2: Balding–Nichols Model: Distribution of α for Increasing Genetic Drift c *This illustrates how* Beta$\left(\frac{\alpha_t(1-c)}{c}, \frac{(1-\alpha_t)(1-c)}{c}\right)$*, the Balding–Nichols model's approximation to the Wright–Fisher model, develops with increasing c, the parameter for genetic drift. This is shown for an initial value of* $\alpha_t = 0.1$*. Here, note how with increasing c, the mode shifts left but there is a more exaggerated skew than for the Wright–Fisher model and although it appears probability density is collecting first at 0 and eventually at 1, this is only because the histogram has a resolution governed by the bin width. The values are very close to 0 (and 1) but an exact 0 or 1 cannot be drawn from a beta distribution such as this.*

A Directed Acyclic Graph (DAG) of the model used is shown in figure 3.3. Here

$$x_{ij}|n_{ij}, \alpha_{ij} \sim \text{Binomial}\,(n_{ij}, \alpha_{ij}), \text{ independently,}$$

$$\alpha_{ij}|\pi_i, c_j \sim \text{Beta}\left(\tfrac{\pi_i(1-c_j)}{c_j}, \tfrac{(1-\pi_i)(1-c_j)}{c_j}\right), \text{ independently,}$$

with priors

$$\pi_i|a \sim \text{Beta}\,(a, a), \text{ independently,}$$

$$c_j \sim \text{Beta}(b_{1j}, b_{2j}), \text{ independently,}$$

where

i labels the locus: $1 \leqslant i \leqslant L$,

j labels the subpopulation $1 \leqslant j \leqslant P$,

n_{ij} is the total number of alleles observed at locus i in subpopulation j,

x_{ij} is the number of one of the two alleles observed at locus i in subpopulation j,

α_{ij} is the population proportion of that allele at locus i in subpopulation j,

π_i is the proportion of that allele at locus i in the ancestral population,

c_j is the amount of genetic drift in subpopulation j.

a is a hyperparameter in the prior of π_i.

b_{1j}, b_{2j} are hyperparameters in the prior of c_j and assigned the value 1 unless otherwise stated.

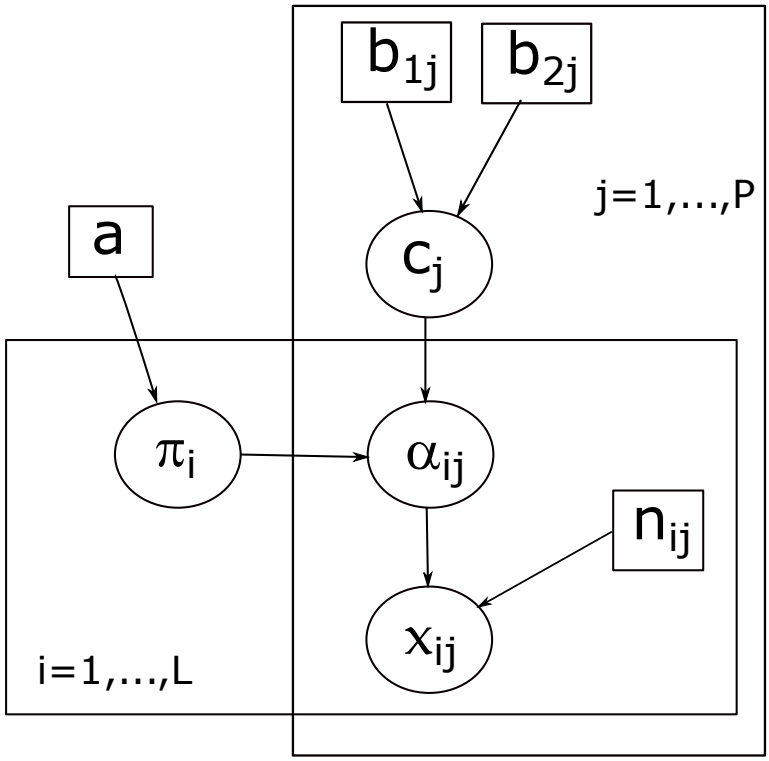

Figure 3.3: DAG of Variant of Nicholson–Donnelly Model

This differs from the model described by Nicholson et al. (2002) in that they used $\alpha_{ij}|\pi_i, c_j \sim N(\pi_i, c_j\pi_i(1-\pi_i))$, with the mass of the distribution below 0 and above 1, atomised at 0 and 1, respectively, to keep the values of α in $[0,1]$. That model will be considered in the next section. The beta distribution used here has the same mean and variance as the Nicholson–Donnelly model but avoids the analytical problems that would arise from the atoms at 0 and 1. The binomial distribution is a natural choice of distribution for x_{ij} where it is a count of the number of times out of a possible n_{ij} that one of two possible alleles can be drawn. The π_i can take values in $(0,1)$ and so a beta distribution prior is a natural choice. Since the decision about which of two variants are counted is an arbitrary one, a symmetric distribution is also a natural choice, hence the repeated hyperparameter, a. There is no reason a-priori to believe that any locus should be different from any other, so a is the parameter for all π_i. However, this can easily be changed if there was a particular reason to do so. The c_j can also take values in $(0,1)$ so

a beta distribution prior is again a natural choice. Here, it is easier to envisage a situation where there could be good arguments for a non-symmetrical distribution to allow the experimenter flexibility to set a strong prior on the drift for a particular period of drift j. Hence, the hyperparameters, b_{1j}, b_{2j} are allowed to differ from each other. Nonetheless, unless otherwise stated these hyperparameters will be taken to be one to represent a prior where all values of the π_i and c_j are equally likely. Often, a case can be made for other weak priors such as the Jeffreys prior, which in this case would be beta$(0.5, 0.5)$. However, the reason for doing that would be to make it invariant to alternative choices of scale. In the case of π_i and c_j, there are no obvious alternative choices of scale so this was not considered to be a worthwhile choice at this stage before considering robustness of the model to alternative choices of prior later.

3.3 The Nicholson–Donnelly Model

3.3.1 Drift of Rare Alleles in the Nicholson–Donnelly Model

Nicholson et al. (2002) argue that since a normal distribution provides a good approximation to a binomial distribution for all but small population sizes that modelling genetic drift with some form of normal distribution is appropriate. However, since the proportion of an allele cannot vary beyond 0 and 1, the normal distribution needs to be rectified at these points so that the whole of the normal distribution below 0 counts as 0 and the whole of the distribution above 1 counts as 1. So an alternative to the beta distribution used by Balding and Nichols is to use a normal distribution rectified at 0 and 1. Nicholson et al. (2002) use a normal distribution with the same mean and variance as Balding and Nichols' beta distribution so that it is $N^{R[0,1]}\left(\alpha_t, c\alpha_t\left(1-\alpha_t\right)\right)$. However, rectifying a Normal at 0 and 1 results in shifting the mean of the new distribution towards 0.5, while the median remains at α_t. So, as well as being analytically awkward, which was why

it was not used for the earlier models, it also does not have the property that the future expected proportion of the allele is the last known proportion α_t; in fact it will be slightly closer to 0.5. The median will be α_t and so the proportion will be as likely to increase as decrease. However, as noted above, this is not a property of the Wright–Fisher model. Some notes on rectified normal distributions, including the mean, variance and a notation for describing rectified normal distributions is included in appendix B.

However, the Nicholson–Donnelly model does have a number of desirable properties. First of all, it does allow the alleles to become fixed with proportions at 0 or 1. Importantly, it can be seen from figure 3.4, that the shape of the rectified normal distribution for a rare allele does look much more similar in shape to that for the Wright–Fisher model as the amount of genetic drift (represented by c) increases (compare figure 3.4 with figure 3.1). The main difference is that the central mode of the distribution that can be seen shifting slightly to the left towards 0 in the Wright–Fisher model of figure 3.1 with increasing numbers of generations, remains fixed in the Nicholson–Donnelly model of figure 3.4.

3.3.2 Implementation of the Nicholson–Donnelly Model

Attention moved to implementing the Nicholson–Donnelly model as described by Nicholson et al. (2002) with the idea of then moving on to extend it. This is very similar to the Balding–Nichols model described earlier with the key difference that the proportion of an allele α_{ij} at locus i for the present-day subpopulation j is modelled by a rectified normal distribution rather than a beta distribution. Before rectification, the normal distribution has the same first two moments as the beta distribution used in the Balding–Nichols model (mean π_i and variance $\pi_i\,(1-\pi_i)\,c_j$). However, the act of rectification perturbs these moments. The

model is now

$$\alpha_{ij}|\pi_i, c_j \sim \mathrm{N}^{\mathrm{R}[0,1]}\left(\pi_i, \pi_i\left(1-\pi_i\right)c_j\right), \text{ independently,} \tag{3.3}$$

with the rest of the model remaining the same as in the previous section.

$$x_{ij}|n_{ij}, \alpha_{ij} \sim \text{Binomial}\left(n_{ij}, \alpha_{ij}\right), \text{ independently,}$$

with priors

$$\pi_i|a \sim \text{Beta}\left(a, a\right), \text{ independently,}$$

$$c_j \sim \text{Beta}(b_{1j}, b_{2j}), \text{ independently,}$$

where

i labels the locus: $1 \leqslant i \leqslant L$,

j labels the subpopulation $1 \leqslant j \leqslant P$,

n_{ij} is the total number of alleles observed at locus i in subpopulation j,

x_{ij} is the number of one of the two alleles observed at locus i in subpopulation j,

α_{ij} is the population proportion of that allele at locus i in subpopulation j,

π_i is the proportion of that allele at locus i in the ancestral population,

c_j is the amount of genetic drift in subpopulation j.

a is a hyperparameter in the prior of π_i.

b_{1j}, b_{2j} are hyperparameters in the prior of c_j and assigned the value 1 unless otherwise stated.

The DAG also remains the same (figure 3.3). The reasons for the choices of priors also remain the same as in section 3.2.2.

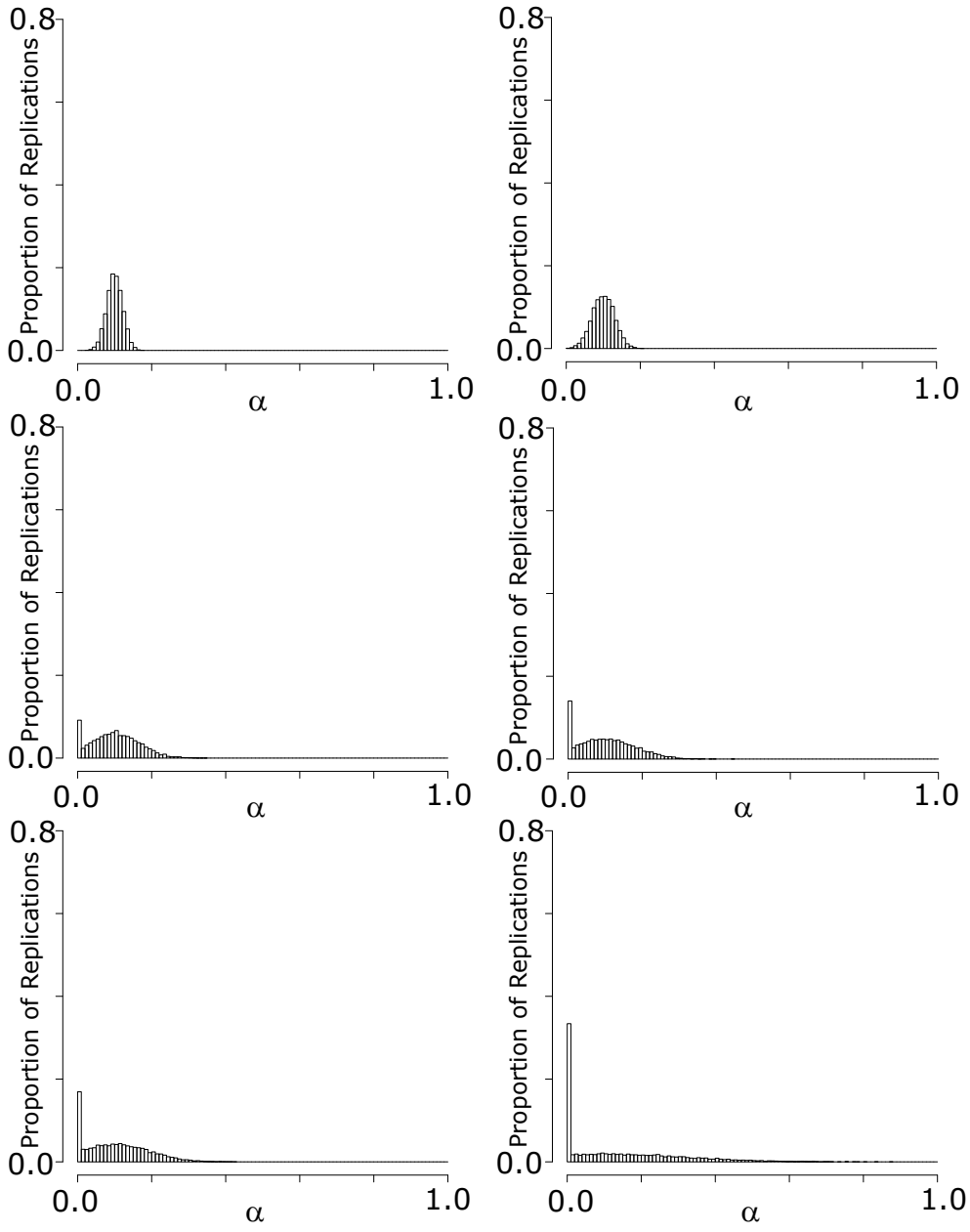

Figure 3.4: Nicholson–Donnnelly Model: Distribution of α for Increasing Genetic Drift c

This illustrates how the rectified normal distribution, $N^{R[0,1]}(\alpha_t, c\alpha_t(1-\alpha_t))$, rectified at 0 and 1, the Nicholson–Donnelly Model's approximation to the Wright–Fisher model, develops with increasing c, the parameter for genetic drift. This is shown for an initial value of $\alpha_t = 0.1$. The spike that collects at 0 (and later 1) is mostly made up of exact 0s (or 1s). Here, note how, with increasing c, unlike the Wright–Fisher model, a mode remains at 0.1 and the mean of the distribution shifts towards 0.5.

3.3.3 Interpretation of the Model Parameters

A virtue of the model developed by Nicholson et al. (2002) was that the parameters had clear and interpretable meanings. This is a virtue that it is an aim for this thesis to maintain throughout. There is a split between the locus-specific parameter, π_i, the allele frequency at the ancestral population and the subpopulation specific drift parameter, c_j. The α_{ij} represent the different allele frequencies for each subpopulation in a clear way. The allele frequency parameters, π_i and α_{ij} have an obvious and intuitive interpretation. The c_j, however, encapsulates effective population size and time in terms of generations and perhaps needs a little more explanation.

Genetic drift in the Wright–Fisher model was explained in Chapter 1 in terms of a new generation of alleles at a locus being formed by making draws of alleles with replacement from its previous generation. Hartl and Clark (1997) give the variance of the allele frequency change as $\frac{\pi(1-\pi)}{2N}$ where N is the population size of a diploid species. The variance in the Balding–Nichols and Nicholson–Donnelly models is $\pi(1-\pi)c$ so, in this one generation case, $1-c$ can be interpreted as $1-\frac{1}{2N}$. However, the purpose of these models is to model drift over a great many generations. Hartl and Clark (1997) describe what happens to Wright's F statistic over a number of generations. Taking F_t to mean the value of the F statistic at time t, they give the formula for its change over 1 generation as $1-F_1 = \left(1-\frac{1}{2N_0}\right)(1-F_0) = (1-c)(1-F_0)$. If the population stays constant in size at N_0 then over t generations, $1-F_t = \left(1-\frac{1}{2N_0}\right)^t (1-F_0)$. However, if the population fluctuates then N_0 is replaced with the effective population, N_e. Over one generation, $N_0 = N_e$ but over t generations $\frac{1}{N_e} \approx \frac{1}{t}\left(\frac{1}{N_0}+\frac{1}{N_1}+\cdots+\frac{1}{N_{t-1}}\right)$. $1-F_t = \left(1-\frac{1}{2N_e}\right)^t (1-F_0) = (1-c)(1-F_0)$. So c can be thought of as $c = 1-\left(1-\frac{1}{2N_e}\right)^t \approx 1-\exp\left(-\frac{t}{2N_e}\right)$. So, in this way, c encapsulates population fluctuations and time. There are a couple of interesting points to note here. First, when $F_0 = 0$, such as would be the case for a pairwise F of a population with itself,

one part of which then splits off and drifts giving rise to $F_t > 0$, $1 - F_t = 1 - c$ or $F_t = c$, showing the relationship in concept of F_{ST} and c. The second is to note that it takes just one generation with a small population size such as might happen after a natural disaster, war, famine or other such incident, to drastically reduce N_e, even if the population recovers to its earlier size over a small number of generations, so c also reflects the effects of population bottlenecks such as this as well as time and population size. The use of c in this way, provides a population-specific parameter with an interpretable meaning while avoiding the considerable complications that would arise from modelling each generation and its population size explicitly.

3.3.4 Full Conditionals for the Balding–Nichols Model

Using the Balding-Nichols drift model, it is possible to integrate out the α_{ij}s, since the beta is conjugate to the binomial:

$$P\left(x_{ij}|\pi_i,c_j\right) = \int_0^1 P\left(x_{ij}|\alpha_{ij}\right) P\left(\alpha_{ij}|\pi_i,c_j\right) d\alpha_{ij}, \tag{3.4}$$

so that

$$P\left(x_{ij}|\pi_i,c_j\right) = \frac{n_{ij}!}{x_{ij}!\left(n_{ij}-x_{ij}\right)!}\,\frac{1}{B\left(\frac{\pi_i(1-c_j)}{c_j},\frac{(1-\pi_i)(1-c_j)}{c_j}\right)} \tag{3.5}$$

$$\times \int_0^1 \alpha_{ij}^{x_{ij}}\left(1-\alpha_{ij}\right)^{n_{ij}-x_{ij}} \alpha_{ij}^{\left(\frac{\pi_i-c_j\pi_i-c_j}{c_j}\right)}\left(1-\alpha_{ij}\right)^{\left(\frac{1-2c_j-\pi_i+c_j\pi_i}{c_j}\right)} d\alpha_{ij}.$$

But, since $\int_0^1 u^{g-1}\left(1-u\right)^{h-1} du = B\left(g,h\right)$ the beta function of arguments g and h,

$$P\left(x_{ij}|\pi_i,c_j\right) = \frac{n_{ij}!}{x_{ij}!\left(n_{ij}-x_{ij}\right)!}\,\frac{B\left(x_{ij}+\pi_i\left[\frac{1-c_j}{c_j}\right],n_{ij}-x_{ij}+(1-\pi_i)\left[\frac{1-c_j}{c_j}\right]\right)}{B\left(\frac{\pi_i(1-c_j)}{c_j},\frac{(1-\pi_i)(1-c_j)}{c_j}\right)}. \tag{3.6}$$

Thus x_{ij}, unsurprisingly, follows a beta-binomial distribution given π_i and c_j.

Taking $\theta = \{\pi_1, ..., \pi_L, c_1, ..., c_P\}$, the full-conditional probabilities for π_i and c_j that are needed for sampling from the posterior by Gibbs sampling can be obtained from

$$P(\theta|x) \propto P(\pi) P(c) P(x|\pi, c) = \prod_{i=1}^{L} P(\pi_i) \prod_{j=1}^{P} P(c_j) \prod_{i=1}^{L} \prod_{j=1}^{P} P(x_{ij}|\pi_i, c_j), \quad (3.7)$$

using the product rule, as,

$$P(\pi_i|a, x, c, \pi_{-i}) \propto \pi_i^{a-1} (1-\pi_i)^{a-1}$$

$$\times \prod_{j=1}^{P} \left(\frac{n_{ij}! \, B\left(x_{ij} + \pi_i \left[\frac{1-c_j}{c_j}\right], n_{ij} - x_{ij} + (1-\pi_i)\left[\frac{1-c_j}{c_j}\right]\right)}{x_{ij}! \, (n_{ij} - x_{ij})! \, B\left(\frac{\pi_i(1-c_j)}{c_j}, \frac{(1-\pi_i)(1-c_j)}{c_j}\right)} \right) \quad (3.8)$$

and

$$P(c_j|x, c_{-j}, \pi, b) \propto \prod_{i=1}^{L} \left(\frac{n_{ij}! \, B\left(x_{ij} + \pi_i \left[\frac{1-c_j}{c_j}\right], n_{ij} - x_{ij} + (1-\pi_i)\left[\frac{1-c_j}{c_j}\right]\right)}{x_{ij}! \, (n_{ij} - x_{ij})! \, B\left(\frac{\pi_i(1-c_j)}{c_j}, \frac{(1-\pi_i)(1-c_j)}{c_j}\right)} \right)$$

$$\times c_j^{b_{1j}-1} (1-c_j)^{b_{2j}-1}. \quad (3.9)$$

Since these full-conditionals cannot be directly sampled, a Metropolis-Hastings-within-Gibbs MCMC approach was taken and implemented in R. A truncated normal proposal distribution for π_i and c_j was chosen where the parts of the normal distribution outside $[0, 1]$ are discarded and the remainder renormalised. This was done because the only allowed values for π_i and c_j are in $[0, 1]$. This also allowed more direct control of the variance of the proposal. In order to ensure reasonable acceptance rates [of about 0.44 as suggested by Rosenthal (2010), an adaptive algorithm for setting the variance of the proposal, as described by Rosen-

thal (2012) and Roberts and Rosenthal (2009), was used, with an adaptation and burn-in time of 10,000 iterations. In testing, by examining traces from the posterior chains from each parameter, the chains appeared to converge much sooner than 10,000 iterations (in fact less than 2,000 iterations) but 10,000 is a round number that is comfortably large enough to feel comfortable about convergence before checking with more formal methods such as Gelman's R. The proposal variances were then fixed and a further 10,000 iterations were taken to provide samples from the posterior distributions of the parameters. The model was run on the HapMap data for each of the 22 autosomes and for all 22 autosomes together.

3.3.5 Full Conditionals for the Nicholson–Donnelly Model

Unlike for the Balding–Nichols model, α_{ij} cannot be integrated out analytically. Nevertheless, all the α_{ij}s can be sampled too with the cost of increasing the computational time. Full conditionals need to be found not only for π_i and c_j but also for α_{ij}, if the sampling is done by Gibbs sampling.

The full conditional for π_i is,

$$P\left(\pi_i|\alpha, c, \pi_{-i}\right) \propto \pi_i^{a-1}\left(1-\pi_i\right)^{a-1} \prod_{j=1}^{P} g\left(c_j, \pi_i, \alpha_{ij}\right), \tag{3.10}$$

where

$$g\left(c_j, \pi_i, \alpha_{ij}\right) = \begin{cases} \left[c_j\pi_i\left(1-\pi_i\right)\right]^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\pi_i)^2}{2c_j\pi_i(1-\pi_i)}\right) dr, & \alpha_{ij}=0, \\ \left[c_j\pi_i\left(1-\pi_i\right)\right]^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ij}-\pi_i)^2}{2c_j\pi_i(1-\pi_i)}\right), & 0<\alpha_{ij}<1, \\ \left[c_j\pi_i\left(1-\pi_i\right)\right]^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\pi_i)^2}{2c_j\pi_i(1-\pi_i)}\right) dr, & \alpha_{ij}=1 \end{cases} \tag{3.11}$$

and P is the number of subpopulations.

The full conditional for c_j is

$$P\left(c_j|\alpha,\pi,c_{-j},b\right) \propto \prod_{i=1}^{L} g\left(c_j,\pi_i,\alpha_{ij}\right) \times c_j^{b_{1j}-1}\left(1-c_j\right)^{b_{2j}-1}, \tag{3.12}$$

where L is the number of loci.

However, the full conditional for α_{ij} is a bit more awkward:

$$P\left(\alpha_{ij}|c_j,\pi_i,\alpha_{-ij},x_{ij},n_{ij}\right) \propto h\left(n_{ij},x_{ij},\alpha_{ij}\right) g\left(c_j,\pi_i,\alpha_{ij}\right), \tag{3.13}$$

where, this time, the $\left[c_j\pi_i\left(1-\pi_i\right)\right]^{-\frac{1}{2}}$ term in $g\left(c_j,\pi_i,\alpha_{ij}\right)$ can be taken out because it does not depend on the value of α_{ij}, and

$$h\left(n_{ij},x_{ij},\alpha_{ij}\right) = \begin{cases} 1, & \alpha_{ij}=0, x_{ij}=0, \\ \alpha_{ij}^{x_{ij}}\left(1-\alpha_{ij}\right)^{n_{ij}-x_{ij}}, & 0<\alpha_{ij}<1, \\ 1, & \alpha_{ij}=1, x_{ij}=n_{ij}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.14}$$

Once again, an MCMC sampler was implemented in R (R Core Team, 2013). Metropolis–Hastings–within–Gibbs and the adaptive MCMC method described by Roberts and Rosenthal (2009) were used to sample c_j and π_i at each iteration but all the α_{ij} were sampled using rejection sampling. A first attempt was made to sample α_{ij} also using Metropolis–Hastings–within–Gibbs. However this implementation led to α_{ij} getting stuck at the values 0 or 1 far more often than it should have in cases where $x_{ij}=0$ or $x_{ij}=n_{ij}$, respectively, leading to unsatisfactory mixing. Using rejection sampling for α_{ij} remedied this problem.

3.3.6 Results of Simulations

Data were simulated under the assumptions of the Nicholson–Donnelly model for random values of the parameters, c_j and π_i. The c_j were selected independently from a continuous uniform(0,0.3) distribution. This is the likely range of values to be met in practice. The π_i were selected independently from a Beta(1,1) distribution to allow for the full range of possibilities. Data for 11 subpopulations each of size 300 (or 150 individuals) and 2400 loci were simulated to make the samples similar to those that might be encountered in the HAPMAP data for a large chromosome. The sampler was used to see if the original parameter values of c_j and π_i were recovered from the data. The results for the c_js are shown in table 2 for one such typical simulation.

Table 3.1: Nicholson–Donnelly Model Estimates of Drift Parameters Compared With True Values

	Nicholson-Donnelly Model Estimates of Simulated c_j		
	95% Central Credible Value Interval Bounds		True Value
j	lower	upper	
1	0.2515	0.2878	0.2614
2	0.2279	0.2618	0.2451
3	0.1394	0.1590	0.1570
4	0.1082	0.1242	0.1195
5	0.0971	0.1112	0.0995
6	0.2453	0.2817	0.2672
7	0.0087	0.0115	0.0102
8	0.1818	0.2090	0.2014
9	0.0947	0.1085	0.1001
10	0.1225	0.1399	0.1267
11	0.1934	0.2199	0.2189

It can be seen that despite there being a wide variety of magnitudes of drift, the 95% credible intervals all contain the true values of c_j in this typical example. In addition, the 95% credible intervals for the π_is contained the true values 2262 times out of 2400, 94.25% of the loci, which is acceptable. The true values of the $11 \times 2400 = 26400$ α_{ij}s were found to be within the 95% credible intervals from the sampler 25091 times or 95.04% of occasions which, again, is much as should

be expected.

3.4 The HapMap Dataset

The data that will be used in this chapter will be from HapMap. The HapMap dataset comes from the HapMap project described by International HapMap Consortium (2003). It contains data on SNPs from throughout the human genome for 988 individuals from 11 subpopulations. There are 4 subpopulations of African origin, African ancestry in southwest USA (ASW), Luhya in Webuye, Kenya (LWK), Maasai in Kinyawa, Kenya (MKK) and Yoruba in Ibadan, Nigeria (YRI) (loosely described as "Africans" subsequently). The remaining 7 subpopulations are Utah residents with North and West European ancestry (CEU), Han Chinese in Beijing (CHB), Han Chinese in Denver, Colorado (CHD), Gujaratis in Houston, Texas (GIH), Japanese in Tokyo (JPT), residents of Los Angeles, California with Mexican ancestry (MEX) and Tuscans, Italy (TSI). (The two Chinese and the Japenese subpopulation will be collectively loosely described as "East Asian" subsequently, while CEU and TSI will be loosely described as "European").

3.4.1 Data Cleaning

A C++ program was written to parse the HapMap data files from HAPMAP phase 3 release 2. There were 242 files to process in all, 1 for each combination of the 22 autosomes and 11 subpopulations. Five of the subpopulations, ASW, CEU, MKK, MEX and YRI contained some immediately related individuals, two parents and a child. The child record was removed from these to ensure that there were no immediately related individuals in the samples. Loci were selected to be at least 100,000 base pairs apart and to have no missing data for any subpopulation. The loci were selected to be at least 100,000 base pairs apart to ensure the assumption that they are independent is not violated by linkage to an important extent.

Pritchard and Przeworski (2001) state that 5-10 markers within 50,000 bases of a locus would be needed to ensure that one was in strong linkage disequilibrium with that locus. While a spacing of at least 100,000 base pairs therefore reduces the chance that adjacent such loci are in strong linkage disequilibrium with each other, no spacing can eliminate the possibility. A judgement was made that this represented the balance between reducing such a risk and the loss of useful allele frequency information.

The data used are genotypes and are stored as allele counts. The loci selected all have exactly two variants. One of two variants at a locus will be counted with the one to be counted chosen at random. So if at a locus the two variants are Guanine and Adenine, Guanine could be chosen at random to be counted out of the two. For an individual in a subpopulation, there are three possibilities, the two homozygotes, GG and AA and the heterozygote, GA (or equivalently, AG). The first two of these cases would score 2 and 0 respectively, and the other would score 1. The scores for the individuals within a subpopulation are summed to produce a total for the subpopulation which will be an integer between or including 0 and twice the number of individuals sampled for that subpopulation. The dataset to be used will thus take the form of allele counts for each of the 11 subpopulations at each locus, which was written into a file of counts data for each selected locus and subpopulation in a format that could be easily read by R.

The total number of loci in each chromosome that was in the dataset after this thinning process is shown in table 3.2. The remaining sample sizes and total numbers of individuals in each subpopulation are shown in table 3.3.

Table 3.2: Number of Loci in Each Chromosome in the Dataset

Chromosome	Loci
1	2063
2	2189
3	1820
4	1737
5	1639
6	1568
7	1417
8	1319
9	1041
10	1216
11	1219
12	1204
13	898
14	804
15	721
16	705
17	719
18	697
19	513
20	565
21	312
22	309

Table 3.3: Sample Sizes for Each Subpopulation in the Dataset

Subpopulation	Sample Size	Individuals
ASW	98	49
CEU	224	112
CHB	168	83
CHD	170	85
GIH	176	88
JPT	172	86
LWK	180	90
MEX	100	50
MKK	286	143
TSI	176	88
YRI	226	113

3.5 Results from Application to the HapMap Dataset

3.5.1 Results for Balding–Nichols Model

Figure 3.5 shows the estimated values of c_j for each chromosome and each subpopulation. The whiskers show the central 95% posterior credible interval for that point estimate. The parameter is the estimated median of the posterior distribution (which was found to be nearly identical to the posterior mean value). There is a lot of variation between the drift estimates for each chromosome; more than would be expected by random variation. In particular, there are prominantly large estimates of drift for the East Asian, Central European, Tuscan and Mexican subpopulations for chromosome 16 compared with those for other chromosomes. Populations that might be expected to be closely related such as the Japanese and Chinese subpopulations show similar patterns.

3.5.2 Results For Nicholson–Donnelly Model

The Nicholson-Donnelly drift model was applied to the same HapMap data to see how the results from each model compared. Figure 3.6 shows the estimated values of c_j for each chromosome and each subpopulation. Some interesting points come from the comparison. The unusually large c values for Chromosome 16 under the Balding–Nichols model (see figure 3.5 for comparison) are gone. There remains more variation between each chromosome's estimates of genetic drift for each subpopulation than would be expected by random variation. In some subpopulations, there is not a value of c_j that would be contained inside enough of the 22 intervals for random variation alone to be a plausible explanation for the differences. Nevertheless, populations that would be expected to be closely related again produce similar patterns. For example, the patterns for the East Asian sub-populations

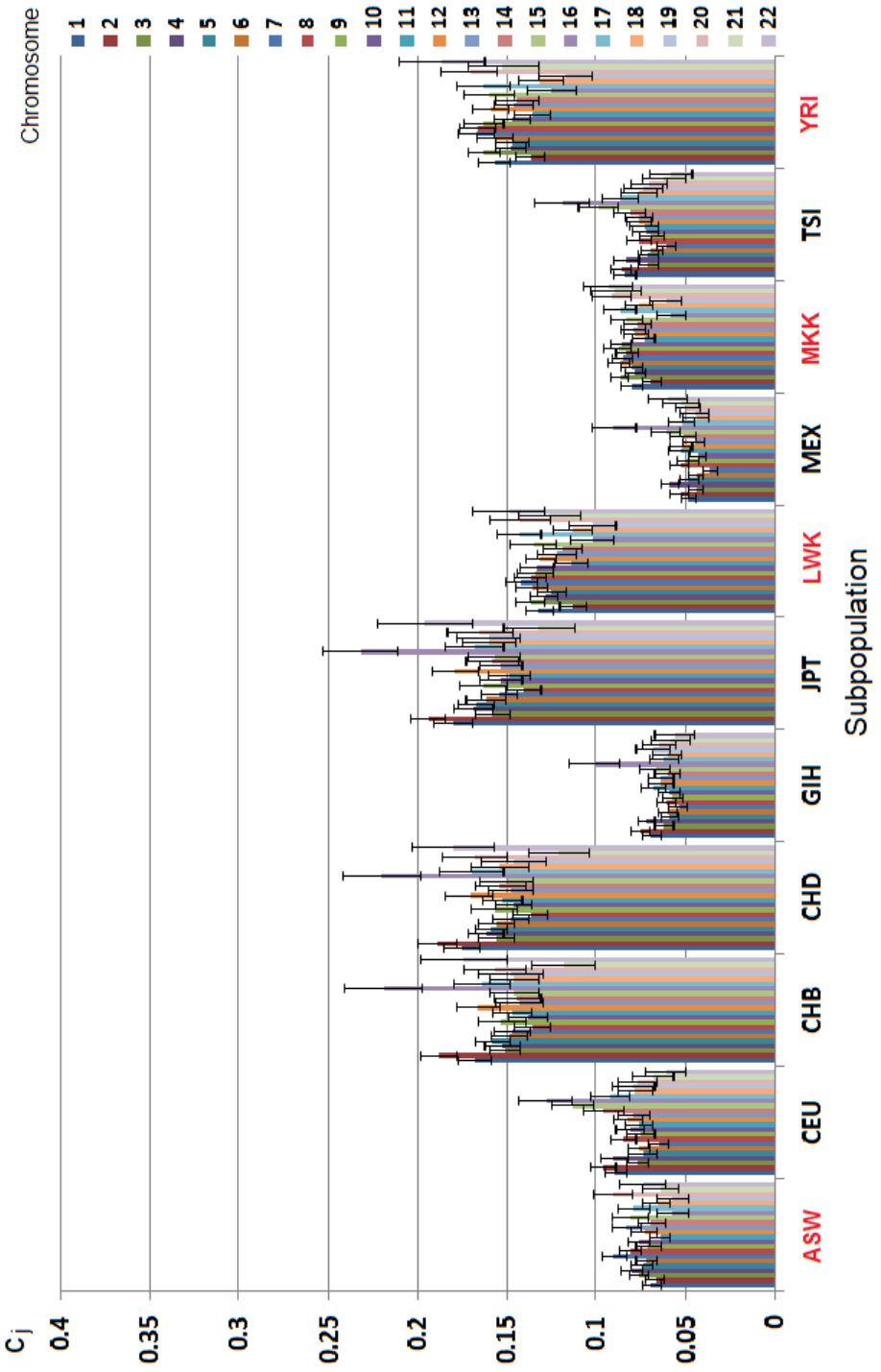

Figure 3.5: Estimated Values of c_j by Subpopulation and Chromosome for the Balding–Nichols Model

Coloured columns represent the different chromosomes. Point estimates (medians) of the genetic drift parameter, c were made for each subpopulation and each chromosome with whiskers representing the central 95% posterior probability density interval in each case.

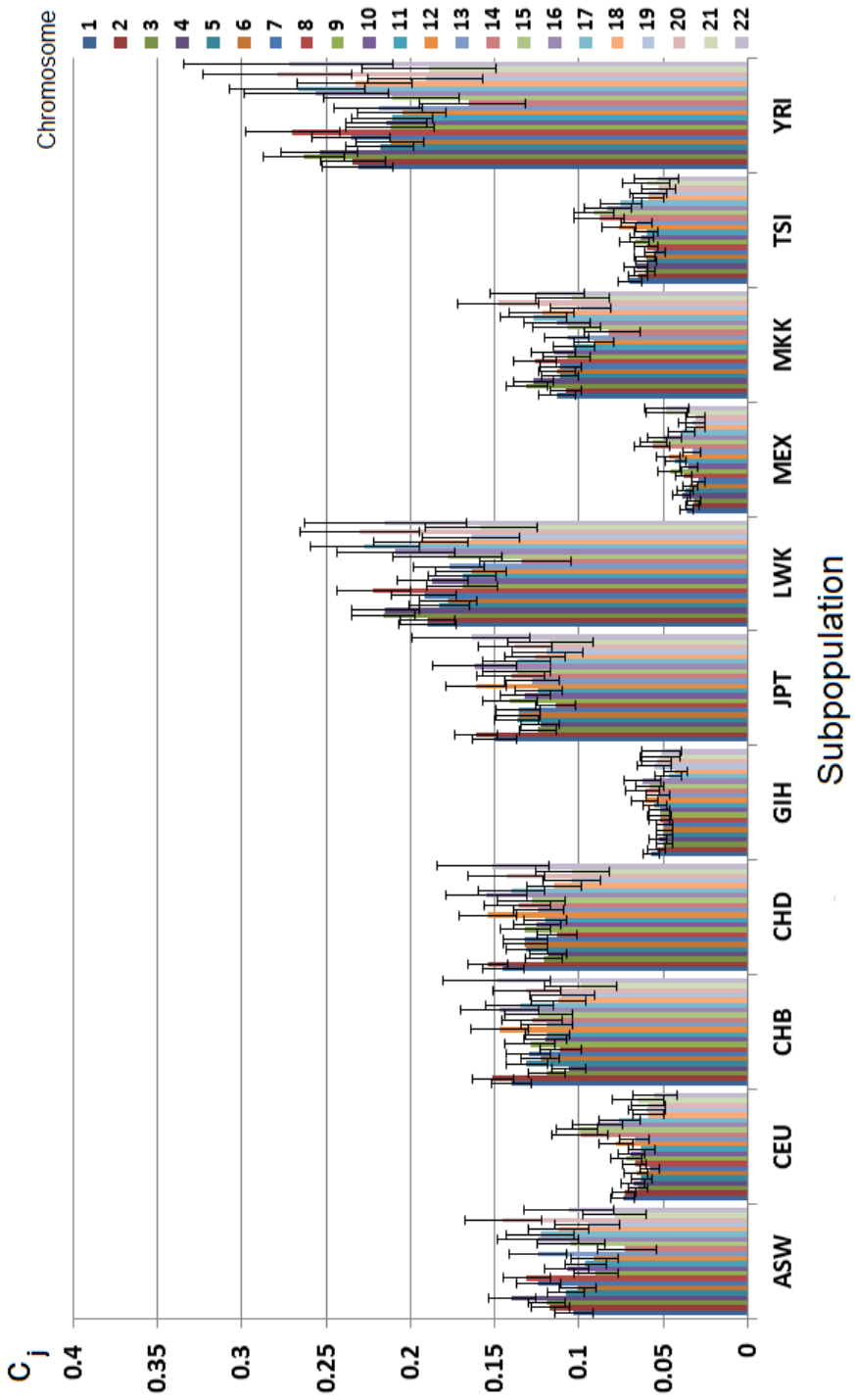

Figure 3.6: Estimated Values of c_j by Subpopulation and Chromosome for the Nicholson-Donnelly Drift Model

Coloured columns represent the different chromosomes. Point estimates (medians) of the genetic drift parameter were made for each subpopulation and each chromosome with whiskers representing the central 95% posterior probability density interval in each case.

remain similar, as do the two European subpopulations. In comparison with the equivalent graph (figure 3.5) for the Balding–Nichols model, the estimated levels of drift have reduced for all subpopulations other than the Africans whose estimates of drift have increased.

3.5.3 Comparison of the Models

Residuals were examined to assess how well each model fits the data. The standardised residuals for the Nicholson–Donnelly model were calculated in the same way as by Nicholson et al. (2002) as

$$e_{ij} = \frac{x_{ij}/n_{ij} - \hat{\pi}_i}{\left[\left\{\hat{c}_j + (1-\hat{c}_j)/n_{ij}\right\}\hat{\pi}_i\left(1-\hat{\pi}_i\right)\right]^{\frac{1}{2}}}, \tag{3.15}$$

where $\hat{\pi}_i$ is the estimated mean of the posterior distribution of π_i and $\hat{c}_j$ is the estimated mean of the posterior distribution of c_j. There were fewer large standardised residuals for the Nicholson–Donnelly model and the sizes of the standardised residuals were smaller in general. However the model fits were compared more formally using the Watanabe Akaike Information Criterion (WAIC) for the two models for each chromosome. These are shown in table 3.4. It can be seen from the table that the WAIC for the Nicholson-Donnelly model is much lower than for the Balding-Nichols model for all 22 autosomes. The WAIC for the Nicholson–Donnelly model is only about three quarters of the size of the WAIC for the Balding-Nichols model in most cases. As with other information criteria, the model with the lower WAIC is preferred. So these results indicate that the Nicholson–Donnelly model is a clearly better model of the data compared to the Balding-Nichols model for all 22 autosome datasets.

Table 3.4: Comparison of the Watanabe–Akaike Information Criterion for the Balding–Nichols and Nicholson–Donnelly Models

	WAIC		
Chromosome	BN	ND	Difference (BN-ND)
1	179049	128111	50938
2	193213	138874	54339
3	162975	117737	45238
4	156198	112482	43716
5	146613	106114	40499
6	139770	101214	38556
7	128023	93045	34978
8	118099	85338	32761
9	93399	67548	25852
10	107725	77906	29819
11	107357	77972	29384
12	105408	75721	29687
13	81727	59352	22376
14	71443	51738	19705
15	65988	47383	18605
16	60640	42915	17725
17	62917	45071	17846
18	61881	45079	16802
19	43472	31922	11550
20	50824	36457	14367
21	28293	20780	7513
22	27717	19863	7854

3.6 Problems with the Models

It would be expected that analysis of different chromosomes would yield similar values of c_j for each subpopulation since drift should affect all chromosomes equally. However, there is not enough overlap between the 95% credible intervals to support this. In particular, chromosome 16 produces unusually high values of c_j in non-African subpopulations (labelled in red in figure 3.5) and unusually low ones for Africans in the Balding–Nichols model. Further, it was found that residuals had a bimodal distribution (e.g., figure 3.7).

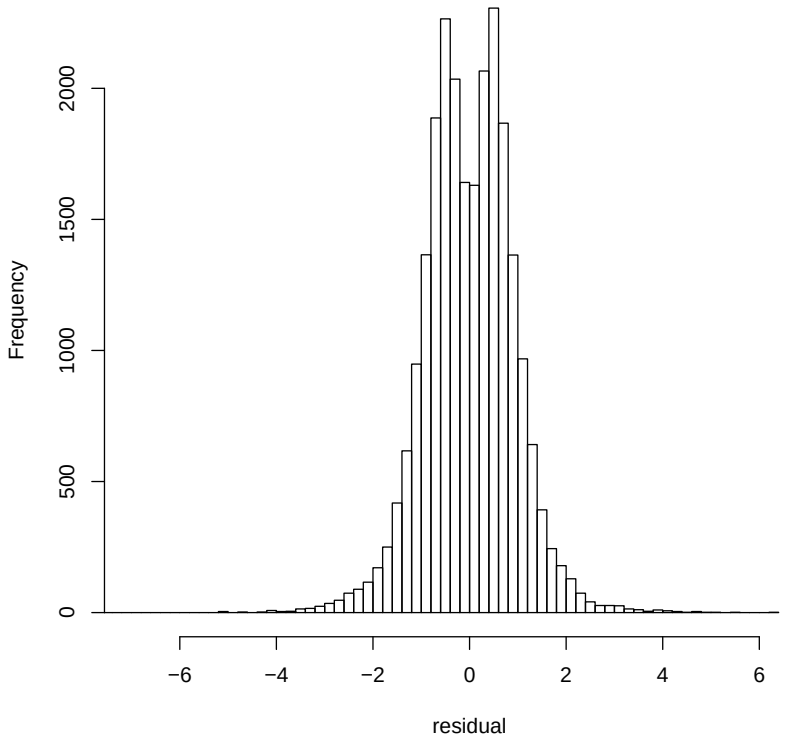

Figure 3.7: Histogram of Standardised Residuals for Chromosome 2 for the Balding–Nichols model

The distribution of the residuals also has heavy tails. The extreme residuals were found to be predominantly from the African subpopulations.

Histograms of standardised residuals had a bimodal pattern and normal QQ plots were distinctly non-linear (e.g., figure 3.8). The former indicated that there was structure within the data that the model did not adequately explain and the latter that the residuals were rather heavy-tailed. This was the case for both the Balding-Nichols (figure 3.8) and the Nicholson–Donnelly (figure 3.9) models as can be seen for a typical example (chromosome 22). However, the normal QQ plot is markedly better behaved for the Nicholson–Donnelly model than it was for the Balding–Nichols model.

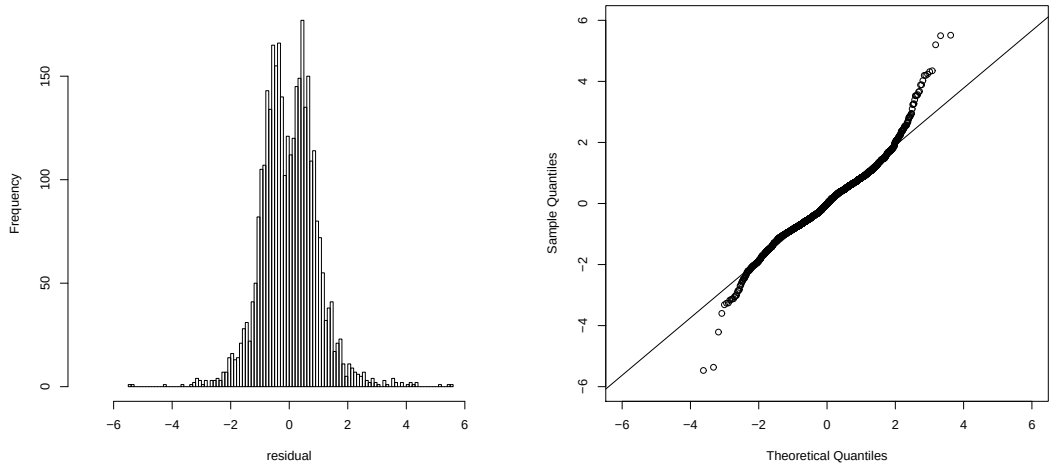

Figure 3.8: Histogram of Standardised Residuals and normal QQ Plot for Chromosome 22 for the Balding–Nichols Model

Diagnostic plots of standardised residuals for the Balding–Nichols model for Chromosome 22. The histogram on the left shows a bimodal pattern suggesting that there is information in the data that the model does not take sufficiently into account. The normal QQ plot on the right shows the heavy tails of the distribution of the residuals.

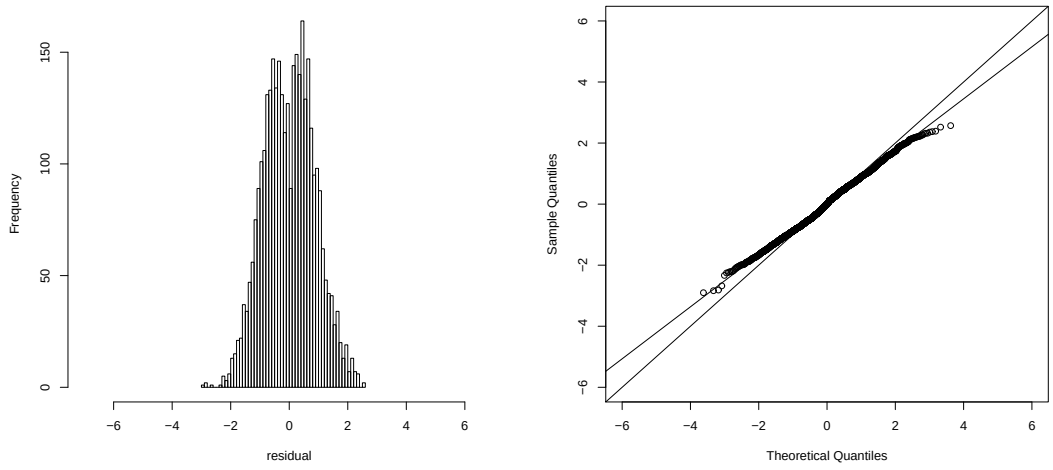

Figure 3.9: Histogram of Standardised Residuals and normal QQ Plot for Chromosome 22 for the Nicholson–Donnelly Model

Diagnostic plots of standardised residuals for the Nicholson–Donnelly models for Chromosome 22. The histogram on the left shows a bimodal pattern suggesting that there are factors in the data that the model does not take sufficiently into account. The QQ plot on the right gives no cause for concern on its own.

As mentioned, the most extreme residuals belonged predominantly to the African

subpopulations. It appeared that the model was not fully able to represent these subpopulations adequately within the full dataset. The model was refitted for both models for a subset of chromosomes to just the African subpopulations and just to non-African subpopulations. The estimated drift parameters are lower and there was an improved overlap between the credible intervals. Chromosome 16 is no longer an outlier even for the Balding–Nichols model (figure 3.10). The residual distributions are still heavy tailed but less markedly so and, in the case of the Africans, the bimodal feature of the distribution is reduced (figure 3.11).

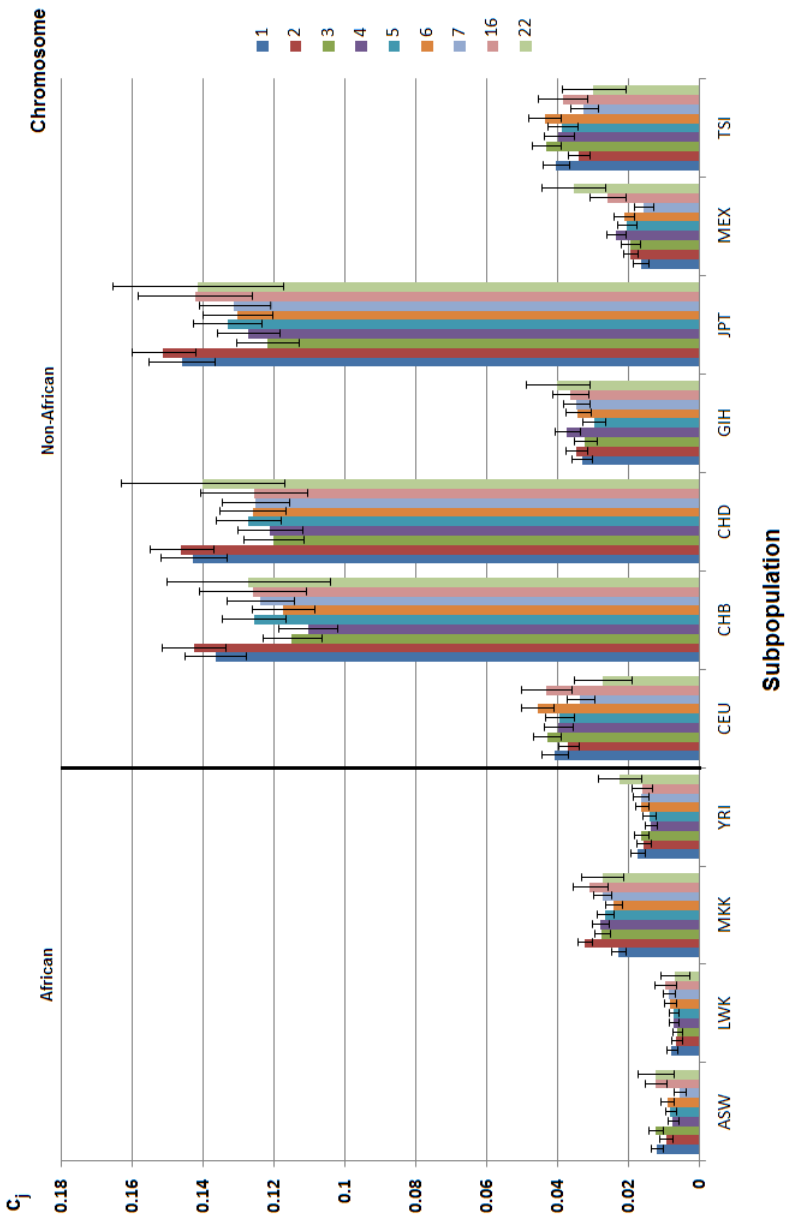

Figure 3.10: Estimated Values of c_j by Subpopulation and Chromosome (African Subpopulations Analysed Separately) for the Balding–Nichols model

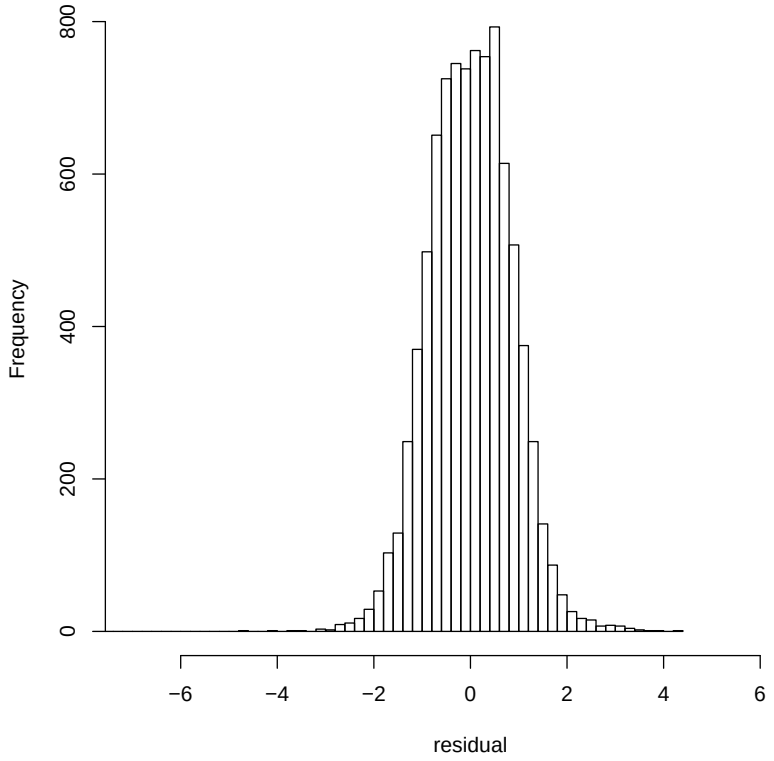

Figure 3.11: Histogram of Standardised Residuals for Chromosome 2 for African Subpopulations Analysed Separately Under the Balding–Nichols Model

However, the more interesting point is that the values for c_j have unambiguously reduced from those previously estimated. This is important because one of the assumptions of the Nicholson–Donnelly model is that the subpopulations diverged from a common ancestor population at much the same time. This seemed to be an adequate assumption for the limited combinations of subpopulations they considered (Nicholson et al., 2002). However, it is stretching credibility to imagine that the Beijing Han Chinese (CHB) subpopulation diverged from the Denver Han Chinese (CHD) at much the same time as it diverged from the Kenyan Maasai (MKK).

If the population tree is closer to figure 3.13 than to figure 3.12 then the model with only Africans will only be estimating c_j from the genetic drift from node C on figure 3.13 rather than from node A on figure 3.12, which is not as long ago and so would lead to a lower estimate of c_j. Similarly, a model with non-Africans only would calculate c_j from node B on figure 3.13 rather than node A on figure 3.12 resulting in a lower estimate of c_j. If figure 3.12 was closer to the true population genealogical tree then the values of c_j for each subset model of subpopulations would be unchanged (apart from some variation around the values due to loss of information about the values of the π_is, the proportions of each nucleotide at locus i in the ancestral population) from that of the full model because the timescale of genetic drift would remain unchanged.

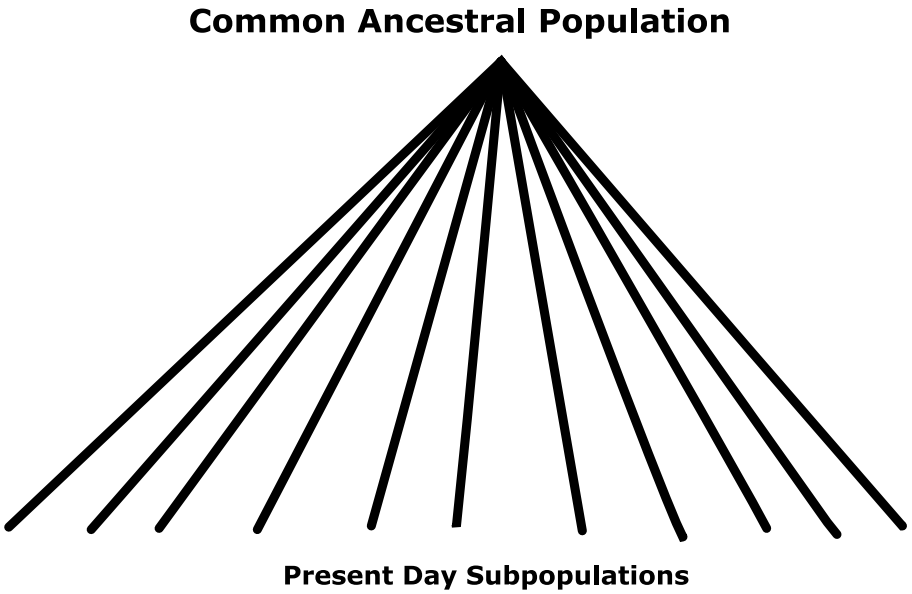

Figure 3.12: Genealogical Tree Assumed by Nicholson–Donnelly Model

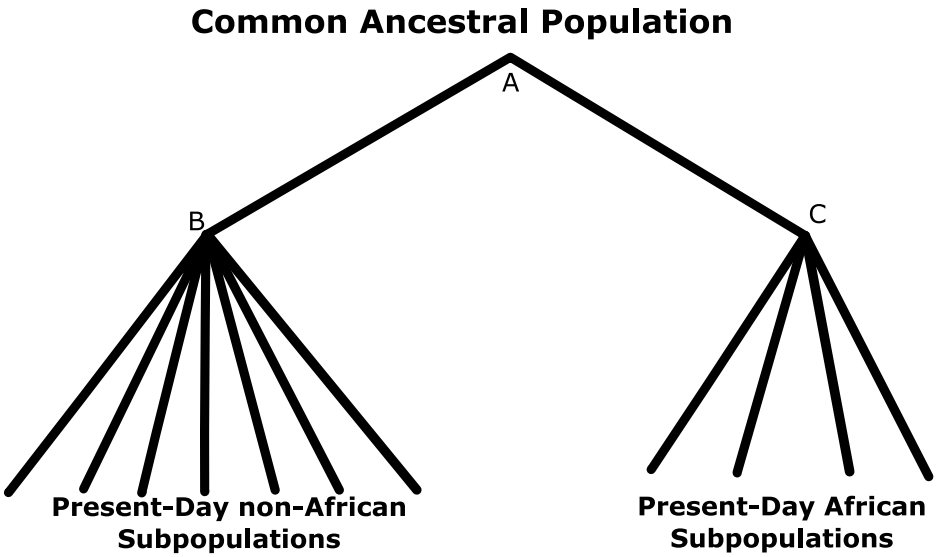

Figure 3.13: Alternative Genealogical Tree to that in figure 3.12

When the Nicholson–Donnelly model was rerun with only the data for African subpopulations, the bimodal distribution of the residuals largely disappeared (figure 3.14). However, if the non-Africans are analysed separately, there is still a suggestion of bimodality in the residuals (figure 3.15), suggesting that further subdivision may be necessary. In both cases, the normal QQ plots remain close to a straight line, giving no cause for concern. As also explained by Nicholson et al. (2002), the straight line deviates slightly from the $x = y$ line because the variance of the standardised residuals will be slightly less than 1; there will be some negative correlation between the P residuals associated with each locus where P is the number of subpopulations. Just as was the case for the Balding–Nichols model, when data from the African subpopulations are analysed alone, and when data with only the non-African subpopulations are analysed, they produce lower values for genetic drift from their ancestral population than when all the data were analysed together. This is consistent with these groups of subpopulations having diverged earlier from each other than the subpopulations within these two groups.

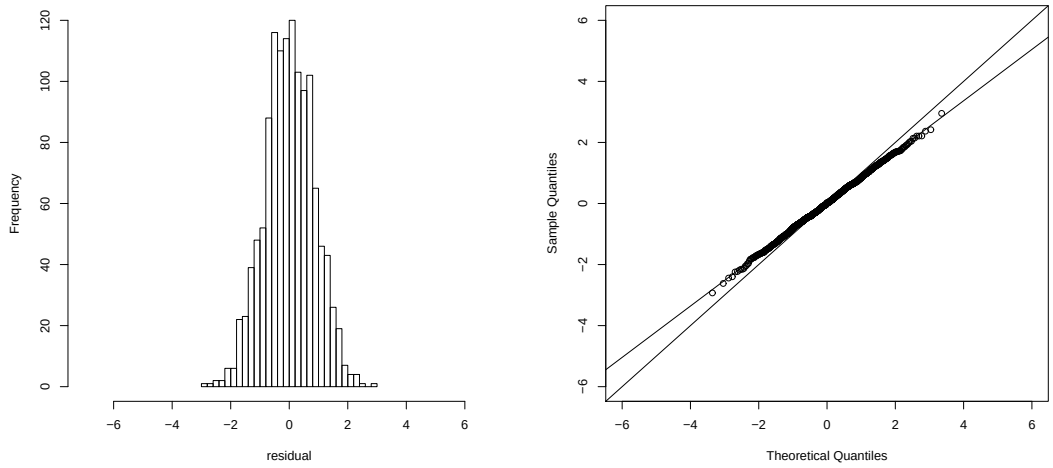

Figure 3.14: Histogram of Standardised Residuals and normal QQ Plot for Chromosome 22 for the Nicholson–Donnelly Model for African Subpopulations Only

The histogram on the left gives no cause for concern. The normal QQ plot on the right shows the residuals lying on an almost perfect straight line consistent with the residuals being normally distributed. The line along which the points on the normal QQ lie close to, deviates from the $x = y$ line there will be some negative correlation between the P residuals associated with each locus.

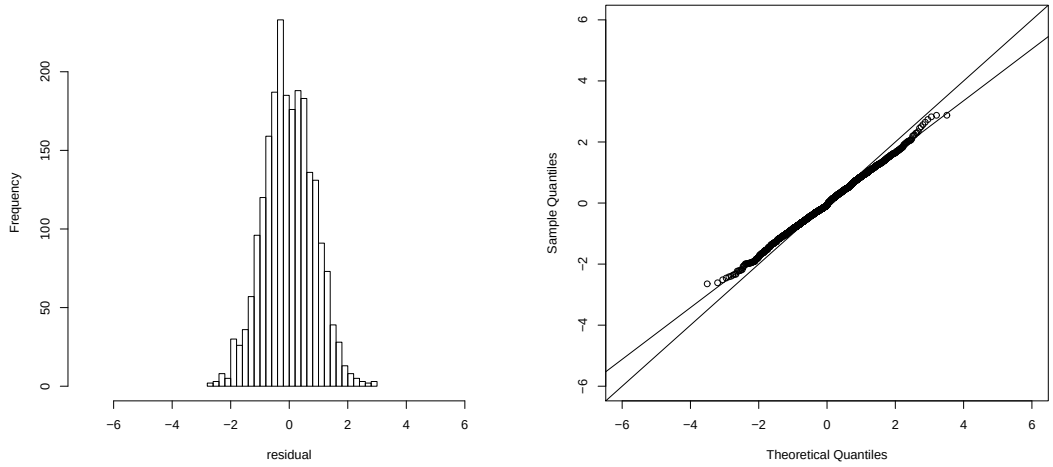

Figure 3.15: Histogram of Standardised Residuals and normal QQ Plot for Chromosome 22 for the Nicholson–Donnelly Model Without Africans

DThe histogram on the left has a hint of bimodality which suggests these data may need to be subdivided further before the model will explain the data properly. The normal QQ plot on the right shows the residuals lying on an approximately straight line. The line along which the points on the normal QQ lie close to, deviates from the $x = y$ line there will be some negative correlation between the P residuals associated with each locus.

Additional evidence for the split between the four African and seven non-African subpopulations was considered. If the African subpopulations were more similar to each other than they were to the non-African subpopulations, then their residuals at each locus would be expected to have the same sign significantly more often than would be expected from random chance. This turned out to be the case. For example, for chromosome 22 above, there are 309 loci in the dataset and the four African subpopulations all had the same sign of their residuals for 237 of these loci. There are 330 combinations of 4 subpopulations from 11 subpopulations. The number of loci for which each of the other 329 combinations of four subpopulations all had the same sign was also counted. Of these, the highest scoring other combination only scored 154. To place these scores in context, the median score was only 26 and interquartile range was 11 to 63. So, the four African subpopulation residuals had the same sign far more often than any other combination of four subpopulations. In the context of the bimodal pattern of residuals, this provides additional evidence that there is something about these subpopulations of which the model was not taking sufficient account.

This leads to the conclusion that a more complex version of the model will be needed to account for the fact that not all the subpopulations would have diverged from the ancestral population simultaneously.

Chapter 4

Models Involving Bifurcating Phylogenetic Trees

This chapter will develop the model of genetic drift in the previous chapter to create a new model which includes more complex relationships between the subpopulations in the form of phylogenetic trees. Each subpopulation will undergo a number of different periods of genetic drift since their common ancestral population, sharing all but the last period of drift in common with other subpopulations. After explaining what phylogenetic trees are and their origin in more detail, the chapter will move on to describing how a model that incorporates them could be built. It will describe the problems that arise from attempting to use the Balding–Nichols model of drift in this context and explain why the Nicholson–Donnelly one is preferred and the importance of the latter's ability to take fixation into account. Some of the properties of the rectified normal distribution used by that drift model will be discussed. After describing the results of using the model on simulated data, the model will be used on the HapMap dataset. Issues arising from examining standardised residuals and observing the effects on these of different choices of prior distributions in the model will be discussed. The use of post predictive checking to evaluate how well the model fits the data will be introduced.

It will be explained how the results of examining the post predictive checks will motivate the further development of the model in the following chapter.

4.1 Phylogenetic Trees

The concept of a phylogenetic tree is as old as the theory of evolution itself. One of the earliest examples of what would now be recognised as a rudimentary phylogenetic tree diagram appears in Darwin (1859)'s *The Origin of Species* (Ch4 pp116-117). The idea is to represent the evolutionary relationship between present-day living organisms by showing their relationships to their common ancestors in the form of a tree diagram. Darwin's theories have since been melded with those of Mendelian genetics so that the bifurcations (or, occasionally, multifurcations) that occur at each node refer to a genetic differentiation in terms of genotype, rather than only diffences in appearance or phenotype.

The "phylo-" of "phylogenetic" refers to phyla, a particular level of biological classification into which groups of organisms are arranged and alludes to the most common use of phylogenetic trees, which is to show the genetic relationships between different species and their ancestors. Figure 4.1, shows a simple version of such a tree. The root of the tree represents the common ancestor of the fly, the mouse and the human. A bifurcation event occurs where the species that will evolve into the modern-day fly branches off from the species that will evolve into the most recent common ancestor of mice and humans. The species that is the common ancestor of mice and humans but not flies is located at the next bifurcation or node. From there, the species that will evolve into modern-day mice and humans become genetically distinct and so are represented as seperate branches.

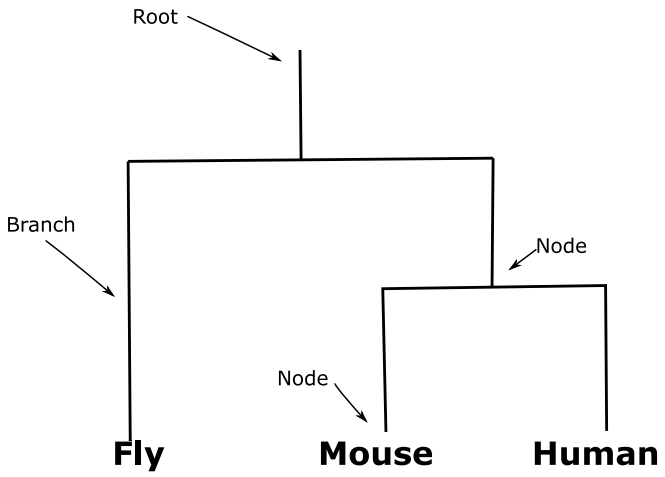

Figure 4.1: A Simple Phylogenetic Tree
Figure adapted from Theobald (2012).

Phylogenetic trees are usually used in this way to describe the ancestral relationships between different species. The bifurcation event at each node happens because the two organisms leading from it are genetically distinct to the extent that they can no longer interbreed and produce fertile offspring. In the context of this thesis, phylogenetic trees will be used to describe the relationships between different subpopulations of one species, that of humans. In doing this, there is a key difference. The subpopulations of humans can, of course, interbreed and produce fertile offspring. However, in general, this interbreeding has been halted by sorts of barrier, leading subpopulations to become somewhat genetically divergent. The most important barrier is geography. If members of two subpopulations of humans don't physically meet, they cannot produce offspring. Another cause might be cultural, where members of the subpopulations could physically meet but interbreeding has almost totally been prevented by a historical cultural or religious taboo. However, unlike different species, when different subpopulations of human diverge and become genetically differentiated from each other in this way, there does remain the possibility or even likelihood, that after some time, maybe hundreds or thousands of years, the subpopulations can meet again and interbreed once more. This situation is not considered further in this chapter but is considered in chapter 5. For the purposes of this chapter, the simplifying assump-

tion is made that the subpopulations do not interbreed after becoming genetically differentiated.

4.2 Applying the Neighbour–Joining Algorithm

In the previous chapter it was found that assuming that all the present-day subpopulations diverged from a common ancestor at roughly the same time, as did Nicholson et al. (2002), led to a model that produced an unsatisfactory fit to the data. Phylogenetic trees can be used to depict a more realistic and complex relationship between the present-day subpopulations and their common ancestors with some pairs of subpopulations becoming genetically differentiated more recently than others. One way to find a phylogenetic tree that might be appropriate for a given set of data is to use the Neighbour Joining Algorithm.

Applying that procedure to the HapMap data for all 22 autosomes produced the unrooted tree shown in figure 4.2. The estimated population pairwise F_{ST} values for each pair of subpopulations were calculated from (2.5) using the data set of all 22 autosomes.

The tree is unrooted but midpoint rooting (Swofford et al., 1996), taking the midpoint of the longest path through the tree as the root, places it midway along a path from the Tokyo Japanese (JPT) to the Nigerian Yoruba (YRI) (filled red circle in 4.2). If the data for each of the 22 autosomes are analysed individually, the resulting tree differs materially only in the placing on the non-African side of the tree of the Houston Gujaratis (GIH) and the Californian Mexicans (MEX). The clear split between Africans and non-Africans is consistent with the analysis in the previous chapter.

The Neighbour Joining algorithm used here is a simple one and one which a 21st century computer can calculate quickly. However, it comes without any estimate

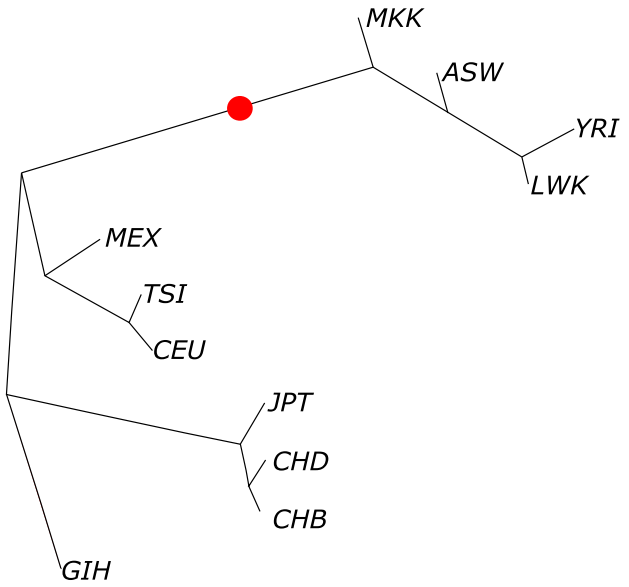

Figure 4.2: Unrooted Neighbour Joining Tree of HapMap Subpopulations, Where a Proposed Root is Shown With a Filled Red Circle

of uncertainty. The fact that the tree remains reasonably consistent when each chromosome is analysed individually suggests it is a tree in which a reasonable amount of confidence can be placed, the only doubt being over the correct placing of the GIH and MEX.

4.3 A Bifurcating Tree Model Incorporating the Balding–Nichols Drift Model

The next stage is to develop the hierarchical model to take the structure of the bifurcating population tree, such as the one produced by the Neighbour Joining algorithm into account, by adding additional parameters for ancestral subpopulations at the nodes of the tree. This produces an enhanced model. The drift model from Balding and Nichols (1995) has the virtue that its beta distributions are easier to work with compared to the rectified normal distributions from Nicholson et al. (2002) and so it was the one for which this new model was implemented first. The justifications for the priors and hyperparameters on π_i and c_j remain

the same as described in section 3.2.2.

Much of the model is similar to last time

$$x_{ij}|n_{ij},\alpha_{ij} \sim \text{Binomial}\,(n_{ij},\alpha_{ij}),\ \text{independently},$$

$$\alpha_{ij}|\pi_i,c_j \sim \text{Beta}\left(\tfrac{\pi_i(1-c_j)}{c_j},\tfrac{(1-\pi_i)(1-c_j)}{c_j}\right),\ \text{independently, for } \alpha\text{s nearest the root of the phylogenetic tree,}$$

$$\alpha_{ij}|\alpha_{ip},c_j \sim \text{Beta}\left(\tfrac{\alpha_{ip}(1-c_j)}{c_j},\tfrac{(1-\alpha_{ip})(1-c_j)}{c_j}\right),\ \text{independently, for other } \alpha\text{s,}$$

where α_{ip} is the alpha for the parent node to node j in the tree.

with priors

$$\pi_i|a \sim \text{Beta}\,(a,a),\ \text{independently},$$

$$c_j \sim \text{Beta}(b_{1j},b_{2j}),\ \text{independently},$$

where

i labels the locus: $1 \leqslant i \leqslant L$,

j labels the subpopulation $1 \leqslant j \leqslant P$,

n_{ij} is the total number of alleles observed at locus i in subpopulation j,

x_{ij} is the number of one of the two alleles observed at locus i in subpopulation j,

α_{ij} is the population proportion of that allele at locus i in subpopulation j,

π_i is the proportion of that allele at locus i in the ancestral population,

c_j is the amount of genetic drift in subpopulation j.

a is a hyperparameter in the prior of π_i.

b_{1j},b_{2j} are hyperparameters in the prior of c_j and assigned the value 1 unless otherwise stated.

4.3.1 Implementation of the Model and Full Conditionals

In contrast to the simpler version of the Balding–Nichols model described in Chapter 3, none of the parameters were integrated out. This was to simplify the initial programming task. Testing was done on a smaller model (DAG shown in figure 4.3) which contains all the ideas of the model but with fewer populations. The code was written in such a way that the model could be scaled up for any arbitrary tree. After testing an R version of the program, it was translated into C++. The faster running time of C++ made analysis of these larger models more feasible.

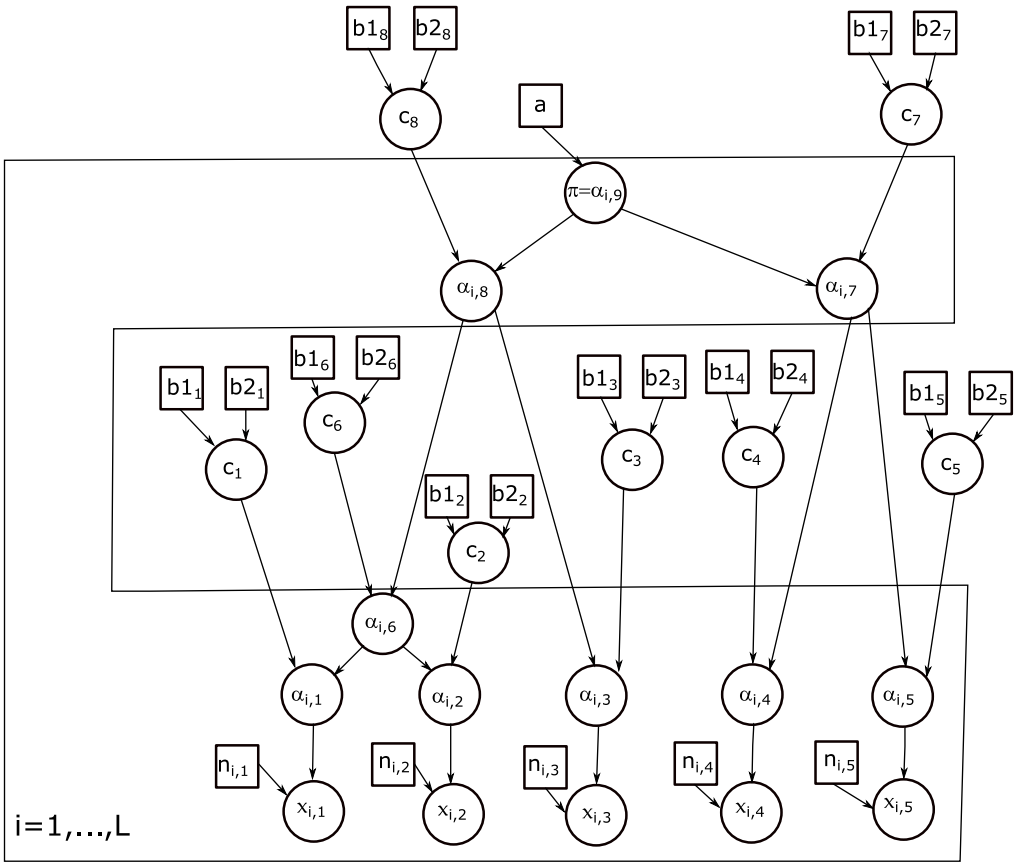

Figure 4.3: Directed Acyclic Graph of the Extended Bifurcating model
DAG of the Extended Bifurcating Balding–Nichols model.

In this smaller version, the second subscript of the α_{ij}s label the populations by starting from the bottom left of the DAG and filling in each level of the hierarchy before proceeding to the next. In this way, the α_{ij} where j is more than the

number of the observed subpopulations, J, is the proportion of the ith allele in the population that is ancestral to the two subpopulations below it in the hierarchy. π_i is a special case of α_{ij} where $j = 2J - 1$. In this case the full conditional for π_i is

$$P\left(\pi_i|\alpha, c, \pi_{-i}\right) \propto \pi_i^{a-1}\left(1-\pi_i\right)^{a-1} \prod_{k\in\Psi}\left(\frac{\alpha_{ik}^{\frac{\pi_i}{c_k}-\pi_i-1}\left(1-\alpha_{ik}\right)^{\frac{1}{c_k}-2-\frac{\pi_i}{c_k}+\pi_i}}{B\left(\frac{\pi_i(1-c_k)}{c_k}, \frac{(1-\pi_i)(1-c_k)}{c_k}\right)}\right), \quad (4.1)$$

where Ψ is the set of values of j for the two ancestral populations coming from the root.

The full conditional for c_j is

$$P\left(c_j|\pi, \alpha, c_{-j}, b\right) \propto \prod_{i=1}^{L}\left(\frac{\alpha_{ij}^{\frac{\alpha_{ip}}{c_j}-\alpha_{ip}-1}\left(1-\alpha_{ij}\right)^{\frac{1}{c_j}-2-\frac{\alpha_{ip}}{c_j}+\alpha_{ip}}}{B\left(\frac{\alpha_{ip}(1-c_j)}{c_j}, \frac{(1-\alpha_{ip})(1-c_j)}{c_j}\right)}\right)$$

$$\times c_j^{b_{1j}-1}\left(1-c_j\right)^{b_{2j}-1}, \quad (4.2)$$

where α_{ip} represents the allele frequency of the parent of α_{ij} (or π_i if $p = 2J - 1$).

There are two cases to treat for α. The first is where the offspring of the α is an x, that is, where $j \leqslant J$. In that case the full conditional is

$$P\left(\alpha_{ij}|\pi, \alpha_{-ij}, c, x, n\right) \propto \alpha_{ij}^{x_{ij}+\frac{\alpha_{ip}}{c_j}-\alpha_{ip}-1}\left(1-\alpha_{ij}\right)^{n_{ij}-x_{ij}+\frac{1}{c_j}-2-\frac{\alpha_{ip}}{c_j}+\alpha_{ip}}, \quad (4.3)$$

which is proportional to Beta$\left(x_{ij}+\frac{\alpha_{ip}}{c_j}-\alpha_{ip}, n_{ij}-x_{ij}+\frac{1}{c_j}-1-\frac{\alpha_{ip}}{c_j}+\alpha_{ip}\right)$ so can be sampled directly.

Finally, there is the case of α_{ij} where $j > J$. This is where there are two offspring of the α that are other αs. In this case the full conditional is

$$P\left(\alpha_{ij}|\pi,\alpha_{-ij},c\right) \propto \left[\prod_{k\in\Psi}\frac{\alpha_{ik}^{\frac{\alpha_{ij}}{c_k}-\alpha_{ij}-1}\left(1-\alpha_{ik}\right)^{\frac{1}{c_k}-2-\frac{\alpha_{ij}}{c_k}+\alpha_{ij}}}{B\left(\frac{\alpha_{ij}(1-c_k)}{c_k},\frac{(1-\alpha_{ij})(1-c_k)}{c_k}\right)}\right]$$

$$\times\,\alpha_{ij}^{\frac{\alpha_{ip}}{c_j}-\alpha_{ip}-1}\left(1-\alpha_{ij}\right)^{\frac{1}{c_j}-2-\frac{\alpha_{ip}}{c_j}+\alpha_{ip}}. \tag{4.4}$$

4.3.2 A Failure of the Model

To test the larger model, data were simulated under the assumptions of the Balding–Nichols model for 11 subpopulations of size 200 (100 individuals each) and 2400 loci. This kept the number of subpopulations, their sizes and the number of loci similar to those in the HAPMAP dataset for a long chromosome. However, the model was consistantly unable to recover the c_js. The parameter estimates produced by the bifurcating Balding–Nichols model were consistently lower than they should have been. This was the case even when all parameters other than the values of drift, c, were kept fixed at their true values. The inability to retrieve parameter values of data that have been generated under the model assumptions was a serious failure of the test. A typical example is shown in table 4.1.

Table 4.1: Underestimation of Drift Parameters by the Bifurcating Balding–Nichols Model.

	Bifurcating Balding–Nichols estimates of c_j			
	Central 95% Credible Value Interval Bounds		True Value	Underestimate
j	lower	upper		
1	0.0156	0.0174	0.0183	*
2	0.0144	0.0160	0.0163	*
3	0.0146	0.0163	0.0173	*
4	0.0188	0.0210	0.0201	
5	0.0207	0.0230	0.0249	*
6	0.0125	0.0139	0.0141	*
7	0.0176	0.0195	0.0201	*
8	0.0206	0.0229	0.0230	*
9	0.0130	0.0145	0.0151	*
10	0.0199	0.0221	0.0228	*
11	0.0106	0.0119	0.0128	*
12	0.0090	0.0101	0.0102	*
13	0.0114	0.0127	0.0131	*
14	0.0164	0.0182	0.0189	*
15	0.0171	0.0190	0.0178	
16	0.0215	0.0239	0.0237	
17	0.0207	0.0230	0.0223	
18	0.0098	0.0110	0.0103	
19	0.0179	0.0193	0.0186	
20	0.0179	0.0193	0.0186	

To understand why this was happening, it was necessary to consider the way in which genetic drift is modelled and some of the practical computational issues that arise from implementing the model, particularly in the case of rare alleles.

4.4 Comparison of Genetic Drift Models

4.4.1 Full Conditional for c in the Balding–Nichols Model

Recall from chapter 3 that the Balding–Nichols (BN) model is being used as an approximation to the Wright–Fisher (WF) model of genetic drift. There it was

shown that the BN model shares the property of the WF model that the expected future proportion of an allele under drift is the same as its last observed value. However, the WF and BN distributions are far from identical. The beta distribution of the BN model makes it impossible for an allele to die out entirely, whereas it can under the WF model, where an allele can become extinct or can become fixed. So the proportion of an allele can never be 0 or 1 under the BN model but it can under the WF model. Also importantly, the probability of a rare allele becoming still rarer under the BN model is greater than under the WF model. There is a strong tendency for rare alleles to become even rarer under the BN model but never actually die out. It is therefore interesting to look at the shape of the full conditional for the amount of drift c under such circumstances.

Recall from (4.2) that the full conditional for c under the bifurcating Balding-Nichols model was (for a single locus and a particular subpopulation)

$$P\left(c|\alpha_k, \alpha_p, b\right) \propto \frac{\alpha_k^{\frac{\alpha_p(1-c)}{c}-1}\left(1-\alpha_k\right)^{\frac{(1-\alpha_p)(1-c)}{c}-1}}{B\left(\frac{\alpha_p(1-c)}{c}, \frac{(1-\alpha_p)(1-c)}{c}\right)} \times c_j^{b_{1j}-1}\left(1-c_j\right)^{b_{2j}-1}. \tag{4.5}$$

Taking α_p, the proportion of the allele at the parent node to be 0.01, which is a rare but not extremely rare allele, and α_k, the proportion of the allele at the child node, to be 0.001, so that it has become rarer as is typical for this model, the full conditional for c is plotted in the top left of figure 4.4. Thinking about this in terms of point estimates, the BN model interprets this as a signal for a value of genetic drift close to c= 0.021. If α_k had been 0.0001, which is only a little smaller, the top right of figure 4.4 shows the model would have taken that as a signal for a larger maximum a-posteriori (MAP) estimate of c of about 0.047. However the point estimate of c in the previous example was less than half the size of this drift estimate. The point estimate of c increases greatly as α_k is decreased towards 0 by small amounts. If the allele has become almost as close to fixed as machine

precision can allow at a value of α_k of 10^{-300}, then this produces a signal of a very large amount of genetic drift with a MAP estimate of 0.87 as shown in the bottom left of figure 4.4. The result of this is that very small changes in the proportion α_k can lead to signals for much larger amounts of genetic drift c. However, this does not explain the *underestimates* of drift that were observed in the case of the bifurcating Balding–Nichols model. To understand this, it is necessary to consider the case where an allele has already become fixed. Because of the practical problem with the beta distribution being undefined for parameters of 0, an artificial barrier close to 0 had to be set for α. However, in reality any amount of drift from a value of α_k that has become fixed should leave it fixed because of the assumption that there are no mutations. In this situation, the lack of drift between two proportions which have become fixed should be uninformative about c. However, in the bifurcating Balding–Nichols model, wherever the artificial barrier is placed on α, the model instead takes this as a very strong signal indeed of no drift at all ($c = 0$) as shown in the bottom right of figure 4.4. There, the signal for a value of c very close to 0 is so strong that the values for c that have been displayed are just from 0 to 10^{-14}, rather than from 0 to 1 as in the other graphs, so that the behaviour can be seen. The number and strength of these signals for a low value of c are what lead to a downward bias in the estimate of the parameters for drift. This also explains why the underestimates were more likely to be observed for low values of j (i.e., near the leaves of the population hierarchy) because these are where the αs are most likely to have become proximate to 0 or 1 because they have drifted further from the ancestral proportion of the allele at the root.

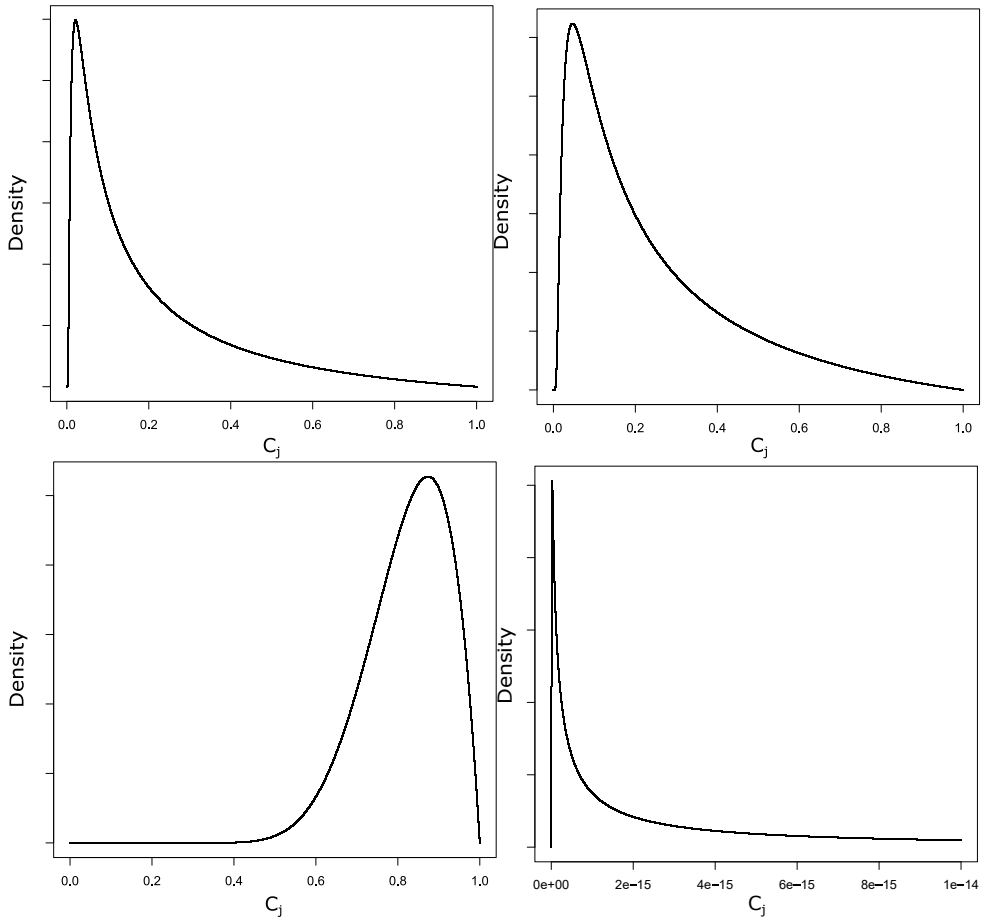

Figure 4.4: Balding–Nichols Full Conditional Examples.
Balding–Nichols Full Conditionals for c where (top left) α_p= 0.01 *and* α_k= 0.001*; (top right)* α_p= 0.01 *and* α_k= 0.0001*; (bottom left)* α_p= 0.01 *and* α_k =10^{-300} *;(bottom right, with different horizontal axis scale)* α_p = 10^{-16} *and* α_k= 10^{-16}*. The Balding–Nichols full conditional for c where* α_p *is the initial proportion of an allele at a locus before a period of genetic drift.* α_k *is the final proportion of an allele at that locus and c is the parameter encapsulating the amount of genetic drift between these two proportions. Here the maximum a-posteriori estimates of c are near c=0.021, c=0.047, c=0.87 and very close to c=0 respectively.*

4.4.2 The Full Conditional for c in an Equivalent Nicholson–Donnelly Model

To look for a way forward, the above results can be compared to the analogous full conditional for c that would have come from the ND model.

The full conditional for c in this case is

$$P\left(c|\alpha_k,\alpha_p,b\right)\propto\begin{cases}\left[c\alpha_p\left(1-\alpha_p\right)\right]^{-\frac{1}{2}}\int_{-\infty}^{0}\exp\left(\frac{-(r-\alpha_p)^2}{2c\alpha_p(1-\alpha_p)}\right)dr, & \alpha_k=0,\\ \left[c\alpha_p\left(1-\alpha_p\right)\right]^{-\frac{1}{2}}\exp\left(\frac{-(\alpha_k-\alpha_p)^2}{2c\alpha_p(1-\alpha_p)}\right), & 0<\alpha_k<1,\\ \left[c\alpha_p\left(1-\alpha_p\right)\right]^{-\frac{1}{2}}\int_{1}^{\infty}\exp\left(\frac{-(r-\alpha_p)^2}{2c\alpha_p(1-\alpha_p)}\right)dr, & \alpha_k=1.\end{cases}$$

$$\times c_j^{b_{1j}-1}\left(1-c_j\right)^{b_{2j}-1}. \tag{4.6}$$

This time, taking $\alpha_p = 0.01$ and $\alpha_k = 0.001$, the results of using this model can be seen in the top left of figure 4.5. This does give a signal for a particular value of genetic drift, with a MAP of 0.008. This time if α_k had been lower at 0.0001, as in the top right of figure 4.5, the model would have taken that as a signal for a similar MAP value for c of 0.010. So far the shapes of the full conditional are similar for both models of genetic drift with the ND model being more tolerant to the possibility that the value of the drift is larger than the MAP estimate. The big difference between these two models is that the ND model allows an allele to become either fixed or extinct. For $\alpha_k = 0$ the full conditional for c is shown in the bottom left of figure 4.5. This is a quite different shape compared with the previous plots. Although a small amount of genetic drift is considered unlikely, the model doesn't give a strong signal for any particular value of c, flattening off for all but the smallest values. This makes much more sense from the point of view of producing an analogy to the WF model than the comparable situation for the BN model (bottom left of figure 4.4). The fact that the allele has become extinct gives little information on how much drift there has been beyond there having been enough for it to have become extinct. For the situation of an allele that has already become extinct so that $\alpha_p = 0$ and $\alpha_k = 0$, the full conditional distribution is shown in the bottom right of figure 4.5. Here, the distribution has become perfectly flat. The situation is entirely uninformative about drift, as it should be. The problem of there being an unwanted strong signal for no drift in the roughly equivalent BN model (bottom right of figure 4.4) has been avoided.

The drawbacks of the ND model noted in chapter 3, namely that unlike the WF model, the expected future value of an allele frequency is not its last observed value and that rectified normal distributions are more awkward to work with than beta distributions, may be worth accepting in order to have a model that better reflects genetic drift in a way that would be expected from the WF model and avoids the pitfalls of having to set artificial (and unintentionally influential) thresholds near frequencies of 0 and 1.

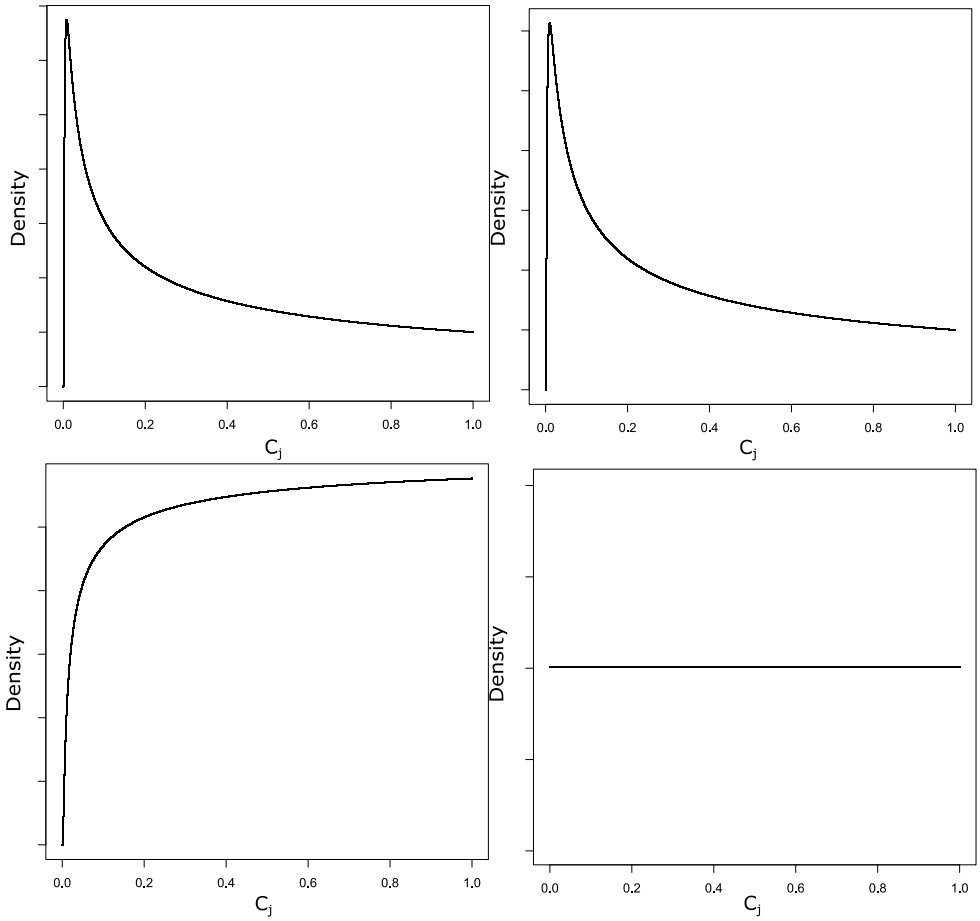

Figure 4.5: Nicholson–Donnelly Full Conditional Examples
Nicholson–Donnelly Full Conditional for c where (top left) α_p= 0.01 and α_k = 0.001; (top right) α_p= 0.01 and α_k = 0.0001; (bottom left) α_p = 0.01 and α_k = 0; (bottom right) α_p = 0 and α_k = 0. The Nicholson–Donnelly full conditional for c where $\alpha_p$is the initial proportion of an allele at a locus before a period of genetic drift. α_k is the final proportion of an allele at that locus and c is the parameter encapsulating the amount of genetic drift between these two proportions. Here the maximum a-posteriori estimates of c are near c=0.008 and c=0.010 in the top left and top right plots respectively.

4.4.3 Revisiting the Comparison of the Single Multifurcation Models of Chapter 3

The Balding–Nichols drift and Nicholson–Donnelly drift models from chapter 3 can now be re-examined. The Balding–Nichols model's inability to allow an allele to become fixed after a period of drift would be expected to lead to it producing higher estimates of drift than the Nicholson–Donnelly model in cases where only one allele is found at a locus for a particular subpopulation. The reason was described above in the commentary on figure 4.4, that the Balding–Nichols model produces high estimates of drift for rare alleles becoming rarer and much more so than the same change in α for a more common allele. On the other hand, the Nicholson–Donnelly model allows for the possibility that an allele can become fixed. Such a case only provides evidence that there must have been at least enough drift to take it there but little information about how much drift there might have been beyond that (bottom left of figure 4.5). The problem encountered with the bifurcating Balding–Nichols model of a fixed allele remaining fixed (figure 4.4 bottom right) cannot happen in this simpler model because the common parent population cannot have a fixed allele at any locus in the sample because it is not fixed in at least one of the subpopulations.

The key difference between the two models is in how they treat the situation where an allele is approaching becoming fixed. It would be expected that this situation would be most likely to arise for loci and subpopulations where $x_{ij} = 0$ or $x_{ij} = n_{ij}$, that is where the data contained only one of the two variant alleles. As the situation where α_{ij} is approaching 0 can arise in this simple model (the situations in the bottom left plots of figures 4.4 and 4.5) but not the ones where α_{ij} starts the period of drift at 0 (bottom right plots of figures 4.4 and 4.5), it would be expected that the Balding–Nichols model would tend to produce larger estimates of c_j than the Nicholson–Donnelly model where a large proportion of the data contains only one of the two possible variants. This is because as noted above,

the Balding–Nichols model interprets the situation where α_{ij} is approaching 0 as evidence for a large value of c_j. The proportion of loci where $x_{ij} = 0$ or $x_{ij} = n_{ij}$ was counted for each chromosome and subpopulation j. These are situations where it is possible for an α_{ij} to have reached or approached 0 or 1 respectively. This was plotted against the difference in the point estimates of c_j for each chromosome and subpopulation from the two models (with the difference calculated as the estimate from the Balding–Nichols model less the estimate from the Nicholson-Donnelly model), in figure 4.6.

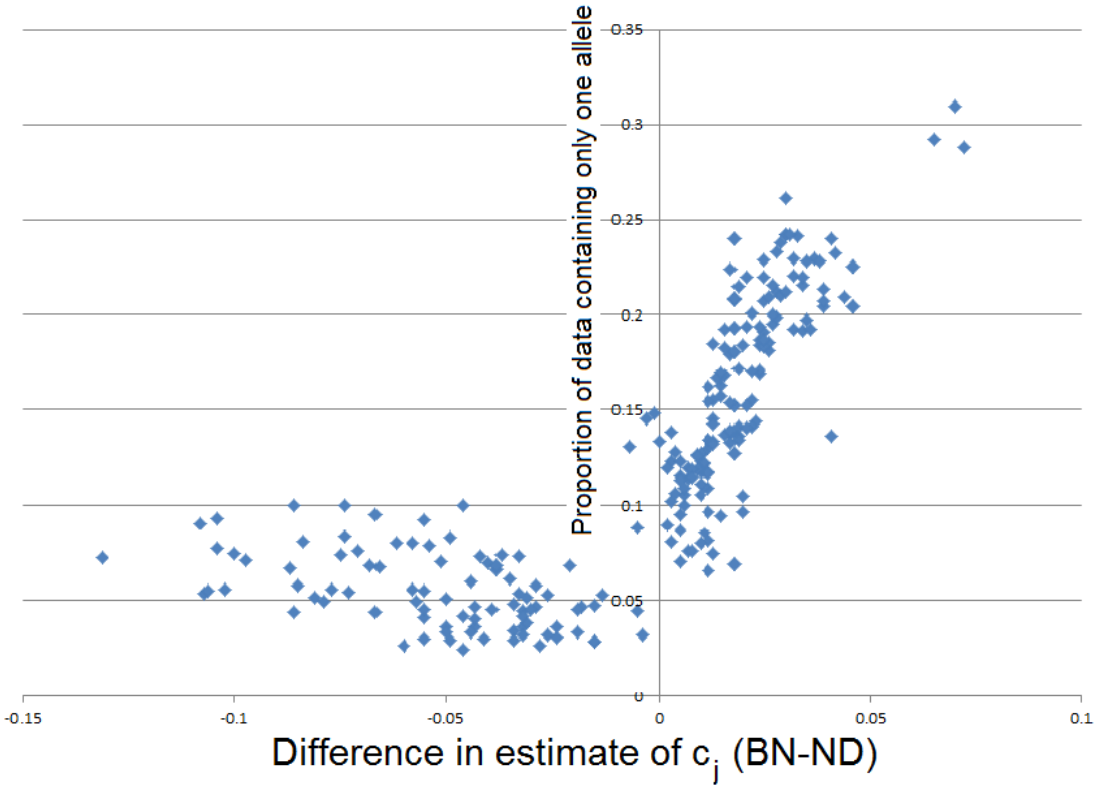

Figure 4.6: Proportion of the Data Containing Only One Allele for the Chromosome and Subpopulation against Difference in Estimate of c_j (Balding–Nichols minus Nicholson–Donnelly)

Proportions of the data for combinations of chromosome and subpopulation where only one allele variant was recorded is plotted against the difference in point estimate of the drift parameter, c_j (Balding–Nichols minus Nicholson–Donnelly) between the Balding–Nichols model and the Nicholson–Donnelly model. The graph illustrates the tendency of the Balding–Nichols model to produce higher estimates of drift where a larger proportion of the data has only one allele variant present. This is consistant with the observation that the Balding–Nichols model interprets situations where one of two allele variants is getting close to dying out in a subpopulation as being evidence for unrealistically high estimates of the drift parameter.

It clearly shows that where the Balding–Nichols estimate of drift is greater than the Nicholson–Donnelly estimate of drift, a greater proportion of the data for that chromosome and subpopulation has only one allele in the count, which is consistent with the explanation given above.

4.4.4 Conclusions from the Comparison

While, as noted in chapter 3, the ND model differs in key respects from the WF model, it is able to represent the extinction or fixation of an allele that is possible under the WF model and the resulting graphs of full conditionals for the drift for the contribution of a single allele make intuitive sense both at and near extinction and fixation. The BN model has some desirable properties but does not produce full conditionals that make sense in these cases. This is because the beta distribution only has support (0,1) and cannot take 0 as a parameter. The beta distributions that typically arise when one of the parameters is close to 0 have very steep gradients near 0 or 1 that lead the model to associate very small changes in the proportion of an allele with a very strong signal for a particular value of genetic drift. When the proportion of an allele starts and ends at a small value of similar magnitude for genetic drift, that signal is for a very small amount of genetic drift. Such a situation must arise because minimum values near 0 but not equal to 0 and maximum values near 1 but not equal to 1 must be set because of the granularity of the digital (usually 64 bit) representation of floating point numbers. 64 bit numbers can only represent 2^{64} points on the real number line. They can never represent the entire set of real numbers. Using more bits or a transformation from the 2^{64} points in $\mathbb{R}$ to 2^{64} points in $(0, 1)$ would just move the problem nearer 0 or 1 rather than solve it.

A number of attempts were made to overcome the problems with the bifurcating BN model. As described, thresholds had to be put in place on the αs to prevent the parameters of the beta distributions becoming computationally indistinguishable

from 0. If the α moved beyond that threshold during an updating step, it was reset to the threshold. Adjusting these thresholds did not help much. In an attempt to solve this problem a possible solution was tried whereby if an α_p had moved too close to 0 or 1, then the contribution of that locus to the full conditional for c was ignored. This was intended to reflect the idea that the allele at that locus had, to all intents and purposes, become fixed for that subpopulation and so should contribute no information to estimates of the genetic drift. However this was done at the expense of losing potentially useful information if the threshold was set too far from 0 or 1 and of not really making much difference if it was set too close to 0 or 1. The choice of such a threshold was arbitrary and although, by trial and error, an optimum level could have been found for a particular set of simulated data, there was not thought to be a way of guaranteeing that it would be the best choice for real data sets. The decision was therefore made to abandon the BN model and make an attempt to rebuild a similar model using Nicholson et al. (2002)'s rectified normal model, despite the analytical and computational complications that were expected to arise from it.

4.5 A Bifurcating or Multifurcating Tree Model Incorporating the Nicholson–Donnelly (ND) Model

4.5.1 Description of the Model

The modification to the model is in the way the αs are modelled. The justifications for the priors and hyperparameters on π_i and c_j remain the same as described in section 3.2.2.

$$x_{ij}|n_{ij}, \alpha_{ij} \sim \text{Binomial}\,(n_{ij}, \alpha_{ij}), \text{ independently,}$$

$$\alpha_{ij}|\pi_i, c_j \sim \mathrm{N}^{\mathrm{R}[0,1]}\left(\pi_i, \pi_i\left(1-\pi_i\right)c_j\right), \text{, independently, for } \alpha\text{s nearest the root of the phylogenetic tree,}$$

$$\alpha_{ij}|\alpha_{ipi}, c_j \sim \mathrm{N}^{\mathrm{R}[0,1]}\left(\alpha_{ip}, \alpha_{ipi}\left(1-\alpha_{ip}\right)c_j\right), \text{, independently, for other } \alpha\text{s,}$$

where α_{ip} is the alpha for the parent node to node j in the tree.

with priors

$$\pi_i|a \sim \mathrm{Beta}\left(a, a\right), \text{ independently,}$$

$$c_j \sim \mathrm{Beta}(b_{1j}, b_{2j}), \text{ independently,}$$

where

i labels the locus: $1 \leqslant i \leqslant L$,

j labels the subpopulation $1 \leqslant j \leqslant P$,

n_{ij} is the total number of alleles observed at locus i in subpopulation j,

x_{ij} is the number of one of the two alleles observed at locus i in subpopulation j,

α_{ij} is the population proportion of that allele at locus i in subpopulation j,

π_i is the proportion of that allele at locus i in the ancestral population,

c_j is the amount of genetic drift in subpopulation j.

a is a hyperparameter in the prior of π_i.

b_{1j}, b_{2j} are hyperparameters in the prior of c_j and assigned the value 1 unless otherwise stated.

4.5.2 Implementation of the Model

The DAG for this model would be similar to that shown in figure 4.3. However, this time the model would allow for the possibility of an ancestral population having more than two offspring populations. This would allow for more complex but parsimonious structures of relationships between the subpopulations and their ancestral populations to be modelled. In addition, it was believed that since there is little power to make inferences about the overall ancestral population, that the population at the root of the population tree should have more than two immediate offspring populations. Having three or more offspring populations would anchor it more firmly without loss of generality. This was hoped to lead to the sampler having better mixing. With the root ancestral population having only two offspring populations, it could be that the combined drift that has occured between that ancestral population and its two offspring populations can be estimated but uncertainty remains about how much of that total drift is attributable to each of the two branches. This would be a case of weak identifiability. The danger is that such a situation could lead to slow mixing as the Gibbs sampler tries to explore a ridge of combinations of drifts from the root ancestral population that have similar probability density but are at roughly 45 degrees to the parameter axes. It was thought that collapsing one of the edges corresponding to one of these two periods of drift would result in the root ancestral population becoming identified with one of its two previous offspring populations. It would retain the other offspring population but inherit at least two more offspring, resulting in it having at least three offspring which would be expected to ameliorate the identifiability problem. The drift that had occured in the collapsed edge would be expected to reappear in the other uncollapsed edge keeping the total amount of drift in the tree much the same. It turned out during testing that the model does not suffer so badly from the weak identifiability problem at its root and that artificially enforcing a trifurcation there did not lead to the genetic drift being transferred to other edges of the network in the way that was expected. So while the additional generality

was not needed for the reason that was expected, it remained possible to treat multifurcations using the model. While strictly speaking only bifurcations happen in phylogenetic trees, if two bifurcations happened in a short period of time and it cannot be confidently discerned which happened first, a trifurcation could be a reasonable approximation to the situation.

Changing to the ND drift model in this situation involves more than simply substituting a rectified normal distribution where there was a beta distribution before. The ND drift model allows the allele at a node to become fixed or extinct. This introduces a range of complications and new conditions to be set to prevent the Gibbs sampler updating the model in such a way that it becomes logically inconsistent. There is also the problem of evaluating when the proportion of an allele is in the atom (discrete part) of the rectified normal distribution (i.e., is 0 or 1) or the continuous part. For example, an allele can only become extinct at a node if the proportions of that allele below it towards the leaves of the tree (towards the data and away from the root) are also 0. Logically, it can't have become extinct and then reappeared again because the assumption of no mutation precludes a variant being reintroduced. Similarly if the proportion of an allele at a node is being updated but the proportion at the next node towards the root has already become fixed, then the proportion at that node must also remain fixed for similar reasons. These issues don't arise in the case of the BN drift model.

With that in mind, the full conditional for the ancestral proportion π_i is much the same as in the standard ND model:

$$P\left(\pi_i|\alpha, c, \pi_{-i}\right) \propto \pi_i^{a-1}\left(1-\pi_i\right)^{a-1} \prod_{m=1}^{s} g_1\left(c_{k_m}, \pi_i, \alpha_{ik_m}\right), \tag{4.7}$$

where $\{k_1, ..., k_s\}$ is the set of child nodes (k for "kinder" or children) of the ancestral node, s is the number of immediate offspring populations (number of members of

the set of child nodes) of the ancestral population, and

$$g_1(c_k, \pi_i, \alpha_{ik}) = \begin{cases} [c_k\pi_i(1-\pi_i)]^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\pi_i)^2}{2c_k\pi_i(1-\pi_i)}\right) dr, & \alpha_{ik} = 0, \\ [c_k\pi_i(1-\pi_i)]^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ik}-\pi_i)^2}{2c_k\pi_i(1-\pi_i)}\right), & 0 < \alpha_{ik} < 1, \\ [c_k\pi_i(1-\pi_i)]^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\pi_i)^2}{2c_k\pi_i(1-\pi_i)}\right) dr, & \alpha_{ik} = 1. \end{cases} \tag{4.8}$$

The full conditional for c_j is again similar,

$$P(c_j|\alpha, \pi, c_{-j}) \propto \prod_{i=1}^{L} g_2(c_j, \alpha_{ip}, \alpha_{ij}) \times c_j^{b_{1j}-1}(1-c_j)^{b_{2j}-1}, \tag{4.9}$$

where

$$g_2(c_j, \alpha_{ip}, \alpha_{ij}) = \begin{cases} [c_j\alpha_{ip}(1-\alpha_{ip})]^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right) dr, & \alpha_{ij} = 0, \\ [c_j\alpha_{ip}(1-\alpha_{ip})]^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ij}-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right), & 0 < \alpha_{ij} < 1, \\ [c_j\alpha_{ip}(1-\alpha_{ip})]^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right) dr, & \alpha_{ij} = 1. \end{cases} \tag{4.10}$$

L is the number of loci, and α_{ip} is the α for the ith locus from the parent node of j. In the case where j is one of the child nodes of the root then $\alpha_{ip} \equiv \pi_i$.

There are two cases for the full conditional for α_{ij}. The first case is the one where there are no further αs arising as offspring of the α in question. In this case there are the data, x_{ij}, n_{ij} below that α in the hierarchy and the full conditional is of a similar form to the one in the model of chapter 3:

$$P(\alpha_{ij}|c_j, \pi_i, \alpha_{-ij}, x_{ij}, n_{ij}) \propto h_1(n_{ij}, x_{ij}, \alpha_{ij})\, g_3(c_j, \alpha_{ip}, \alpha_{ij}), \tag{4.11}$$

where again, in the case where j is one of the child nodes of the root node, $\alpha_{ip} \equiv \pi_i$

and

$$g_3\left(c_j, \alpha_{ip}, \alpha_{ij}\right) = \begin{cases} \int_{-\infty}^{0} \exp\left(\frac{-(r-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right) dr, & \alpha_{ij} = 0, 0 < \alpha_{ip} < 1, \\ \exp\left(\frac{-(\alpha_{ij}-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right), & 0 < \alpha_{ij} < 1, 0 < \alpha_{ip} < 1, \\ \int_{1}^{\infty} \exp\left(\frac{-(r-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right) dr, & \alpha_{ij} = 1, 0 < \alpha_{ip} < 1, \\ 1, & \alpha_{ij} = 1, \alpha_{ip} = 1, \\ 1, & \alpha_{ij} = 0, \alpha_{ip} = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.12}$$

The novelty here is that, as discussed above, if the parental α_{ip} has become fixed (i.e., has the value 0 or 1), then α_{ij} must equal α_{ip}. Since it is assumed that there is no mutation, once an allele becomes fixed at one node in the hierarchy of populations, it must be fixed for all subpopulations that are offspring of that population.

Also, just as in the model described in chapter 3,

$$h_1\left(n_{ij}, x_{ij}, \alpha_{ij}\right) = \begin{cases} 1, & \alpha_{ij} = 0, x_{ij} = 0, \\ \alpha_{ij}^{x_{ij}} \left(1-\alpha_{ij}\right)^{n_{ij}-x_{ij}}, & 0 < \alpha_{ij} < 1, \\ 1, & \alpha_{ij} = 1, x_{ij} = n_{ij}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

The new element in the extended model is the other case of an α, which is one which has two or more other αs as children of its node in the hierarchy. Then,

$$P\left(\alpha_{ij} | c, \pi, \alpha_{-ij}\right) \propto h_2\left(\alpha_{ij}, \alpha_{-ij}, c\right) g_3\left(c_j, \alpha_{ip}, \alpha_{ij}\right), \tag{4.14}$$

where $g_3\left(c_j, \alpha_{ip}, \alpha_{ij}\right)$ is as in (4.12).

An important point to note here is that this α_{ij} can take the value 1 only if all s_j

of its child αs all have the value 1. Similarly, α_{ij} can take the value 0 only if all s_j of its child αs all have the value 0. In other cases, $h_2(\alpha_{ij}, \alpha_{-ij}, c)$ is the product of rectified normals. That is,

$$h_2(\alpha_{ij}, \alpha_{-ij}, c) = \begin{cases} 1 & \alpha_{ij} = 0, \alpha_{ik_1} = \alpha_{ik_2} = \ldots = \alpha_{ik_{s_j}} = 0, \\ 1 & \alpha_{ij} = 1, \alpha_{ik_1} = \alpha_{ik_2} = \ldots = \alpha_{ik_{s_j}} = 1, \\ \prod_{m=1}^{s_j} f(c_{k_m}, \alpha_{ij}, \alpha_{ik_m}) & 0 < \alpha_{ij} < 1, \\ 0 & \text{otherwise}, \end{cases} \tag{4.15}$$

where $\{k_1, ..., k_{s_j}\}$ is the set (of size s_j) of child nodes of the node (j) in question and

$$f(c_k, \alpha_{ij}, \alpha_{ik}) = \begin{cases} [c_k \alpha_{ij}(1-\alpha_{ij})]^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\alpha_{ij})^2}{2c_k\alpha_{ij}(1-\alpha_{ij})}\right) dr, & \alpha_{ik} = 0, \\ [c_k \alpha_{ij}(1-\alpha_{ij})]^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ik}-\alpha_{ij})^2}{2c_k\alpha_{ij}(1-\alpha_{ij})}\right), & 0 < \alpha_{ik} < 1, \\ [c_k \alpha_{ij}(1-\alpha_{ij})]^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\alpha_{ij})^2}{2c_k\alpha_{ij}(1-\alpha_{ij})}\right) dr, & \alpha_{ik} = 1. \end{cases} \tag{4.16}$$

The πs and cs could have been sampled by Metropolis–Hastings–within–Gibbs as before, but the complexity created by the possibility of an allele becoming fixed at various steps in the chain made it easiest to sample the αs by rejection sampling. The model was originally built to use Metropolis–Hastings but it was found during testing that proportions of alleles could become stuck at 0 or 1 for large numbers of consecutive iterations and so were in those states much more often than they should have been. This may have been due to an undetected error in the computer code rather than the result of an inherent property of Metropolis–Hastings. Nevertheless, since the problem was found to be remedied by using a system of rejection sampling this sampling method was preferred.

The above describes the full conditionals up to proportionality. However, for the problem of determining whether an α is in the probability atom (i.e., where $\alpha = 1$ or $\alpha = 0$) an actual probability is needed. At a given locus (for clarity

the i subscript will temporarily be dropped), there is no problem when the parent alpha, α_p, is in the atom. If $\alpha_p = 1$ then $\alpha_j = 1$ and if $\alpha_p = 0$ then $\alpha_j = 0$ because no mutation is assumed. Also if $0 < \alpha_p < 1$ and at least one of the child alphas is also not in the atom, $0 < \alpha_k < 1$, then α_j cannot be in an atom because, again, no mutation is assumed. So the problem arises only when all the child alphas $\alpha_{k_1}, \alpha_{k_2}, ..., \alpha_{k_{s_j}}$ of the alpha being updated, α_j, are all in the atom but the parent alpha is not, $0 < \alpha_p < 1$. Here a two-stage process is followed. First it is determined whether α_j is in the atom. Second. if it has been determined that it is not, then the usual sampling procedure is followed for choosing a value in the $(0, 1)$ interval. The first stage needs the probability of that event. To see how this probability is determined, suppose at node j that $\alpha_{k_1} = \alpha_{k_2} = \cdots = \alpha_{k_{s_j}} = 1$, all the child alphas were equal to 1 and the parent alpha $0 < \alpha_p < 1$. So here α_j can be in the $(0, 1)$ interval or it can be 1 but it cannot be 0. The chance of the drift between α_p and α_j, governed by c_j carrying α_j to 1 would be $y(\alpha_p, c_j) \equiv 1 - \Phi\left(\frac{1-\alpha_p}{\sqrt{\alpha_p(1-\alpha_p)c_j}}\right)$. However, that is not the only possibility. drift from the parent, α_p, may not have fixed α_j, but subsequent drift from α_j could have led to fixation for all of the child alphas. The probability density of the drift c_j carrying α_j to some value, $r \in (0, 1)$, is $v_1(\alpha_p, c_j, r) \equiv \frac{1}{\sqrt{\alpha_p(1-\alpha_p)c_j}}\phi\left(\frac{r-\alpha_p}{\sqrt{\alpha_p(1-\alpha_p)c_j}}\right)$ where ϕ is the standard normal distribution pdf. The probability of such a value, $\alpha_j = r$ then resulting in all the child alphas being 1 is then $v_2(c_{k_1}, ..., c_{k_{s_j}}, r) = \prod_{m=1}^{s_j}\left[1 - \Phi\left(\frac{1-r}{\sqrt{r(1-r)c_{k_m}}}\right)\right]$. The probability of α_j being 1 is then the chance of c_j carrying α_j to 1, divided by sum of the probabilities of all the possibilities,

$$Pr(\alpha_j = 1) = \frac{y(\alpha_p, c_j)}{y(\alpha_p, c_j) + \int_0^1 v_1(\alpha_p, c_j, r)\, v_2(c_{k_1}, ..., c_{k_{s_j}}, r)dr}. \tag{4.17}$$

The integral can be done numerically e.g., by using the trapezium method. Analogous reasoning can be used to obtain the probability of $\alpha_j = 0$ when $\alpha_{k_1} = \alpha_{k_2} = \cdots = \alpha_{k_{s_j}} = 0$ and $0 < \alpha_p < 1$.

Similar reasoning can be used in the cases of alphas which have no child alphas and which are instead adjacent to the data. Such an alpha(α_j) must be in the same atom as its parent alpha if that alpha is in the atom. Also α_j can only be 0 if $x_j = 0$ and can only be 1 if $x_j = n_j$. So the problem only arises when $0 < \alpha_p < 1$ and either $x_j = 0$ or $x_j = n_j$. Thinking of the case where $x_j = n_j$ and $0 < \alpha_p < 1$, $Pr(\alpha_j = 1)$ is the same as that described in equation 4.17 with $v_2(c_{k_1}, ..., c_{k_{s_j}}, r)$ replaced by r^{x_j}, the binomial probability that $x_j = n_j$ if $\alpha_j = r$.

The functions being sampled from could be very sharply peaked. This makes a simple rejection sampler inefficient because it will take many attempts before it hits a value in the peak. A more efficient rejection sampling scheme was devised and tested and found to be faster. The interval [0,1] was cut into a number of slices (with a default number of 100, the best performing of the three options tested in section 2.5.2 over what was thought to be the most likely function shapes). The values of the function at the left and right of each slice were recorded and the higher of the two values assigned to the slice. The maximum value of a narrowly peaked function could be higher than either edge for the slice containing the maximum. So the two slices sharing the edge at which the highest such value was found were divided again into a number (again defaulting to 100) of slices each and the values of the function at each of these points calculated, the highest of which was then assigned to the two slices. The values assigned to the 100 slices are added together and a total found. The proportion of the total then becomes the probability that slice will be selected. A random uniform(0,1) number is generated to select a slice. The selected slice then has a simple rejection sample performed within it by choosing a uniform random point within its width and a uniform random number up to the value that had been assigned to the slice. If this is a value that is less than the value of the function at that point then that becomes the sampled value. If it is more than the value of the function, the process repeats by randomly selecting a slice in the way described, again. Sometimes, particularly for the c parameters, the function can be extremely narrowly peaked close to 0. After a

particular number of iterations of the sampler (default is set to 7,000 within the burn-in period but not too early that the chain hasn't settled down at all or too late to place it too near the end of the burn-in period) the program checks to see if extremely low numbers are being sampled for that parameter . If they are, then the number of slices used is increased to 1,000 at each step instead of 100, for greater accuracy (also 1,000 is the largest number tested in section 2.5.2 and found to be effective for the most extremely sharply peaked functions. The sharpest peaks in full conditional functions for c were observed most often in testing at very low values),. The program was written to allow the numbers of slices and point at which the program checks for low values to be easily changed.

4.5.3 Results from Application to the HapMap Dataset

The 20 edges corresponding to periods of drift in the model are shown in figure 4.7 numbered 0 to 19. This tallies with the tree obtained from the neighbour joining algorithm in figure 4.2.

The mean genetic drift, c, and the 95% central credible interval from the posterior distribution for each chromosome are given in Tables C.1-C.22 in appendix C.

These results are fairly consistent with each other but there are still too many cases where the 95% credible intervals of genetic drift for a particular edge in the graph do not overlap for different chromosomes to be sensibly attributable to randomness. These multiple large tables of numbers are also difficult to digest. To aid interpretation, trees like figure 4.7 were produced using a custom-built JAVA program, but with the edge lengths proportional to the point estimate of genetic drift. Figure 4.8 is such a tree created by taking the weighted average of the point estimates for the 22 autosomes. The weightings used were the number of loci in the sample for each chromosome.

Some of the main features of the tree are unsurprising. The east Asian subpopu-

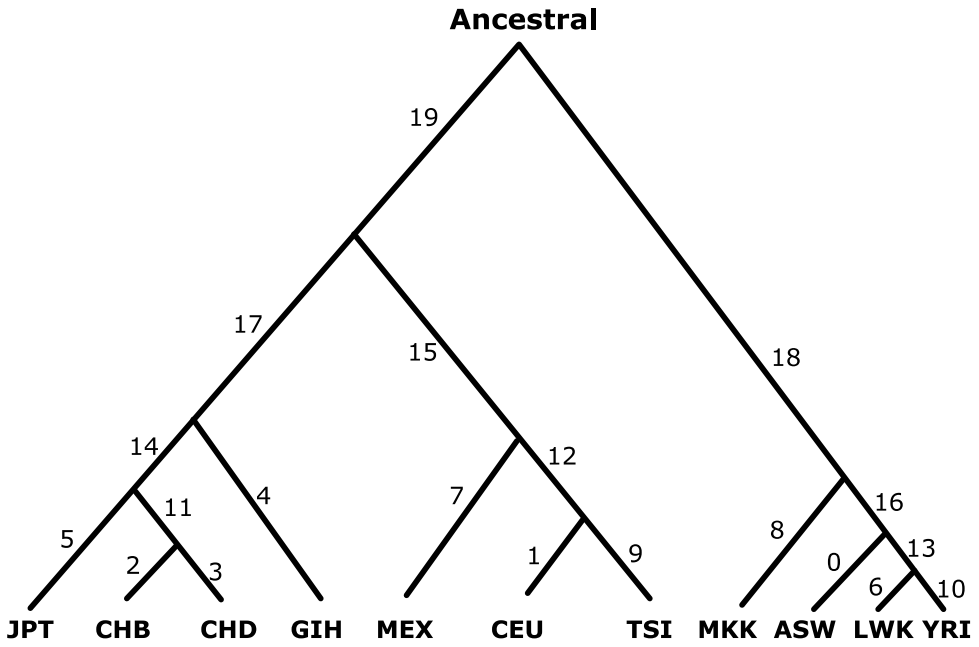

Figure 4.7: The Twenty Periods of Genetic Drift to be Modelled
The twenty periods of genetic drift to be modelled correspond to the edges in the graph and are numbered from 0 to 19. The ancestral population is at the top of the diagram and the present-day subpopulations in the HapMap dataset are shown at the bottom of the diagram.

lations (CHB, CHD, JPT) are arranged very close to each other with the two Han Chinese subpopulations (CHB, CHD) being so close together as to be almost indistinguishable. The two European subpopulations (CEU, TSI) are also very close to each other. The African subpopulations (YRI, LWK, ASW, MKK) are closer to each other than any other subpopulation but are more spread out than the two previous clusters, reflecting greater genetic differentiation among them. The remaining two subpopulations Mexicans (MEX) and Gujuratis (GIH) are nearer to the Europeans than they are to the other groups but not very close to them. In the case of the Mexicans, this is not particularly surprising because they are likely to have some European admixture from over five centuries of European colonial influences. One interesting feature is just how much genetic drift there is between the Europeans, Gujaratis and Mexicans and the east Asians. This possibly suggests that the east Asians are historically descended from a small number of individuals making up their ancestral population (a founder effect) between the nodes 17 and 14. There is also a lot of genetic drift between nodes 20 and 19. This would be

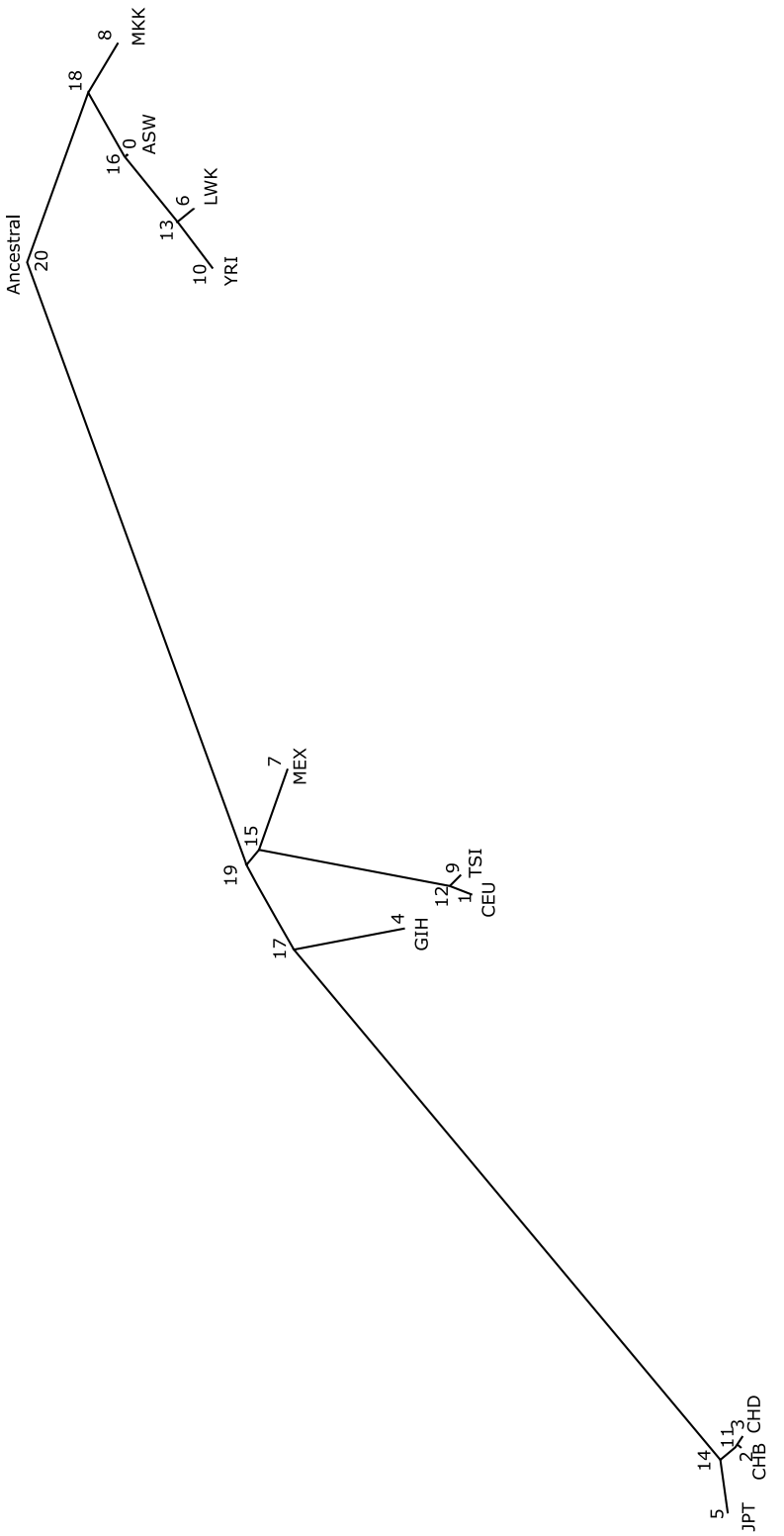

Figure 4.8: Phylogenetic Tree with Edge Lengths Proportional to Estimated Genetic Drift

The twenty periods of genetic drift for the HapMap dataset with the edge lengths proportional to the average (weighted by number of loci) of the posterior mean estimates of drift for the 22 autosomes.

consistent with the theory, such as that described by Macaulay et al. (2005) that all the populations outside of Africa are descended from a relatively small number of individuals who left Africa in a wave of migration 60,000-80,000 years ago, with all humans descended from an original ancestral population that was located somewhere in east Africa.

Many features of the model make story-telling sense but there are, nevertheless, some issues with it that will be explored in the following subsections.

4.5.4 Mixing and Convergence Issues

One of the problems that arose when looking at the Markov chains from the sampler was that in some cases the traces showed evidence of slow mixing and autocorrellation as in the left panel of figure 4.9. Examples of good mixing were also found as in right panel of figure 4.9. The problem of poor mixing was not universal but was found to be most apparent for the genetic drift parameter, c, when it was small. The sluggish mixing would need to be addressed in subsequent models.

4.5.4.1 Solution to the Problems Arising From Autocorrelation

The problem that slow mixing causes is that the resulting chain may be unrepresentative of the posterior distribution. This can be remedied by running the process for a greater number of iterations. However, this has practical difficulties. For the longest chromosomes, such as number 2, 20,000 iterations take approximately two days to obtain, even in C++. Running the process for longer would take proportionately longer. Nevertheless to show that this, in principle, provides a solution to the problem the chain was run for longer using a shorter chromosome, number 22.

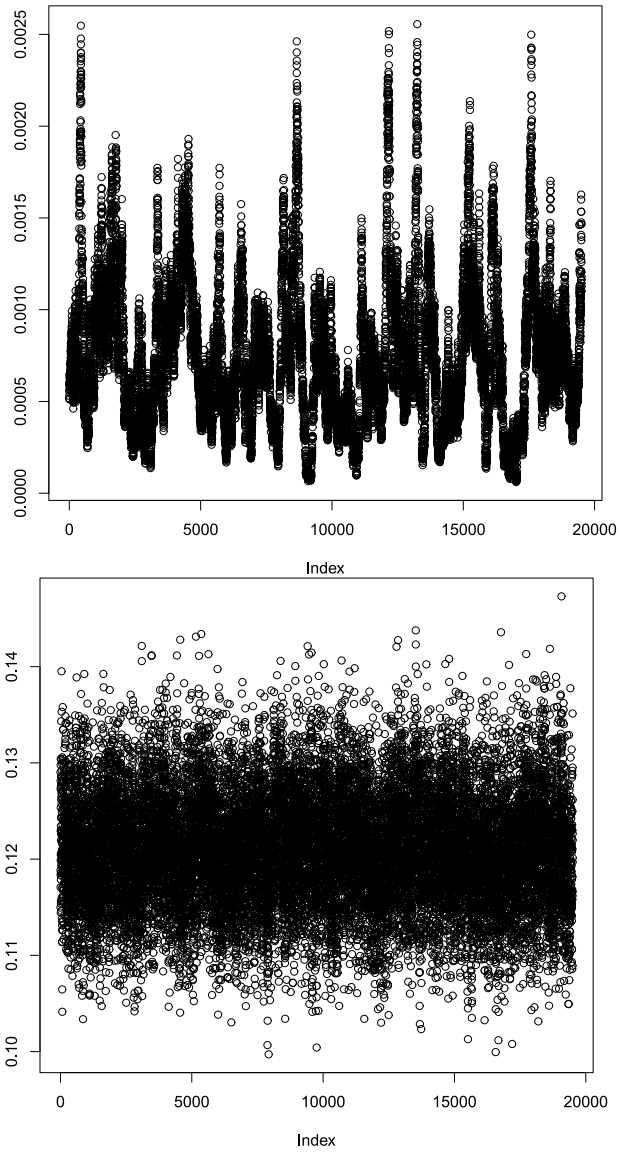

Figure 4.9: Traces of the chains produced by the Gibbs sampler for different periods of genetic drift.

Traces from the Gibbs sampler for two periods of genetic drift. The panel on the top shows an example of poor mixing which tended to occur when the estimated value of the drift parameter, c, was very small. The panel on the bottomt shows an example of good mixing characterised by rapid movement centred on a particular value.

The left hand plots in figure 4.10 show the trace plot and a histogram of the samples for the drift parameter, c_{15} after a 20,000 iteration chain. The first 10,000 of these were discarded as burn-in before the histogram below it was made. The chain was then allowed to continue for a further 80,000 iterations. The total being 100,000, an arbitrary round number that was found to be adequate. The right hand plots of

figure 4.10 show the trace plot and histogram for the same parameter after these additional iterations. It can be seen that the histogram for the longer chain is smoother, and although there is still considerable autocorrelation, there is some reassurance from the longer trace that the whole range of plausible values for the parameter has been explored. Nevertheless, the distribution after 20,000 iterations does provide a passable approximation to it. A longer chain would therefore be desirable for use by a final model. However, although not ideal, the finite amount of time available suggests that 20,000 iterations is adequate for the purposes of testing and evaluating the performance of the model at this stage.

4.5.5 Assessment of Model Fit

4.5.5.1 Standardisation of Residuals in the Context of Rectified Normal-Distribution-Based Models

To assess the fit of a model it is traditional to look at the values the model would predict at a data point ("fitted values") compared to the observed data at that point. However, simply looking at the difference between the two, the residual, is usually not enough. If it is assumed, as it usually is, that the residuals are normally distributed (at least approximately) then to compare the residuals, they need to be standardised. To standardise them, the residual is divided by its standard error. If they can be assumed to be approximately normally distributed, the distribution of these standardised residuals will be approximately $N(0,1)$ and so about 95% of them should be in the interval $[-2,2]$. This method of standardisation is used by Nicholson et al. (2002) for their simpler model. However, in that paper, the simplifying assumption was made when calculating these standardised residuals, without being explicitly stated, that the mean of the $N^{R[0,1]}(\mu,\sigma^2)$ remains μ and its variance is σ^2.

In fact, rectifying the distribution at 0 and 1 shifts the value of the mean to-

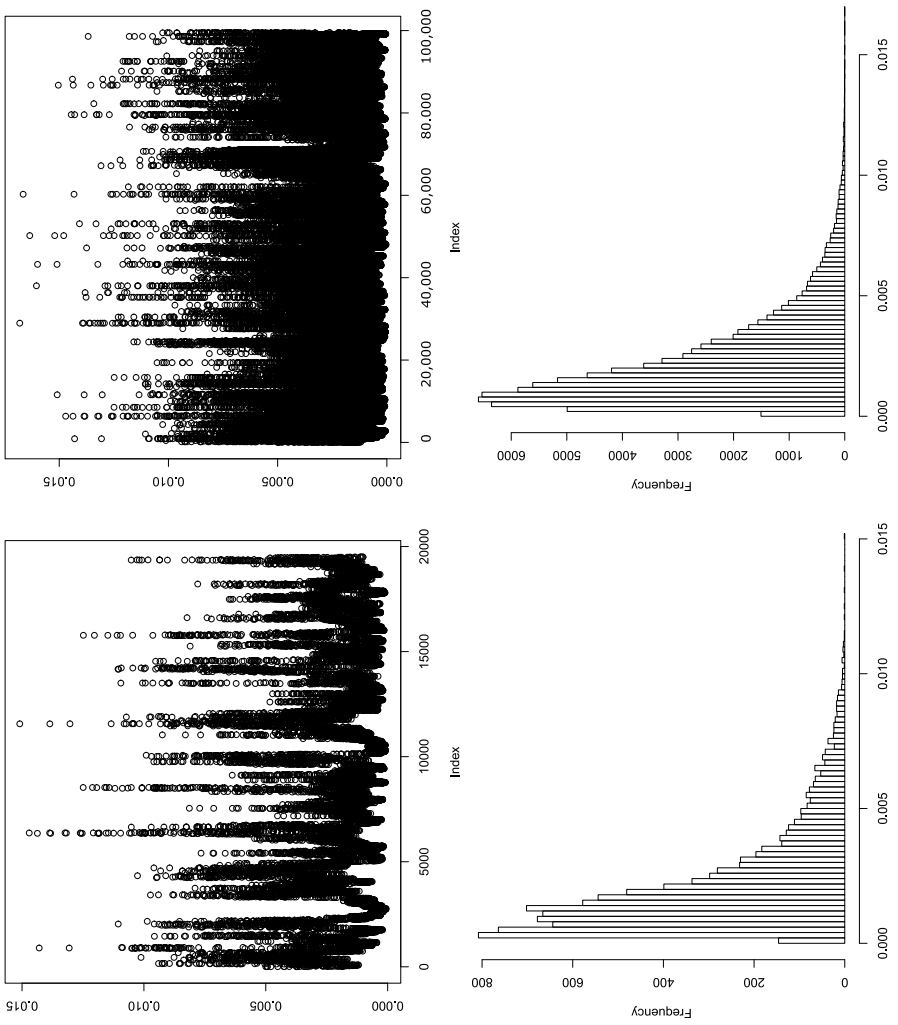

Figure 4.10: Trace plot and histogram of c_{15}, one of the smaller drift parameters in the model for Chromosome 22 at 20,000 and 100,000 iterations

A trace plot from the Gibbs sampler for a small period of drift parameter, c_{15} for chromosome 22 and the associated histogram for the first 20,000 iterations (left) and 100,000 iterations (right). At 20,000 iterations, there is sluggish mixing and it is not clear from the histogram what the shape of the posterior distribution is. At 100,000 iterations there is still sluggish mixing but the histogram shows a much smoother shape for the posterior distribution.

wards 0.5 and reduces the value of the variance compared to σ^2. It can be seen intuitively that the variance must be less than σ^2 because the range of values over which $N^{R[0,1]}(\mu, \sigma^2)$ can vary is restricted to $[0, 1]$ in this rectified normal distribution compared to $\mathbb{R}$ for the unrectified normal distribution. The variance of $N^{R[0,1]}(\mu, \sigma^2)$ approaches $\frac{1}{4}$ from below as $\sigma^2 \to \infty$. That the mean must be nearer $\frac{1}{2}$ than μ can also be understood intuitively. In the case where $\mu > \frac{1}{2}$, in the unrectified normal distribution, more probability density is above 1 than below 0. The act of rectifying the distribution at 0 and 1 "moves" all the probability above 1 to 1 and below 0 to 0. Since more probability has been reduced to value 1 than increased to value 0, the act of rectifying in this way must reduce the value of the mean. However, since the original distribution was a normal distribution, symmetric about μ and $\mu > \frac{1}{2}$, the resulting rectified normal distribution will still have a mean above $\frac{1}{2}$ because for any value δ where $0 \leq \delta \leq \frac{1}{2}$ there will be more probability density at $\frac{1}{2} + \delta$ than at $\frac{1}{2} - \delta$. Similar reasoning can be used in the case where $\mu < \frac{1}{2}$.

The upshot of this is that there are two ways of standardising the residual depending on which values are used for the mean and variance, approximate ones as used by Nicholson et al. (2002) or the "true" ones which have been derived for this thesis (Appendix B). In the context of a phylogenetic tree model, there are several periods of drift to take into account between the ancestral population and the observed allele counts rather than just one in Nicholson et al. (2002) and in chapter 3 of this thesis. However, if the simplifying approximation that Nicholson et al. (2002) make, namely that the mean of a rectified normal distribution is approximately μ and variance is approximately σ^2, is used then a general formula for d periods of genetic drift can be derived for the variance of the proportion of an allele at locus i:

$$Var\left(\frac{x_i}{n_i} \mid \pi_i, c\right) = \frac{\pi_i\left(1-\pi_i\right)}{n_i}\left[1 + \left(n_i - 1\right)\left\{1 - \left(1 - c_d\right)\left(1 - c_{d-1}\right)...\left(1 - c_1\right)\right\}.\right] \tag{4.18}$$

This formula was derived for the purposes of this thesis and a proof is provided in appendix A. Here $c_1, ..., c_d$ are the drift parameters in each of the d periods of genetic drift in series, π_i is the proportion of the allele at locus i in the ancestral population, x_i is the observed frequency of one of the alleles at locus i and n_i is the total number of both (or all) variants observed at locus i so that $\frac{x_i}{n_i}$ is the observed proportion of an allele at locus i. This is a simple formula and it can be used to estimate the size of a single period of genetic drift that is equivalent to d periods of genetic drift in series (at least in terms of variance):

$$c_s = 1 - (1 - c_d)(1 - c_{d-1}) \cdots (1 - c_1), \tag{4.19}$$

where c_s is the value of the equivalent effective genetic drift. Encouragingly, this formula also ties in with the discussion of the interpretation of the c parameter in section 3.3.3.

However, if the approximation cannot be made, and the more complicated expressions for the true mean and variance of the rectified normal distribution have to be used, the process of calculating the variance to be used in standardising the residuals after d periods of genetic drift is considerably more complicated and has to be done numerically. The question of whether the simplification can be made depends on how good an approximation to the true values of mean and variance μ and σ^2 make in situations that could realistically arise.

4.5.5.2 Differences Between Approximate Mean and Variance and True Mean and Variance in Rectified Normal Distributions

The problem then turns on how different the approximate means and variances are to the true means and variances over typical values of the genetic drift during a period of drift, c, from an ancestral proportion of the allele, π, at a given locus and subpopulation. The differences have been computed and are shown in Figure

4.11. It can be seen that the differences are smallest for small values of c, becoming larger as c increases. Values of c above 0.4 are not practically realistic and there is relatively little difference between the two over these values. The values of π where the difference is largest are near $\pi = 0.1$ and $\pi = 0.9$, with the difference always being 0 at $\pi = 0.5$. Combined this gives a maximum difference between the two of 0.0360 at $c = 0.4$ and $\pi = 0.904$ or 0.096.

The difference between the true and approximate variance was also calculated over a range of typical values (figure 4.12). Here the maximum difference, unsurprisingly, always occurs at $\pi = 0.5$ and increases with increasing c. Again, taking $c = 0.4$ to be the largest realistic value of that parameter, the maximum absolute difference between the true and approximate variances is 0.0191.

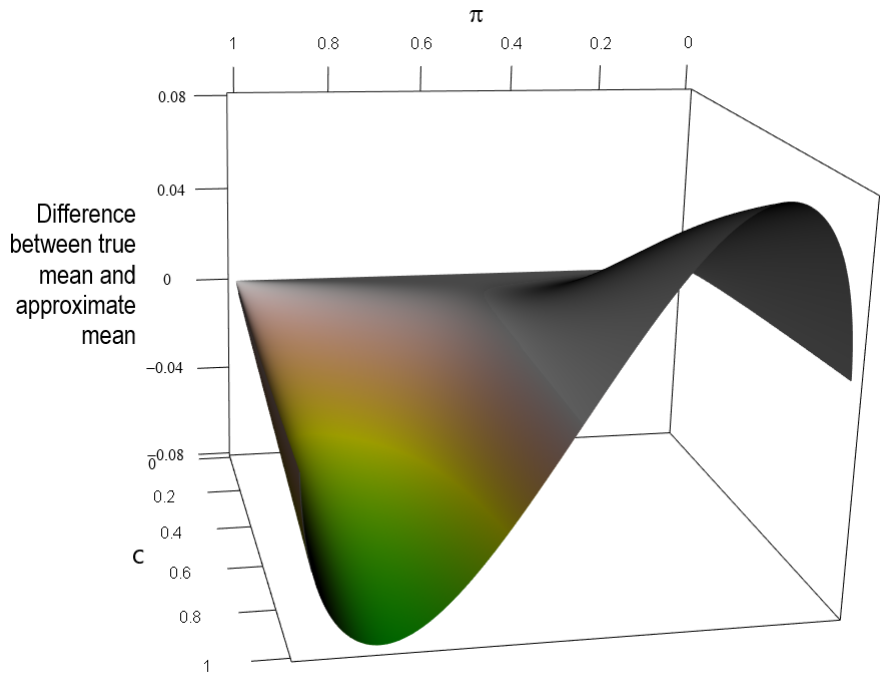

Figure 4.11: Surface showing the difference between the true and approximate means of $N^{R[0,1]}$ for different values of π and c

A surface showing the difference between the true and approximate values of the mean for the $N^{R[0,1]}(\pi, \pi(1-\pi)c)$ *distribution for different values of the ancestral allele proportion* π *and drift parameter* c*.*

Overall, the differences between the true means and variances and their approximations look sufficiently small to make little material difference. From this point

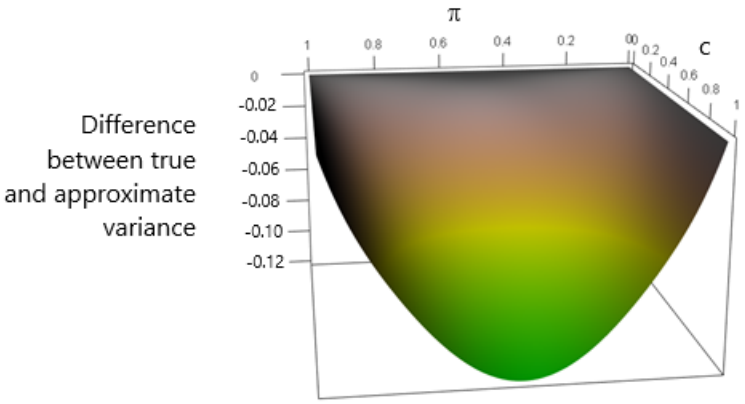

Figure 4.12: Surface showing the difference between the true and approximate variances of $N^{R[0,1]}$for different values of π and c

A surface showing the difference between the true and approximate values of the variance for the $N_{R[0,1]}(\pi, \pi(1-\pi)c)$ distribution for different values of the ancestral allele proportion π and the drift parameter c.

of view, the use of the approximate result by Nicholson et al. (2002) appears to be reasonable. A material difference would arise if examining standardised residuals using each of these methods of standardisation led to different conclusions about how well the model fitted the data.

4.5.5.3 Comparison of Standardised Residuals Using Different Methods of Standardisation

It is worth recounting the events which first motivated such an examination of the difference between the approximate mean and variance and the true mean and variance. During diagnostic checking of the model, Normal QQ plots such as those shown in Figure 4.13 were produced using residuals calculated from the approximate mean and variance.

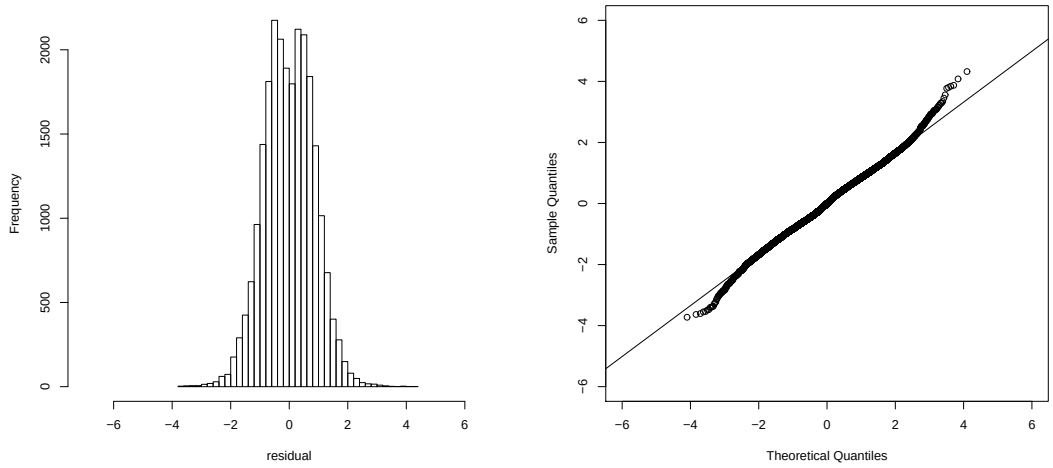

Figure 4.13: Plot of Standardised Residuals using the approximate mean and variance and QQ Plot for Chromosome 2

Diagnostic plots of standardised residuals for the bifurcating ND model for Chromosome 2. The histogram on the left shows a bimodal pattern suggesting that there are factors in the data that the model does not take sufficiently into account. The QQ plot on the right gives no cause for concern on its own.

The line in the QQ plot is not $y = x$, just as was the case in chapter 3. Once again, the variance of the standardised residuals is less than 1 because there is negative correlation between the P residuals associated with each locus, where P is the number of subpopulations. The QQ plot is consistent with the residuals being approximately normally distributed. Histograms of residuals were also produced. Many of these, such as that on the left of Figure 4.13, are still bimodal. The phylogenetic branching structure in the new model clearly had not solved the problem of bimodality of residuals that had, in part, motivated it. Discovering the reason for the bimodality needed further investigation. However, another puzzling pattern emerged when boxplots of the residuals were constructed by subpopulation. One of these is shown in Figure 4.14. An examination of the diagram reveals that the boxes for two of the subpopulations, the Maasai (MKK) and Afro-Americans (ASW) had smaller variances for their residuals than the other subpopulations. Conversely, the two Chinese (CHB and CHD) and the Japanese (JPT) subpopulations have a larger range for their residuals than the other subpopulations.

Referring back to Figure 4.8, what distinguishes ASW and MKK from the other subpopulations is the smaller number of periods of drift from the ancestral population and a lower overall amount of genetic drift. Conversely, those with the largest number of periods of drift and largest amount of drift, CHB, CHD and JPT, had the largest spread among their residuals.

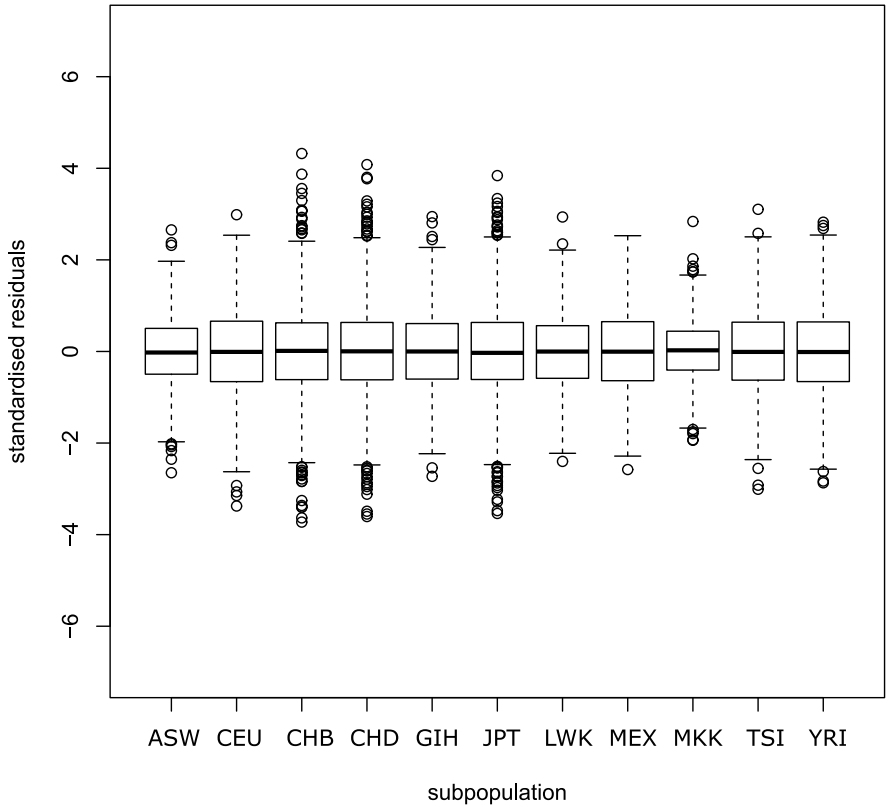

Figure 4.14: Boxplot of Standardised Residuals using the approximate mean and variance by Subpopulation for Chromosome 2

Diagnostic boxplots of standardised residuals for the bifurcating ND model for Chromosome 2. The boxplot shows that the least variance in the residuals occurs for subpopulations that experienced the lowest level of overall drift.

It was this observation that originally prompted a more careful examination of the difference between the approximate and true values of the means and variances of rectified normal distributions and whether the difference might be what was causing this pattern in the standardised residuals. If the use of the approximate values caused the unstandardised residuals of subpopulations which had experienced little genetic drift to be divided by too large a number and those which had

experienced a large amount of genetic drift to be divided by too small a number, it would explain the pattern and the problem would be remedied by using the true values of the mean and variance instead of the approximations. To test this idea, a small five subpopulation dataset with 1000 loci and subpopulation sizes 98, 224, 168, 170 and 176 respectively (49, 112, 84, 85 and 88 individuals respectively) was simulated from the assumption that one of the subpopulations had experienced a very much smaller amount of genetic drift than the other four. These were chosen to be the same as those observed in the HAPMAP dataset. Different sizes were chosen as in the HAPMAP dataset, rather than having them all the same, in case this made a difference. 1,000 loci would be similar to the number for a medium chromosome in the HAPMAP dataset. It was simulated under the non-branching version of the Nicholson–Donnelly model described in chapter 3. The reason for this was that if all the subpopulations experienced the same number of periods of genetic drift (one) and the phenomenon could be replicated, then the reason would be to do with the overall amount of genetic drift experienced by the subpopulations and not the number of periods of genetic drift. A residual plot for such an experiment is reproduced in Figure 4.15.

The intersting point to notice is that the spread of standardised residuals is smaller for population 4, which is the one for which the genetic drift was smallest (0.001 compared with 0.1 for the other 4 subpopulations). The phenomenon is therefore to do with the overall amount of genetic drift and not the number of periods of drift. The residuals were then standardised using the true means and variances of $N^{R[0,1]}$(Figure 4.16).

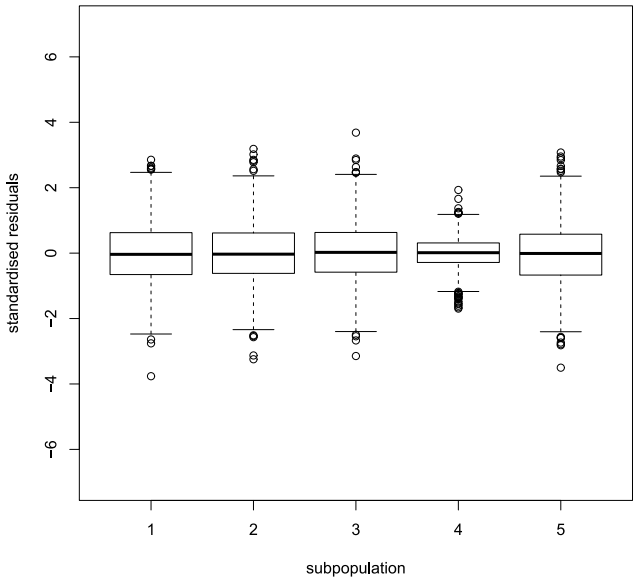

Figure 4.15: Boxplot of Standardised Residuals from Simulated Data using the approximate mean and variance

Diagnostic boxplots of standardised residuals for the bifurcating ND model for simulated data. The boxplot shows that the least variance in the residuals occurs for subpopulations that experienced the lowest level of overall drift.

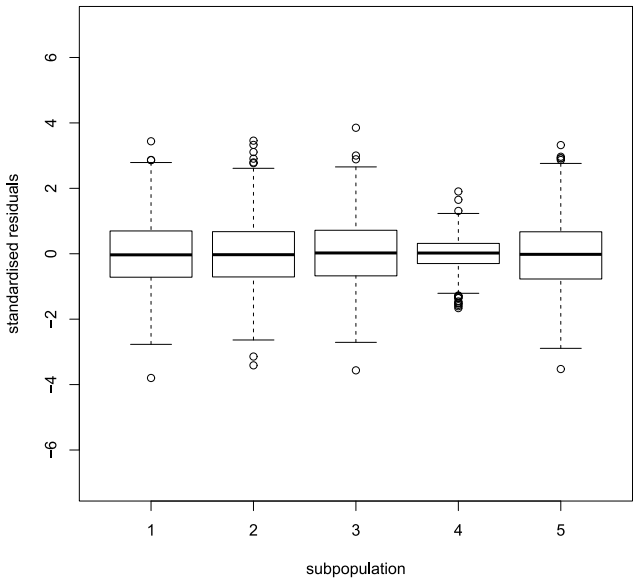

Figure 4.16: Boxplot of Standardised Residuals from Simulated Data using the true mean and variance

Diagnostic boxplots of standardised residuals for the ND model for simulated data. The residuals for the data using the true means and variances were not different enough from those using approximate means and variances to lead to different conclusions about the fit of the model.

This did not solve the problem of the subpopulation that experienced the least drift

having a narrower spread of residuals. The two plots (figures 4.15 and 4.16) are only slightly different, reflecting the slight change in the values of the standardised residuals but not materially so. The conclusions they lead to are the same. Once again, the use of residuals standardised by the approximate variance rather than the exact more complicated form does not make sufficient practical difference to justify the additional complexity.

4.5.5.4 Posterior Predictive Checking

The classical practice of examining residuals to consider how well the model was fitting the data did not appear helpful in this case. The usual assumption is that the boxplots of residuals by subpopulation should be approximately the same in each case. However, the spread of residuals for different subpopulations were related to the amount of genetic drift that the subpopulation had experienced. Should this be a cause for concern? Or is this simply a case where the usual classical assumptions should not be made? To answer this another approach to model checking was examined, posterior predictive checking.

The theory of posterior predictive checking can be found on pages 143-153 of Gelman et al. (2013). The idea behind it is quite simple and elegant. If the model fits the data well, then the data should not look grossly different to datasets simulated under the model assumptions and with typical values of its parameters. The Gibbs sampler gives a set of parameter values at each iteration. For each iteration, a dataset can be simulated under the model's assumptions using the parameter values at that iteration. This generates a number, τ, of simulated datasets equal to the number of iterations that the Gibbs sampler was run for after the burn-in period. Any data set can be summarised using a summary function T. If the summary value for the actual data is unlikely given the distribution of the summary of the simulated data, this is evidence of model misfit. The values of the T function for the τ simulated datasets and 1 true dataset can be ordered.

The proportion of simulated datasets whose value for T is larger than that of the true dataset in effect becomes a sort of p-value. If almost all of the simulated datasets produce a value of T that is higher than that for the true dataset, or if they almost all produce a value that is lower than that for the true dataset then the true dataset is out of place among the simulated datasets. This would not occur if the model was a good fit to the data. So, as for p-values in two tailed tests, proportions of the simulated datasets with T values higher than that of the data close to 1 or 0 are evidence for misfit, those near the middle of the [0,1] interval give no cause for concern. This method is more in keeping with the Bayesian philosophy of fully capturing uncertainty via the posterior distribution. The downside to this approach is that it can be quite computer intensive.

One important point is that the function T could be anything. So a decision needs to be made about what T should be. To decide this some thought needs to be given to what aspects of the data the model is trying to reflect. It is rare for models to need to be a fully accurate reflection of reality. The model needs only to be a good approximation to the properties of the problem that are of interest. The choice of T should, therefore, be determined in relation to and reflect these properties. More than one T can be examined. Gelman et al. (2013) states that in this case, multiple testing is not a problem because the process is not being used for model selection, but to test the limits of the applicability of the model.

Applying this approach to the question of deciding whether the spread of residuals in the ASW and MKK subpopulations (Figure 4.14) should be a cause for concern leads to a choice of T being the variance of the residuals for that subpopulation. It is one of the flexibilities of this approach that since T can be anything, T can be a function that applies to a subset of the data. It is then possible to calculate T functions for more than one subset of the data and so produce a number of different p-values. In this case a T has been defined that can be applied to each subpopulation, so a p-value will be produced for each subpopulation. The p-values arising from this definition of T are shown in Table 4.2. The p-values for the ASW

and MKK subpopulations, the ones which had the smaller spread of residuals in Figure 4.14, are 0.7402 and 0.8646, respectively. These are not extreme and so the small spread of those residuals is not a cause for concern. Nonetheless, there are some interesting aspects to the p-values. All of them are higher than 0.5 indicating that the spread of standardised residuals for the data tends to be higher than for the simulations and the highest p-values come from the two European subpopulations, CEU and TSI, which are closely related.

Subpopulation	p-Value
ASW	0.7402
CEU	0.9702
CHB	0.7402
CHD	0.7670
GIH	0.9485
JPT	0.7918
LWK	0.7338
MEX	0.9145
MKK	0.8646
TSI	0.9683
YRI	0.7300

Table 4.2: p-values for Variances of Residuals Produced from Post Predictive Checking *Results of post predictive checking of the variances of residuals shown in Figure 4.14. The small spread of residuals for the ASW and MKK subpopulations does not cause concern under post predictive checking.*

To examine this further, a more specific question needs to be asked and an appropriate function, T, defined to answer it. The intention of the phylogenetic tree model is to reflect the relationships of the subpopulations between each other and the genetic drift they have experienced since the time of the common ancestral population. A natural choice then is to take T as Wright's pairwise F_{ST} (Wright, 1951). A T function can then be calculated for each pair of subpopulations to see if the relationship between each pair in the τ simulated datasets is similar to the relationships in the actual data. If the p-value is close to 0 then it means that the subpopulations are more closely related in the data than in the simulated datasets and therefore more closely related than the model suggests. Conversely, a p-value

close to 1 would mean that the subpopulations are less closely related in the data than in the simulated datasets. The results for a typical chromosome (chromosome 2) are shown in Table 4.3.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0	0.0077	0.0062	0	0.0048	0.8666	0.0007	0.9514	0	0.0003
CEU	0	X	0.7635	0.9950	0.0598	0.9410	0.9977	0.6643	0.7354	0.6144	1
CHB	0.0077	0.7635	X	0.1456	0.6596	0.2282	0.8546	0.0001	0.3981	0.9125	0.9982
CHD	0.0062	0.9950	0.1456	X	0.9005	0.6566	0.7947	0.0009	0.3836	0.9958	0.9963
GIH	0	0.0598	0.6596	0.9005	X	0.8756	0.8269	0.9963	0.1498	0.0078	0.9987
JPT	0.0048	0.9410	0.2282	0.6566	0.8756	X	0.8205	0.0002	0.3909	0.9743	0.9981
LWK	0.8666	0.9977	0.8546	0.7947	0.8269	0.8205	X	0.9774	0	0.8593	0.5767
MEX	0.0007	0.6643	0.0001	0.0009	0.9963	0.0002	0.9774	X	0.7528	0.9650	0.9999
MKK	0.9514	0.7354	0.3981	0.3836	0.1498	0.3909	0	0.7528	X	0.0074	0.9420
TSI	0	0.6144	0.9125	0.9958	0.0078	0.9743	0.8593	0.9650	0.0074	X	0.9986
YRI	0.0003	1	0.9982	0.9963	0.9987	0.9981	0.5767	0.9999	0.9420	0.9986	X

Table 4.3: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Phylogenetic Tree Model of Chromosome 2

Results of post predictive checking of the p-values for pairwise F_{ST} for each pair of subpopulations produced from post predictive checking of the phylogenetic tree model of Chromosome 2 presented as a symmetric matrix. Numbers near 0 indicate that the subpopulations are more closely related in the data than in the model. Numbers close to 1 indicate the opposite.

Here it can be seen that the model does not provide a good fit to the data in this

respect but the ways in which it does not are interesting. For example, the p-values indicate that the Afro-Americans (ASW) are more closely related to the European subpopulations (CEU, TSI) than the model allows. This may not be too surprising. Most of today's African Americans in the south-west USA are descended from slaves. There are many stories of the sexual exploitation of Afro-American slaves during the era when slavery was legal in the USA. For example, Marable (1999) tells of instances of sexual relations between master and slave occurring to produce slave children for profit. Many of today's Afro-Americans, including those in the HapMap sample, could thus well have some European ancestry. ASW would therefore be an example of an admixed population, a subpopulation produced by the mixing of two or more other subpopulations. There are also low values for the pairing of ASW with Asian subpopulations (CHB, CHD, GIH, JPT) but also with MEX. This could be just because the model poorly describes the ancestry of the ASW subpopulation but it may be that there is also a possibility of Native American ancestry too. That they should be closer to the YRI (Yoruba in Nigeria) subpopulation perhaps reflects that Nigeria is part of the coast of Africa from which slaves bound for America were traded. The Mexicans (MEX) are also interesting. They have low p-values for the pairs with the east Asian subpopulations (CHB, CHD, JPT). In this model, MEX is placed near the European subpopulations (CEU, TSI) but these p-values suggest that MEX should also be nearer the East Asian subpopulations than the model allows. This again could be because the Mexicans are an admixed population. Genetic and archaeological studies such as that by Rasmussen et al. (2014) point to Asians having crossed into North America between 10,000 and 15,000 years ago. These would then have migrated south to present-day Mexico over time. They were later met by European settlers from the time of Columbus, 1492, onwards forming the admixture population that is today's Mexicans. There are also a few points that are difficult to explain. That the p-values suggest the Maasai (MKK) should be closer to the Tuscans (TSI) but not the central Europeans (CEU) is puzzling. The p-values also suggest that the Gujaratis (GIH) should be more closely related to the Europeans (CEU, TSI)

and Afro-Americans. It may be that all these subpopulations are more closely related to another subpopulation that is not in the HapMap dataset such as an Arab subpopulation, rather than directly to each other but without a sample from such a subpopulation to test the idea, this cannot be more than speculation.

The upshot of this is that this analysis highlights that the assumption of the phylogenetic tree model that once subpopulations split, they never have contact with each other or any other subpopulation ever again and develop in isolation is an unrealistic one. The model does not capture the relationships between the subpoplations adequately. To represent the HapMap data more accurately, it is necessary to model subpopulations meeting and merging to form admixed subpopulations. This increases the complexity of the model but it is necessary to explain the data.

This type of analysis can also be used to show that the simpler model from the previous chapter where all subpopulations diverge from the ancestral population simultaneously was also inadequate. Table 4.4 shows the matrix of p-values that would have been produced by the simple model with ND drift from the previous chapter.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.9996	1	1	0.9999	0.9999	0	1	0	1	0
CEU	0.9996	X	0.9862	0.9883	0	0.9772	0.9999	0	0.9963	0	0.9998
CHB	1	0.9862	X	0	0.3924	0	0.9992	0.0170	0.9980	0.9916	0.9978
CHD	1	0.9883	0	X	0.3860	0	0.9993	0.0220	0.9975	0.9915	0.9976
GIH	0.9999	0	0.3924	0.3860	X	0.3445	0.9995	0.0010	0.9960	0	0.9994
JPT	0.9999	0.9772	0	0	0.3445	X	0.9982	0.0083	0.9957	0.9855	0.9963
LWK	0	0.9999	0.9992	0.9993	0.9995	0.9982	X	1	0	0.9999	0
MEX	1	0	0.0170	0.0220	0.0010	0.0083	1	X	1	0	1
MKK	0	0.9963	0.9980	0.9975	0.9960	0.9957	0	1	X	0.9927	0
TSI	1	0	0.9916	0.9915	0	0.9855	0.9999	0	0.9927	X	1
YRI	0	0.9998	0.9978	0.9976	0.9994	0.9963	0	1	0	1	X

Table 4.4: p-values for Pairwise F_{ST} for Each Pair of Subpopulations from Post Predictive Checking of the Simple Model for Chromosome 2

Results of post predictive checking of the p-values for pairwise F_{ST} for each pair of subpopulations produced from post predictive checking of the simple model with the Nicholson-Donnelly drift model from the previous chapter for Chromosome 2 presented as a symmetric matrix. Numbers near 0 indicate that the subpopulations are more closely related in the data than in the model. Numbers close to 1 indicate the opposite.

It can be seen that the p-values suggest the subpopulations form three groups of subpopulations that should be more closely related than the simple model suggests. There is a group of Africans (ASW, LWK, MKK, YRI), a group of east Asians and Mexicans (CHB, CHD, JPT, MEX) and a group of Europeans, Gujaratis and

Mexicans (CEU, GIH, MEX, TSI). These are distinct except the Mexicans appear in both of the last two groups. This surely reflects them needing to be more closely related to both Europeans and to east Asians due to the admixed nature of the subpopulation, as described above. This structure is reflected in the phylogenetic tree model that has been the subject of this chapter. The appearance of more extreme values in table 4.4 compared with table 4.3 reflects that the phylogenetic tree model is a better fit to the data than the model of the last chapter, so while the former is not perfect by any means, it does represent progress towards a better model of the dataset.

4.5.6 Sensitivity to Alternative Choices of Prior

In Bayesian modelling it is useful to find out what effect the assumptions about the prior have on the fitted model and, in particular, the conclusion that would be drawn from it. In this case an assumption is made about the distribution of proportions of alleles in the ancestral population. These are the π_is in the model. It has been assumed, that these have a Beta(1,1) distribution and this value was also used in Nicholson et al. (2002). The effect of varying the parameter in the beta distribution was examined, with two interesting results.

To illustrate the first of these, the results from the model on one of the chromosomes can be examined. This is presented in the form of a graph of the phylogenetic tree with the edge lengths proportional to the point estimates of the size of drift along them. The result for the Beta(1,1) prior is shown in Figure 4.17. Compare that with the figure that is produced when an alternative uninformative prior, Beta(0.5,0.5), which is the Jeffreys prior in this case, is assumed. This is shown in Figure 4.18. There is no material difference in the results with the notable exception that the ancestral population has moved closer to node 18, which is the ancestral population for the African subpopulations.

Next, if the parameter of the beta distribution is increased, so that the prior on π is now Beta(2,2), a bell-shaped distribution, the ancestral population moves away from the African ancestor at node 18. This is shown in Figure 4.19. Again it has no material effect on the rest of the tree. What is happening here is that the more u-shaped the prior distribution, the more the ancestral population becomes closer to the African ancestral population and so needs less genetic drift to get it there. However, flatter and more bell-shaped distributions make it more like the ancestral population for the rest of the world at node 19. But this is only true up to a point. If a stronger bell-shaped prior like Beta(10,10) is used, the ancestral population becomes more unlike either of its two offspring populations at 18 and 19. The result of this can be seen in Figure 4.20.

It is important to note that the scale of the tree has been changed in figure 4.20 to half the size of that of its predecessors to fit on the page. Now the ancestral population has become so unlike its two offspring populations with its allele frequencies being dragged towards 0.5 by the prior that it takes a lot of genetic drift for it to reach either of them. One of the effects of genetic drift is to make the distribution of allele proportions more u-shaped and less bell-shaped as was shown in chapter 3. A lot of drift is needed to flatten out the marked bell-shape of the Beta(10,10) distribution.

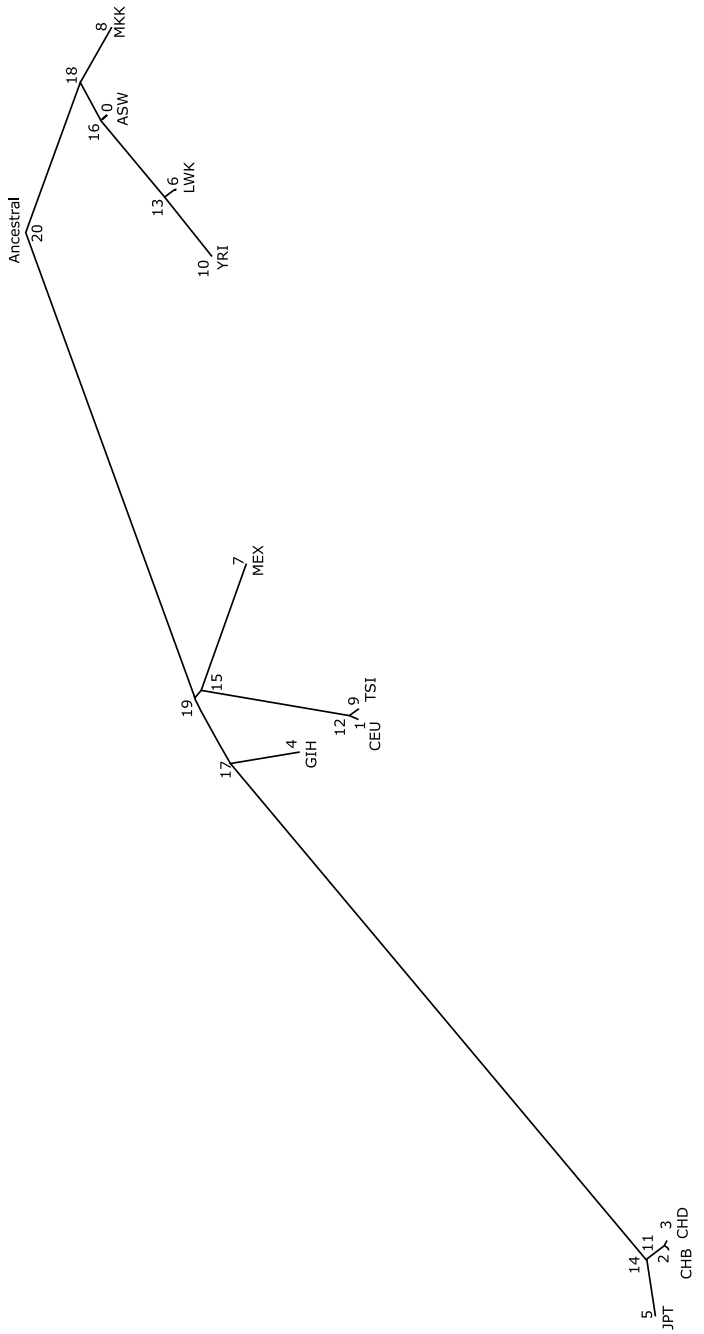

Figure 4.17: Phylogenetic Tree with Edge Lengths Proportional to Estimated Genetic Drift for chromosome 22 with a Beta (1,1) prior on π

The twenty periods of genetic drift for the HapMap dataset with the edge lengths proportional to the posterior mean estimates of drift for chromosome 22. A Beta(1,1) prior on π is assumed.

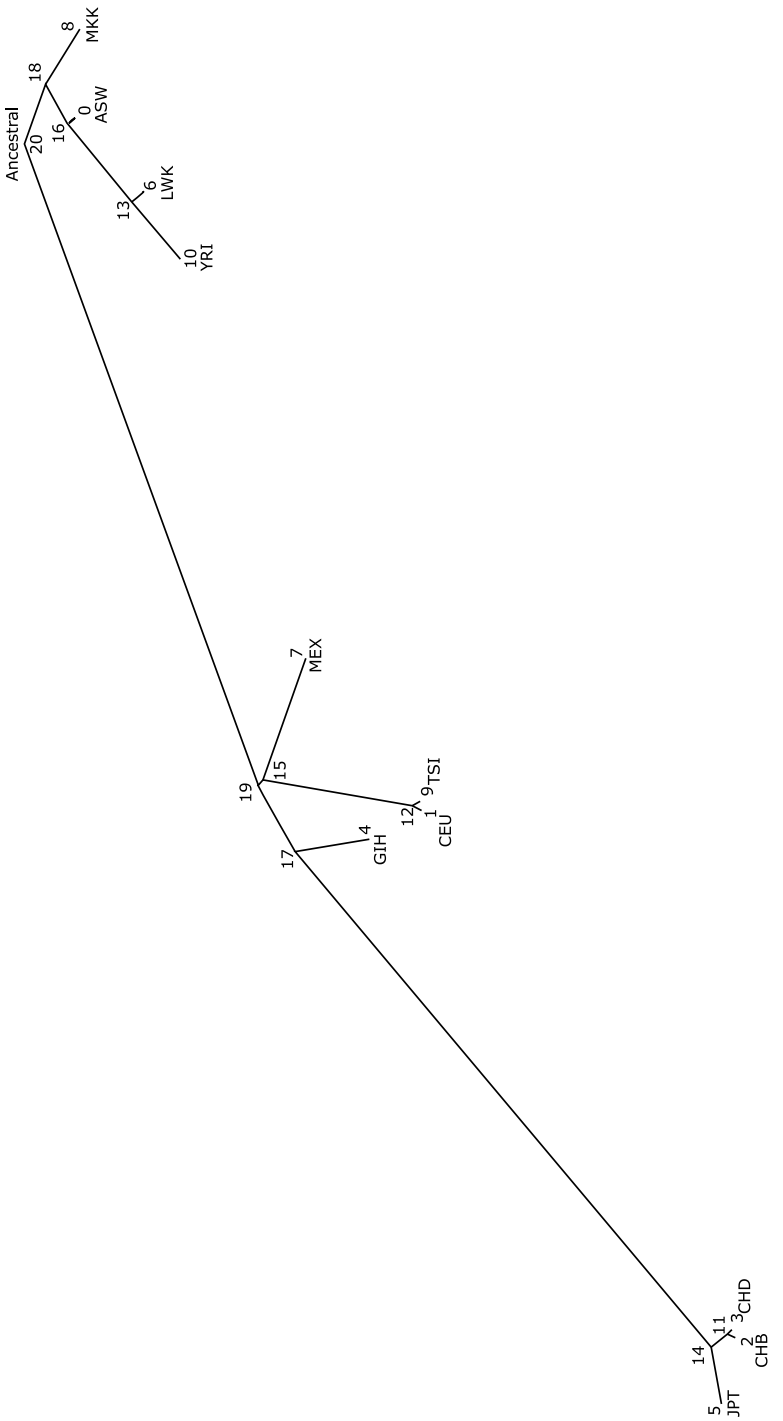

Figure 4.18: Phylogenetic Tree with Edge Lengths Proportional to Estimated Genetic Drift for chromosome 22 with a Beta (0.5,0.5) prior on π

The twenty periods of genetic drift for the HapMap dataset with the edge lengths proportional to the posterior mean estimates of drift for chromosome 22. A Beta(0.5,0.5) prior on π is assumed.

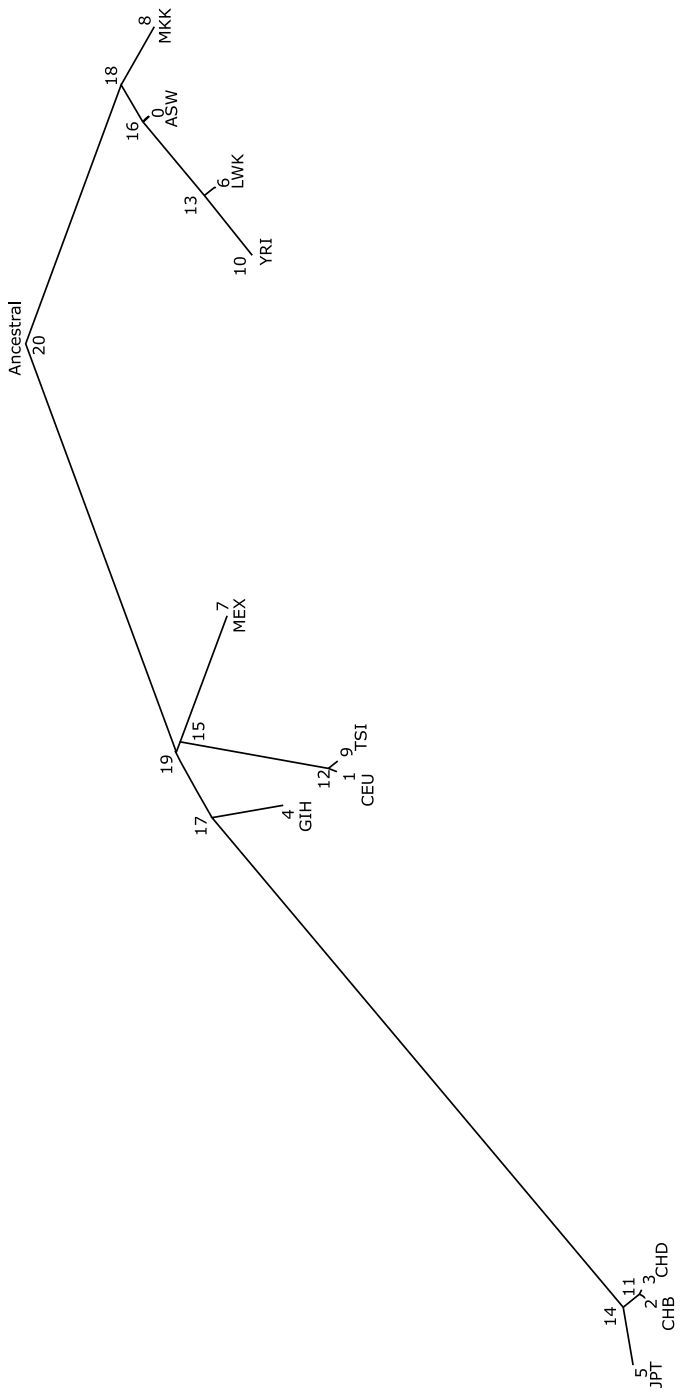

Figure 4.19: Phylogenetic Tree with Edge Lengths Proportional to Estimated Genetic Drift for chromosome 22 with a Beta (2,2) prior on π

The twenty periods of genetic drift for the HapMap dataset with the edge lengths proportional to the posterior mean estimates of drift for chromosome 22. A Beta(2,2) prior on π is assumed.

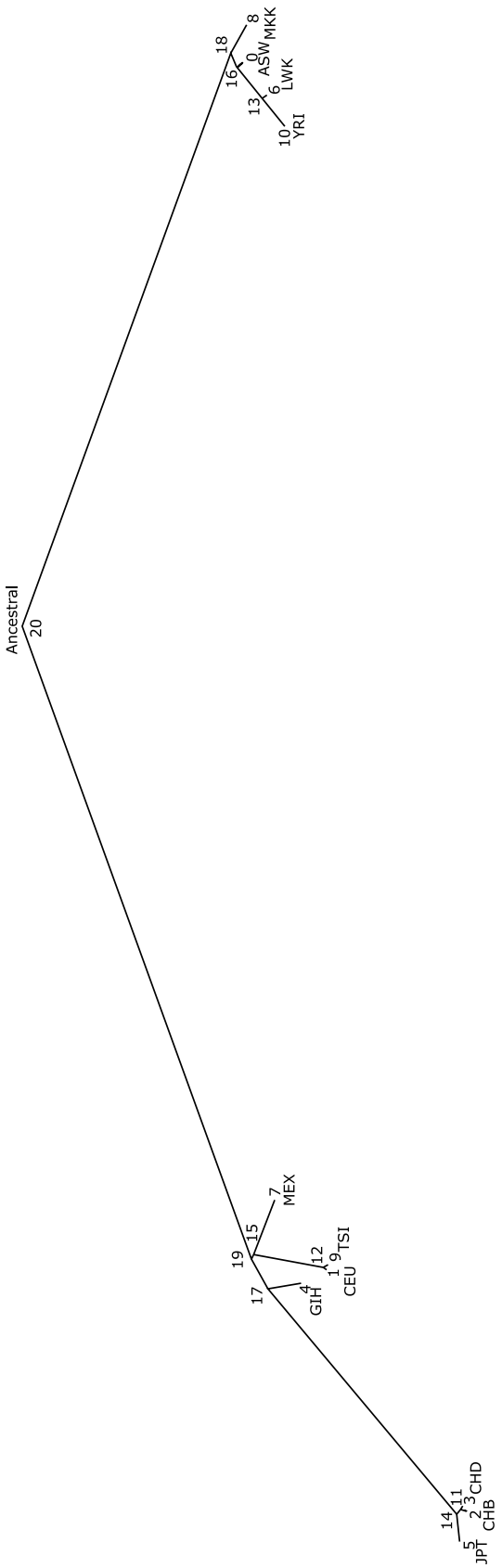

Figure 4.20: Phylogenetic Tree with Edge Lengths Proportional to Estimated Genetic Drift for chromosome 22 with a Beta (10,10) prior on π

The twenty periods of genetic drift for the HapMap dataset with the edge lengths proportional to the posterior mean estimates of drift for chromosome 22. A Beta(10,10) prior on π is assumed.

So, one effect of different choices of prior on the π_is is to move the position of the ancestral population in relation to the subpopulations below it. The remainder of the inferred tree is quite robust. The other interesting effect is the impact this has on the residuals that are calculated from the fitted model. Figure 4.21 shows the histogram of standardised residuals for the model fitted to the chromosome 15 data, along with the boxplot with the prior as Beta(1,1).

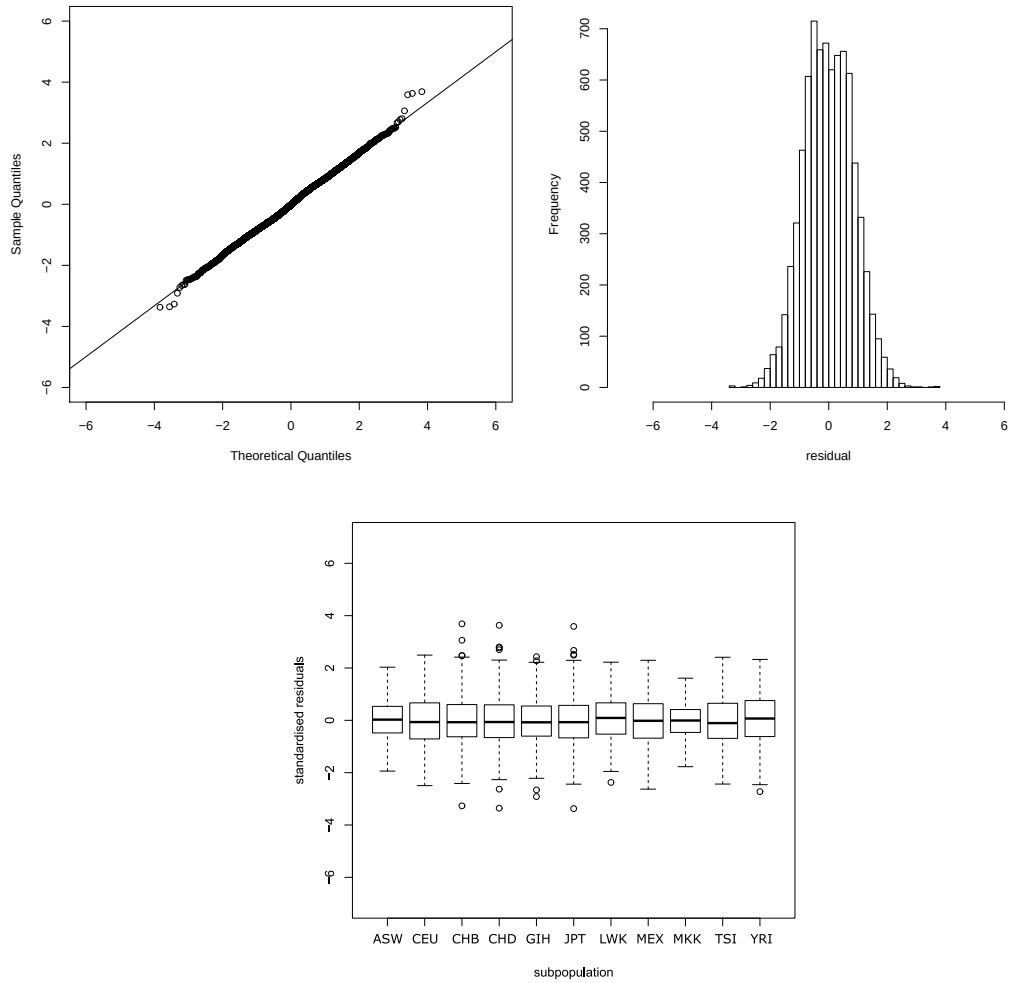

Figure 4.21: Plots of Standardised Residuals using the approximate mean and variance and QQ Plot for Chromosome 15 with the Beta(1,1) prior

Diagnostic plots of standardised residuals for the bifurcating ND model for Chromosome 15. The histogram in the centre shows hints at a possible bimodal pattern suggesting that there are factors in the data that the model does not take sufficiently into account. The QQ plot on the left gives no cause for concern on its own. The boxplot shows less spread of the standardised residuals for the subpopulations experiencing the least drift, the Afro-Americans (ASW) and the Maasai (MKK).

The QQ plot is unremarkable and gives no cause for concern, while the boxplot has a smaller spread of standardised residuals for the two subpopulations experiencing the least drift, the Maasai (MKK) and the Afro-Americans (ASW). The histogram shows a hint of the bimodality that was seen in Figure 4.13 for chromosome 2. Now, if the prior on π is changed to Beta(0.5,0.5), the Jeffreys prior, the plots of residuals are shown in Figure 4.22.

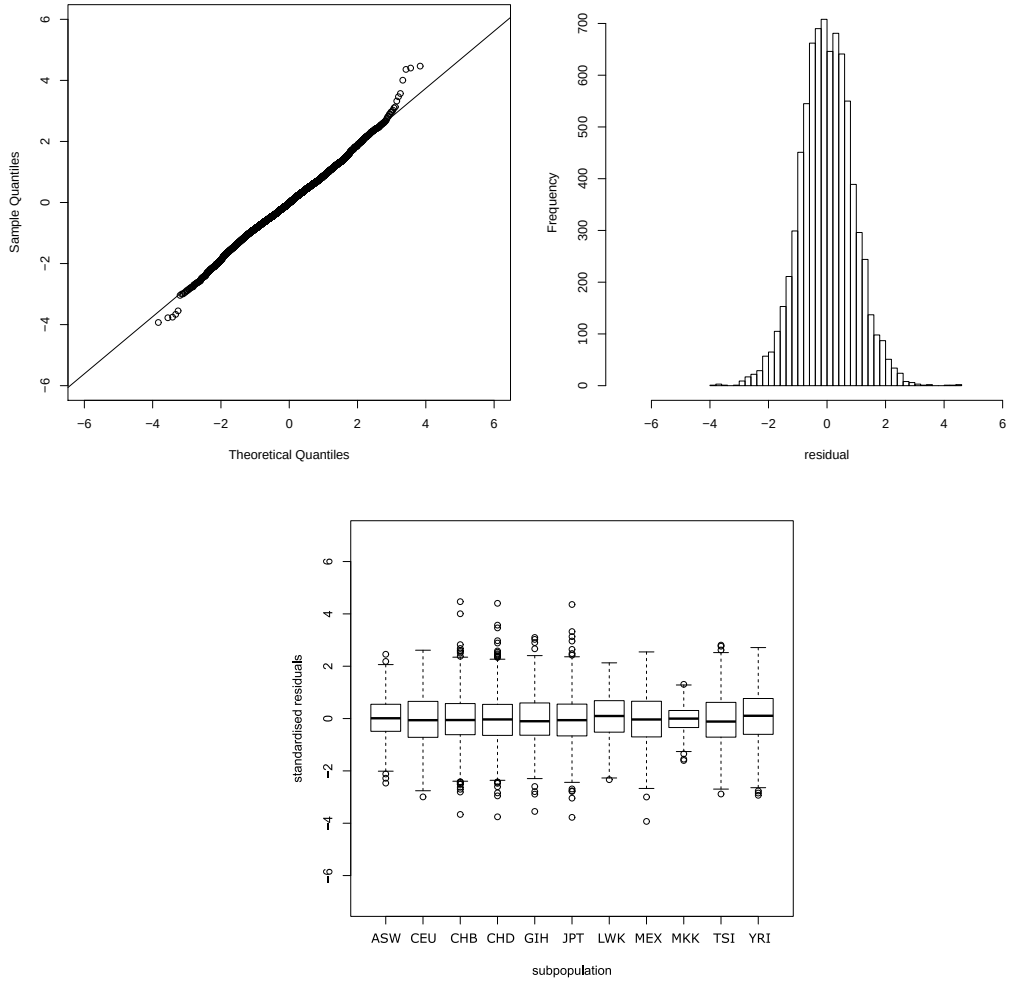

Figure 4.22: Plots of Standardised Residuals using the approximate mean and variance and QQ Plot for Chromosome 15 with the Beta(0.5,0.5) prior

Diagnostic plots of standardised residuals for the bifurcating ND model for Chromosome 15 with the Beta(0.5,0.5) prior on π. The histogram in the centre shows less evidence of a bimodal pattern. The QQ plot on the left gives no cause for concern. The boxplot shows even less spread of the standardised residuals for the subpopulations experiencing the least drift, the Maasai (MKK).

Here the QQ plot shows a departure from linearity in the tails compared to the Beta(1,1) prior. The boxplot for the Maasai (MKK), the subpopulation that has experienced the least drift, reflects an even smaller spread of residuals, but the histogram shows less evidence of bimodality than for Beta(1,1). A poor choice of prior seems to be contributing to the bimodality in the histogram. The narrower boxplot for MKK reflects the fact that this choice of prior moves the ancestral population closer to the African subpopulations as was seen earlier. This means that the MKK subpopultion is now expriencing even less overall drift from the ancestral population and, since the spread of residuals is lower for those subpopulations that have experienced less drift, this leads to its boxplot narrowing. Moving the parameter of the prior in the opposite direction, with a prior on π of Beta(2,2), gives the plots shown in Figure 4.23.

Unsurprisingly, the opposite effects are observed for this change of prior. Now the QQ plot is starting to show the first signs of skew. The residual histogram now has a more obvious bimodal pattern and the boxplot has a more even pattern of spread for the standardised residuals by subpopulation. Indeed if the parameter of the prior is increased the effects become greater. A model for the chromosome 22 data was fitted with the rather extreme Beta(10,10) prior to illustrate the point clearly. The residual plots in that case are shown in Figure 4.24.

Here the QQ plot shows a clear S shape, the bimodal pattern for the standardised residuals is extreme but the boxplots for the standardised residuals by subpopulation are quite even. This reflects the fact that a lot of genetic drift has been experienced by all the subpopulations to get from such an extreme bell-shaped distribution in the ancestral population so, since they have all experienced a lot of genetic drift, their boxplots will this time appear about even.

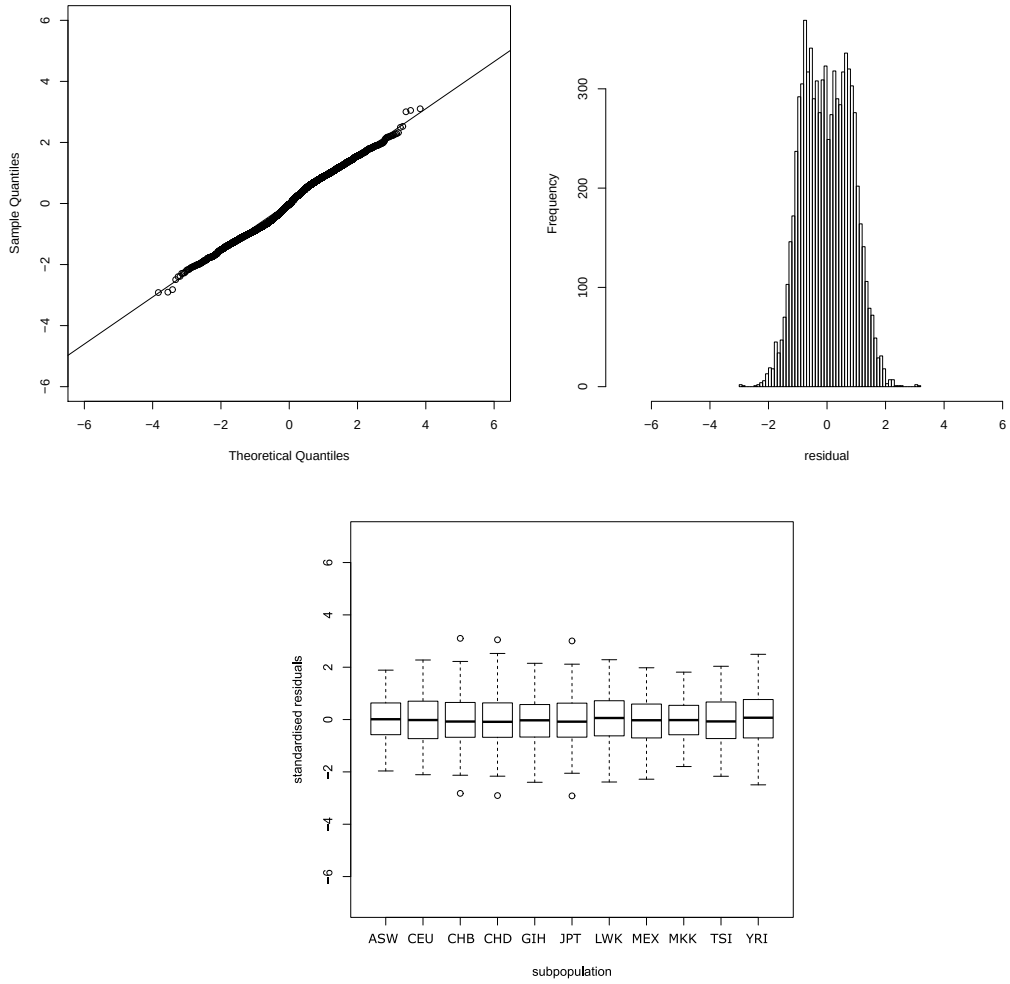

Figure 4.23: Plots of Standardised Residuals using the approximate mean and variance and QQ Plot for Chromosome 15 with the Beta(2,2) prior

Diagnostic plots of standardised residuals for the bifurcating ND model for Chromosome 15 with the Beta(2,2) prior on π. The histogram in the centre shows more marked evidence of a bimodal pattern. The QQ plot on the left gives no cause for concern. The boxplot shows a more even spread of standardised residuals for the subpopulations.

Clearly, the choice of prior on π does affect the position of the ancestral population at the root of the phylogenetic tree in relation to its subpopulations and this impacts on the standardised residuals since the estimated values of π for each locus and the estimated values of c for the genetic drift for each of the ancestral population's immediate subpopulations feed into that calculation and are affected by the choice of prior.

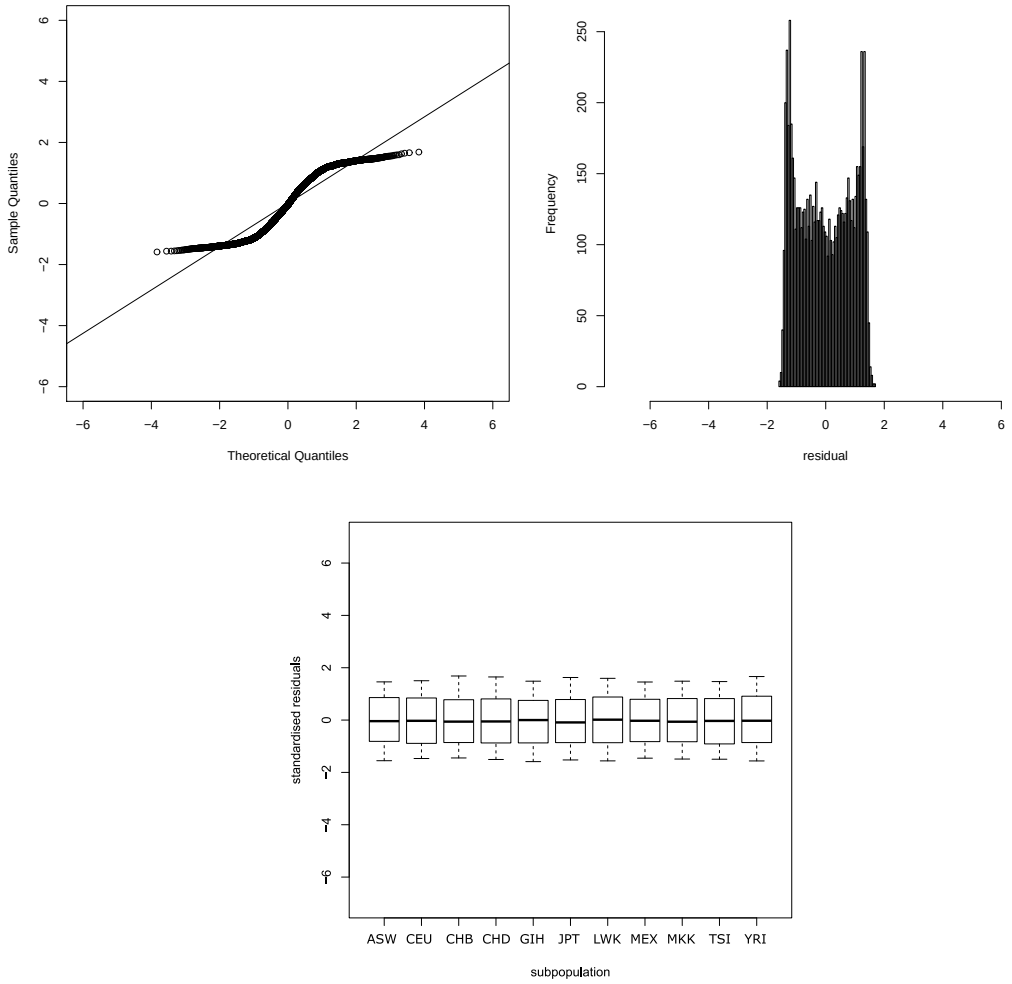

Figure 4.24: Plots of Standardised Residuals using the approximate mean and variance and QQ Plot for Chromosome 15 with the Beta(10,10) prior

Diagnostic plots of standardised residuals for the bifurcating Nicholson–Donnelly model for Chromosome 15 with the Beta(10,10) prior on π. The histogram in the centre shows an extreme bimodal pattern. The QQ plot on the left has a marked S shape. The boxplot shows a very even spread of the standardised residuals for the subpopulations.

However, the rest of the phylogenetic tree seems to be robust to this choice of prior. This will need to be borne in mind and some experimentation with different, but sensible, choices of prior that more accurately reflect the, unfortunately, poorly understood distribution of proportions of alleles that might reasonably be expected in the ancestral population. Section 5.6.2.1 investigates the use of an outgroup to mitigate the effects of a poor choice of prior on π. Choosing a different prior on c

in the same way, Beta$(0.5, 0.5)$ or Beta$(2, 2)$ in place of Beta$(1, 1)$, was not found to make any material difference.

4.6 Conclusions

Tests on the model showed that it works as intended. Nevertheless, it did not model the HapMap data as well as was hoped. Examination of the standardised residuals revealed a persistent bimodal pattern. This turned out to be remedied by using an appropriate choice of prior on the ancestral proportions of the allele counted at each locus. The problem with the residuals for each subpopulation not being equally spread was investigated and found not to be due to the use of approximate values of the mean and variance of the distribution describing drift. Posterior predictive checking revealed that the assumption that they should be equally spread was, in fact, erroneous and that the pattern was not actually a cause for concern. In the light of bimodality of the residuals being explained by choice of prior, it was necessary to consider whether this more complicated phylogenetic tree model was necessary. Posterior predictive checking did show that the phylogenetic part of the model was indeed necessary to better explain the data. Nevertheless, the results of posterior predictive checking also showed that this phylogenetic tree model alone does not describe the data adequately. It is highly plausible that at least two of the subpopulations, the Afro-Americans and the Mexicans, have resulted from admixture events in their past and so are related to other subpopulations in ways that this model does not attempt to take into account. To extend this model to take admixture into account would be an ambitious undertaking that is nevertheless worth trying. It is the development of such a model that is the subject of the next chapter.

Chapter 5

Generalisation to Allow Admixture Events

The new models described in the previous chapter assume that subpopulations never become socially involved with other subpopulations to the extent of producing offspring with parentage shared between the two subpopulations in sufficient numbers to warrant representation in the model. New subpopulations could only be created by two groups within one parent subpopulation becoming isolated from each other so that they experienced genetic drift independently. In this chapter, that assumption is relaxed. Two subpopulations can meet and create a new subpopulation with its genetic character partially inherited from each of the parent subpopulations. The new model introduced in the previous chapter is further developed here to accommodate these admixture events.

5.1 Admixture Events in Genetics

The emergence of a new subpopulation that inherits its genetic character from at least two parent subpopulations is called admixture (Balding et al., 2007). Two

subpopulations or parts of these two subpopulations, which were previously isolated from each other, meet and integrate to the extent of producing offspring who form a new subpopulation which has inherited genetic material from both the two parent subpopulations. The two subpopulations could meet at a particular place and combine at a particular time in history to form the new subpopulation. However, the process of integration may not be instantaneous. It may take place over a number of generations. If the two populations meet geographically, there may be migratory flow from one or both the parent subpopulations over a number of generations. While it is more socially and geographically realistic to suppose an admixture event took place over a period of time, in the interests of simplicity, it will be assumed that it can be modelled as if it were an instantaneous event.

5.1.1 Examples of Different Types of Historical Admixture Events

Historically, large-scale human admixture events have happened at a great many points in history and in many different circumstances. One way that two previously isolated subpopulations could meet in prehistoric times was if a physical geographical barrier between them ceased to exist as a result of climate change. This could happen if a land bridge between two land bodies appears as the result of lower sea water levels e.g., the Bering Strait (Elias et al., 1996). Often technological developments would allow movement by one or other subpopulation over a barrier. Developments in ability to navigate accurately at sea or improved ships are such examples (Rayment, 2017). Improvements in both of these technologies contributed to an age of exploration from the 16th century onwards that brought populations into contact that had been previously isolated. Barriers need not be physical. They can be social. The Amish in North America are descended from central European immigrants. They live without many forms of modern technology and have deep pacifist and religious beliefs. While they live among other

Americans, it is extremely rare for people outwith the Amish to marry into their communities (Hou et al., 2013). Many societies have historically had taboos about admixture. Indeed, it is only as recently as 1967 that the US supreme court, in the case of Loving v Virginia (US Supreme Court, 1967), ruled the laws banning miscegenation (interracial marriage) that still existed in 16 of the 50 states of the USA were unconstitutional. The Immorality Act in South Africa which banned South Africans of European ancestry from intercourse with people from other subpopulations was repealed only as recently as 1985 (Republic of South Africa, 1985). It should be noted that the participants in admixture events may not always have done so willingly. Slavery has featured in many societies throughout human history. In most cases slaves were often taken from a different subpopulation or culture to that of their owners. Slaves were sexually exploited leading to admixed populations (Baptist, 2001). Large colonial migrations can be driven by the possibility of making a better life and stories of riches on offer. Colonial migrations are also not always undertaken voluntarily. Convicts in the British Empire were often transported to penal colonies, first in North America (Ekirch, 1990) and later, in Australia (McConville, 1981). Still others felt compelled to migrate due to hunger. The potato famine which hit Ireland in the 19th century led many to migrate to the colonies of the then British Empire (Foster, 1988). Colonists, whatever the reason for their migration, would then be in proximity to native subpopulations leading to the possibility for admixture.

5.1.2 Features of an Admixture Event in a Simple Context

How could an admixture event be modelled in a way that will fit into the branching genetic drift models that have been described thus far? An admixture model in its simplest context is shown in figure 5.1. Each line represents a period of genetic drift. The ancestral population at A diverges into two subpopulations that experience independent genetic drift until they reach points B and C. At point B, a

subset of that subpopulation is destined to be involved in an admixture event. The remainder of that subpopulation continues to experience genetic drift and becomes the present-day subpopulation represented at G. Similarly the other subpopulation at C also has a subset that is destined to be involved in the same admixture event while the remainder of that subpopulation continues to drift and becomes the present-day subpopulation at J. The subset of the subpopulations at B and C that are destined to become admixed may each experience genetic drift on their journeys towards meeting each other at points D and E, respectively. These two subpopulations now meet and mix to become a new composite third subpopulation at F. This new admixed subpopulation can experience genetic drift itself, becoming the present-day admixed population at H. Each of the two subpopulations at D and E will contribute to a proportion of the new admixed population at F. So that if D contributes a proportion, w, of the admixed subpopulation then E must contribute a proportion, $1 - w$, of the admixed subpopulation. Thus at any particular SNP locus, if the proportion of an allele in the subpopulation at D is α_D and at E is α_E, the proportion of the allele in that new admixed subpopulation at F must be $\alpha_F = w\alpha_D + (1 - w)\alpha_E$. Moreover, the proportion w has the same value at every locus.

This shows how an admixture event could be modelled in a simple setting, but the salient elements of it can be easily incorporated into far more complex phylogenetic trees which could have several such admixture events. This model allows there to be genetic drift along any or all of the edges BD, CE and FH and so is quite general. In reality, there may be very little genetic drift along one or more of these edges, if the time periods represented by them are very short and the subpopulation sizes not too small. For example, the edge BD could represent the transportation of convicts to a penal colony on an island who are later released and mix with the native population there. This would take only a single generation. The only drift would result from the sampling effect of taking a number of convicts from a much larger colonial parent population at point B. The proportions of alleles at each

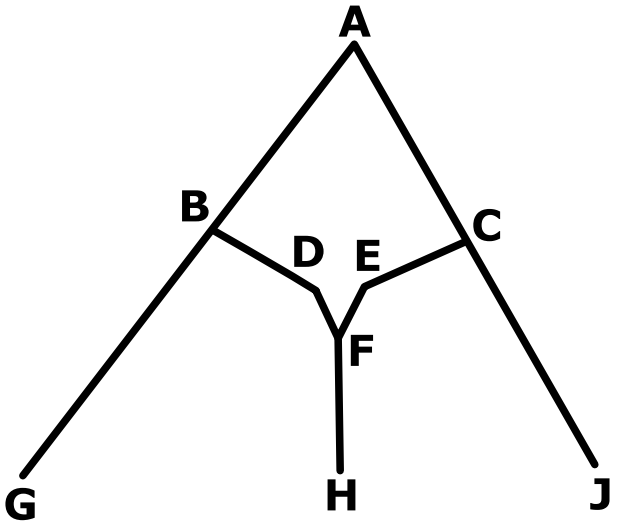

Figure 5.1: A Simple Admixture Model

Each line represents a period of genetic drift. The ancestral population at A diverges into two subpopulations that experience independent genetic drift until they reach points B and C. At point B, a subset of that subpopulation is destined to be involved in an admixture event. The remainder of that subpopulation continues to experience genetic drift and becomes the present day subpopulation represented at G. Similarly the other subpopulation at C also has a subset that is destined to be involved in the same admixture event while the remainder of that subpopulation continues to drift and becomes the present-day subpopulation at J. The subset of the subpopulations at B and C that are destined to become admixed may each experience genetic drift on their journeys towards meeting each other at points D and E, respectively. These two subpopulations now meet and mix to become a new composite third subpopulation at F. This new admixed subpopulation can experience genetic drift itself, becoming the present-day admixed population at H.

locus in the convict population at D from the colonial parent population at B would not be the same and therefore would appear to have drifted slightly but would only differ because of sampling variation rather than genetic drift, strictly speaking, because no reproduction has taken place along the edge BD. However, a lot of people in close proximity to each other on board a ship would be vulnerable to the spread of disease, so it could also be argued that those that survive are more different from the populations they are taken from than this drift would explain, due to selection effects.

5.2 Description of the Model

The new parameter is w, the admixture parameter. This takes values in $(0, 1)$ and so a beta prior, which covers those values, seems a reasonable choice. Giving it two hyperparameters, ω_1, ω_2 allows the possibility for a strong prior to be set by the experimenter where outside knowledge or previous studies make a particular value more likely. However setting these to 1 sets a weak prior where all values are as likely as each other a-priori. Alternative weak priors such as the Jeffreys prior beta(0.5,0.5) could be used but since there is no obvious alternative scale on which to measure the admixture parameter, this seems unnecessary. The justification for the priors on c_j and π_i remain as they were described in the earlier model of section 3.2.2.

$$x_{ij}|n_{ij}, \alpha_{ij} \sim \text{Binomial}\left(n_{ij}, \alpha_{ij}\right), \text{ independently,}$$

$$\alpha_{ij}|\pi_i, c_j \sim \text{N}^{\text{R}[0,1]}\left(\pi_i, \pi_i\left(1-\pi_i\right)c_j\right), \text{, independently, for } \alpha\text{s nearest the root of the phylogenetic network,}$$

$$\alpha_{ij} = w_j\alpha_{ip_1} + \left(1-w_j\right)\alpha_{ip_2}, \text{ deterministically, for the } \alpha\text{s immediately following an admixture event.}$$

where α_{ip_1} and α_{ip_2} are the alphas for the two parent nodes in the network that feed into the admixture event.

$$\alpha_{ij}|\alpha_{ipi}, c_j \sim \mathrm{N}^{\mathrm{R}[0,1]}\left(\alpha_{ip}, \alpha_{ipi}\left(1-\alpha_{ip}\right)c_j\right), \text{ independently, for other } \alpha\text{s},$$

where α_{ip} is the alpha for the parent node to node j in the tree.

with priors

$$w_j|\omega_1, \omega_2 \sim \mathrm{Beta}\left(\omega_1, \omega_2\right), \text{ independently,}$$

$$\pi_i|a \sim \mathrm{Beta}\left(a, a\right), \text{ independently,}$$

$$c_j|b_{1j}, b_{2j} \sim \mathrm{Beta}(b_{1j}, b_{2j}), \text{ independently,}$$

where

i labels the locus: $1 \leqslant i \leqslant L$,

j labels the subpopulation $1 \leqslant j \leqslant P$,

n_{ij} is the total number of alleles observed at locus i in subpopulation j,

x_{ij} is the number of one of the two alleles observed at locus i in subpopulation j,

α_{ij} is the population proportion of that allele at locus i in subpopulation j,

π_i is the proportion of that allele at locus i in the ancestral population,

c_j is the amount of genetic drift in subpopulation j.

w_j is the admixture proportion for the admixture event that results in subpopulation j where that subpopulation is the direct result of an admixture event.

a is a hyperparameter in the prior of π_i.

b_{1j}, b_{2j} are hyperparameters in the prior of c_j and will be assigned the value 1 unless otherwise stated.

ω_1 and ω_2 are hyperparameter in the prior of w_j and will be assigned the value 1 unless otherwise stated.

5.3 Implementation of the Model

5.3.1 Hierarchical Model of an Admixture Event

The DAG of the simple admixture event described in figure 5.1 is shown in figure 5.2.

The new parameters of this model are the admixture proportion, w, its prior ω, and the allele frequencies at each SNP (generically i) in the newly admixed subpopulation, $\alpha_{i,F}$, with a deterministic relationship to w, $\alpha_{i,D}$ and $\alpha_{i,E}$. These are shown in this simple context for illustrative purposes but are generalisable in the obvious way to more complicated phylogenetic trees which could have several such admixture events. The said deterministic relationship is $\alpha_{i,F} = w\alpha_{i,D} + (1-w)\alpha_{i,E}$. The remaining features of this model are directly analogous to those described in chapters 3 and 4.

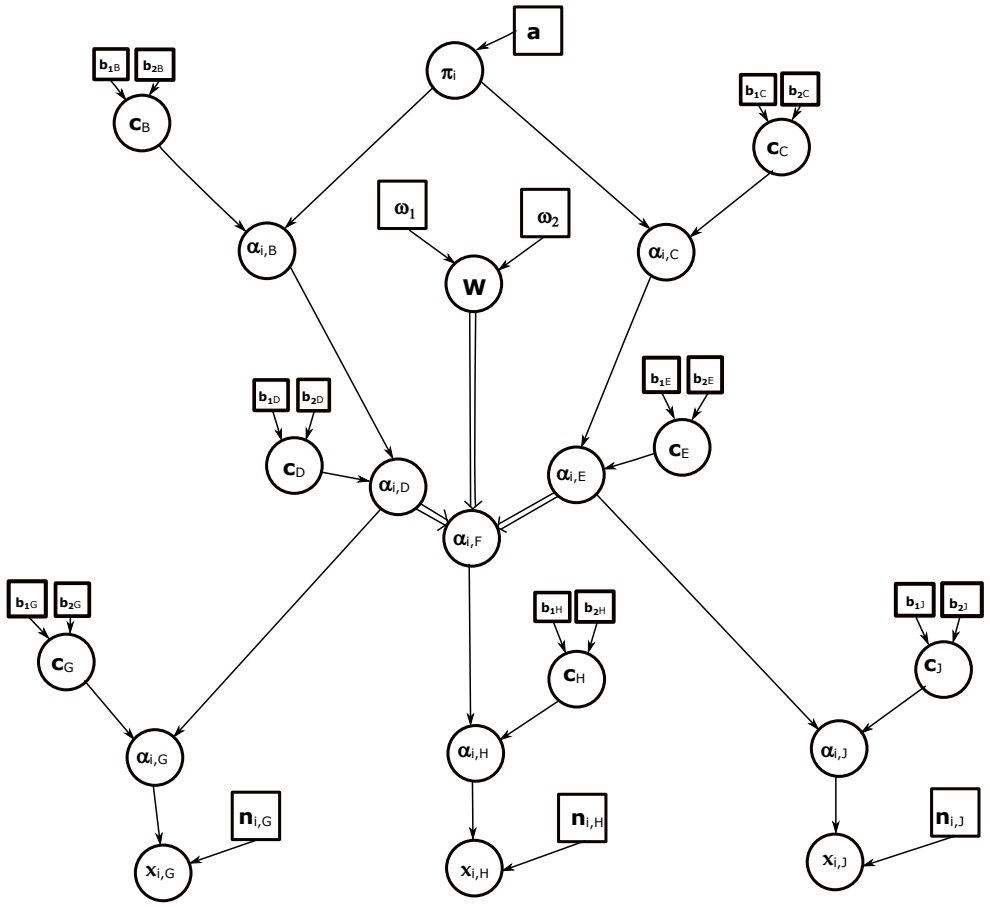

Figure 5.2: DAG of an Admixture Model in a simple context

DAG of the simple model including an admixture event where
i labels the locus: $1 \leqslant i \leqslant L$,
Letters A,B,...H,J correspond to the nodes in the phylogeneitic tree shown in figure 5.1.
n_{ij} is the total number of alleles observed at locus i in subpopulation j,
x_{ij} is the number of one of the two alleles observed at locus i in subpopulation j,
α_{ij} is the population proportion of that allele at locus i in subpopulation j,
π_i is the proportion of that allele at locus i in the ancestral population, with a as a parameter within its prior,
c_j parameterises the amount of genetic drift in subpopulation j,
w is the proportion of the admixed subpopulation at F that is contributed by D and consequently $(1-w)$ is the proportion of the admixed population contributed by E. ω is a parameter in its prior.

Once again, the full conditional for the ancestral allele frequency,π_i in this simplified case is

$$P\left(\pi_i|\alpha, c, \pi_{-i}\right) \propto \pi_i^{a-1}\left(1-\pi_i\right)^{a-1} g_1\left(c_B, \pi_i, \alpha_{iB}\right) g_1\left(c_C, \pi_i, \alpha_{iC}\right), \tag{5.1}$$

and in a more general setting where s subpopulations rather than just the two at B and C are directly descended from it becomes

$$P\left(\pi_i|\alpha, c, \pi_{-i}\right) \propto \pi_i^{a-1}\left(1-\pi_i\right)^{a-1} \prod_{m=1}^{s} g_1\left(c_{k_m}, \pi_i, \alpha_{ik_m}\right), \tag{5.2}$$

where $\{k_1, ..., k_s\}$ is the set of child nodes of the ancestral node as before, and

$$g_1\left(c_k, \pi_i, \alpha_{ik}\right) = \begin{cases} \left[\pi_i\left(1-\pi_i\right)\right]^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\pi_i)^2}{2c_k\pi_i(1-\pi_i)}\right) dr, & \alpha_{ik}=0, \\ \left[\pi_i\left(1-\pi_i\right)\right]^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ik}-\pi_i)^2}{2c_k\pi_i(1-\pi_i)}\right), & 0<\alpha_{ik}<1, \\ \left[\pi_i\left(1-\pi_i\right)\right]^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\pi_i)^2}{2c_k\pi_i(1-\pi_i)}\right) dr, & \alpha_{ik}=1. \end{cases} \tag{5.3}$$

The full conditional for c_j is also unchanged,

$$P\left(c_j|\alpha, \pi, c_{-j}, b\right) \propto \prod_{i=1}^{L} g_2\left(c_j, \alpha_{ip}, \alpha_{ij}\right) \times c_j^{b_{1j}-1}\left(1-c_j\right)^{b_{2j}-1}., \tag{5.4}$$

where

$$g_2\left(c_j, \alpha_{ip}, \alpha_{ij}\right) = \begin{cases} c_j^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right) dr, & \alpha_{ij}=0, \\ c_j^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ij}-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right), & 0<\alpha_{ij}<1, \\ c_j^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right) dr, & \alpha_{ij}=1. \end{cases} \tag{5.5}$$

Here, L is the number of loci, and α_{ip} is the allele frequency for the ith locus at the parent node of j. As before and also below, in the case where j is one of the child nodes of the ancestral root then $\alpha_{ip} \equiv \pi_i$.

The full conditionals for α_{ij} at the tips of the phylogenetic tree such as α_{iG}, α_{iH} and α_{iJ} in the above simplified example, that represent present-day subpopulations also remain unaffected. These are the αs nearest the data, x_{ij}, n_{ij}, in the hierarchy:

$$P\left(\alpha_{ij}|c_j, \pi_i, \alpha_{-ij}, x_{ij}, n_{ij}\right) \propto h_1\left(n_{ij}, x_{ij}, \alpha_{ij}\right) g_3\left(c_j, \alpha_{ip}, \alpha_{ij}\right) \tag{5.6}$$

and

$$g_3\left(c_j, \alpha_{ip}, \alpha_{ij}\right) = \begin{cases} \int_{-\infty}^{0} \exp\left(\frac{-(r-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right) dr, & \alpha_{ij} = 0, 0 < \alpha_{ip} < 1, \\ \exp\left(\frac{-(\alpha_{ij}-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right), & 0 < \alpha_{ij} < 1, 0 < \alpha_{ip} < 1, \\ \int_{1}^{\infty} \exp\left(\frac{-(r-\alpha_{ip})^2}{2c_j\alpha_{ip}(1-\alpha_{ip})}\right) dr, & \alpha_{ij} = 1, 0 < \alpha_{ip} < 1, \\ 1, & \alpha_{ij} = 1, \alpha_{ip} = 1, \\ 1, & \alpha_{ij} = 0, \alpha_{ip} = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5.7}$$

$$h_1\left(n_{ij}, x_{ij}, \alpha_{ij}\right) = \begin{cases} 1, & \alpha_{ij} = 0, x_{ij} = 0, \\ \alpha_{ij}^{x_{ij}}\left(1-\alpha_{ij}\right)^{n_{ij}-x_{ij}}, & 0 < \alpha_{ij} < 1, \\ 1, & \alpha_{ij} = 1, x_{ij} = n_{ij}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.8}$$

Full conditionals for the αs not at the tips of the phylogenetic tree and that are not directly involved with the admixture such as α_{iB} and α_{iC} in the simple example above are also unaffected. These are

$$P\left(\alpha_{ij}|c, \pi, \alpha_{-ij}\right) \propto h_2\left(\alpha_{ij}, \alpha_{-ij}, c\right) g_3\left(c_j, \alpha_{ip}, \alpha_{ij}\right), \tag{5.9}$$

and

$$h_2\left(\alpha_{ij}, \alpha_{-ij}, c\right) = \begin{cases} 1 & \alpha_{ij} = 0, \alpha_{ik_1} = \alpha_{ik_2} = ... = \alpha_{ik_{s_j}} = 0, \\ 1 & \alpha_{ij} = 1, \alpha_{ik_1} = \alpha_{ik_2} = ... = \alpha_{ik_{s_j}} = 1, \\ \prod_{m=1}^{s_j} f\left(c_{k_m}, \alpha_{ij}, \alpha_{ik_m}\right) & 0 < \alpha_{ij} < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.10}$$

where again, $\left\{k_1, ..., k_{s_j}\right\}$ is the set (of size s_j) of child nodes of the node (j) in

question and

$$f\left(c_k, \alpha_{ij}, \alpha_{ik}\right) = \begin{cases} \left[\alpha_{ij}\left(1-\alpha_{ij}\right)\right]^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\alpha_{ij})^2}{2c_k\alpha_{ij}(1-\alpha_{ij})}\right) dr, & \alpha_{ik} = 0, \\ \left[\alpha_{ij}\left(1-\alpha_{ij}\right)\right]^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ik}-\alpha_{ij})^2}{2c_k\alpha_{ij}(1-\alpha_{ij})}\right), & 0 < \alpha_{ik} < 1, \\ \left[\alpha_{ij}\left(1-\alpha_{ij}\right)\right]^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\alpha_{ij})^2}{2c_k\alpha_{ij}(1-\alpha_{ij})}\right) dr, & \alpha_{ik} = 1. \end{cases} \tag{5.11}$$

That leaves the αs directly involved in the admixture event. The case of α_{iF} is relatively simple. As noted before, it is determined from the values of the admixture parameter w and the proportions of the allele from the two subpopulations that make up the admixture $\alpha_{iF} = w\alpha_{iD} + (1-w)\alpha_{iE}$. More generally, if that α is labelled α_{ij} to be consistent with the notation above and its two contributing parent αs as α_{ip_1} and α_{ip_2} this deterministic relationship becomes

$$\alpha_{ij} = w\alpha_{ip_1} + (1-w)\alpha_{ip_2}. \tag{5.12}$$

The remaining cases are those αs that contribute to an admixture such as α_{iD} and α_{iE} in the simple case. In these cases the full conditionals are

$$P\left(\alpha_{ij}|c, w, \alpha_{-ij}\right) \propto g_4\left(\alpha_{ij}, \alpha_{-ij}, w, c\right) g_3\left(c_j, \alpha_{ip}, \alpha_{ij}\right), \tag{5.13}$$

where $g_3\left(c_j, \alpha_{ip}, \alpha_{ij}\right)$ is as above and,

$$g_4\left(\alpha_{ij}, \alpha_{-ij}, w, c\right) = \begin{cases} \left[\alpha_{im}\left(1-\alpha_{im}\right)\right]^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\alpha_{im})^2}{2c_k\alpha_{im}(1-\alpha_{im})}\right) dr, & \alpha_{ik} = 0, \alpha_{ij} \neq \alpha_{iv}, \\ \left[\alpha_{im}\left(1-\alpha_{im}\right)\right]^{-\frac{1}{2}} \int_{-\infty}^{0} \exp\left(\frac{-(r-\alpha_{im})^2}{2c_k\alpha_{im}(1-\alpha_{im})}\right) dr, & \alpha_{ik} = 0, \alpha_{ij} = \alpha_{iv} < 1, \\ \left[\alpha_{im}\left(1-\alpha_{im}\right)\right]^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ik}-\alpha_{im})^2}{2c_k\alpha_{im}(1-\alpha_{im})}\right), & 0 < \alpha_{ik} < 1, \alpha_{ij} \neq \alpha_{iv}, \\ \left[\alpha_{im}\left(1-\alpha_{im}\right)\right]^{-\frac{1}{2}} \exp\left(\frac{-(\alpha_{ik}-\alpha_{im})^2}{2c_k\alpha_{im}(1-\alpha_{im})}\right), & 0 < \alpha_{ik} < 1, \alpha_{ij} = \alpha_{iv}, 0 < \alpha_{ij} < 1, \\ \left[\alpha_{im}\left(1-\alpha_{im}\right)\right]^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\alpha_{im})^2}{2c_k\alpha_{im}(1-\alpha_{im})}\right) dr, & \alpha_{ik} = 1, \alpha_{ij} \neq \alpha_{iv}, \\ \left[\alpha_{im}\left(1-\alpha_{im}\right)\right]^{-\frac{1}{2}} \int_{1}^{\infty} \exp\left(\frac{-(r-\alpha_{im})^2}{2c_k\alpha_{im}(1-\alpha_{im})}\right) dr, & \alpha_{ik} = 1, \alpha_{ij} = \alpha_{iv} > 0 \\ 0 & \textit{otherwise} \end{cases} \tag{5.14}$$

where $\alpha_{im} = w\alpha_{ij} + (1-w)\alpha_{iv}$ in the case where α_{ij} is the proportion of the allele in the first population contributing to the admixture, analogous to α_{iD} in figure 5.2. In this case α_{iv} is the proportion of the allele in the second population contributing

to the admixture. Similarly, in the case where α_{ij} is the proportion of the allele in the second population contributing to the admixture, analogous to α_{iE} in figure 5.2, $\alpha_{im} = w\alpha_{iv} + (1-w)\alpha_{ij}$. This time, α_{iv} is the proportion of the allele in the first population contributing to the admixture. In these cases, the parameters w and the α for the other subpopulation contributing to the admixture enter the full conditional through the relationship 5.12. It should be noted that if $\alpha_{iv} = 0$ and $\alpha_{ik} \neq 0$ then $\alpha_{ij} \neq 0$ (or in figure 5.2, if $\alpha_{iE} = 0$ and $\alpha_{iH} \neq 0$ then $\alpha_{iD} \neq 0$) because the combination $\alpha_{iv} = \alpha_{ij} = 0$ (or $\alpha_{iD} = \alpha_{iE} = 0$ in figure 5.2), i.e., fixation in the two subpopulations that contribute to the new admixed population would imply that in the admixed population $\alpha_{im} = 0$ (or in figure 5.2 $\alpha_{iF} = 0$) but the admixed population has reaching fixation cannot be true if $\alpha_{ik} \neq 0$ (or in figure 5.2 $\alpha_{iH} \neq 0$) in the absence of mutation. By analogous reasoning, if $\alpha_{iv} = 1$ and $\alpha_{ik} \neq 1$ then $\alpha_{ij} \neq 1$. This is enforced by the conditions in $g_4(\alpha_{ij}, \alpha_{-ij}, w, c)$.

This leaves the full conditional for w. Since w can vary between 0 and 1, its prior should reflect that, so a beta prior is a reasonable choice. It may be useful to allow the possibility for flexibility in setting strong priors on w that using both parameters allows, so a $\text{Be}(\omega_1, \omega_2)$ prior seems reasonable. To make it a weak prior, there is no a-priori reason to assume it should be asymmetrical, $\omega_1 = \omega_2 = 1$ is one reasonable choice but is not the only reasonable choice. This leads to a full conditional for w of

$$P(w|\omega, \alpha_{ip_1}\alpha_{ip_2}, \alpha_{ik}, c_k) \propto w^{\omega_1 - 1}(1-w)^{\omega_2 - 1} \prod_{i=1}^{L} g_4(c_k, w\alpha_{ip_1} + (1-w)\alpha_{ip_2}, \alpha_{ik}), \tag{5.15}$$

where $\alpha_{ip_1}\alpha_{ip_2}$ are the αs from the two subpopulations contributing to the admixture (α_{iD} and α_{iE} in figure 5.2). Here k indexes the population descending from the admixed population (H in figure 5.2).

These describe the full conditionals up to proportionality. A similar process is followed for determining whether the alphas are in the atoms (equal to 0 or 1) as that described in section 4.5.2 of the previous chapter. There is, nonetheless,

an additional situation that has not been covered; the αs that contribute to an admixture such as α_D and α_E (dropping the i subscript for clarity) can also enter the atoms at 0 and 1 and an actual probability is, again, needed. Looking at this from the point of view of determining whether α_D is in the atom, when the parent alpha, α_B, is 0 or 1 then α_D must be in the same state because no mutation is assumed. That is a straightforward situation. But this time, if $0 < \alpha_B < 1$ then α_D can be 0 or 1 even if α_H is not. Again, a two-stage process is followed. First it is determined whether α_D is 0 or 1. Second, if it is not, then the usual sampling procedure is again followed for choosing a value in the $(0, 1)$ interval. The first stage needs probabilities for $\alpha_D = 0$ and for $\alpha_D = 1$. Taking the case of $\alpha_D = 0$, it must have got there by drift from α_P and $\alpha_F = 0 + (1 - w)\alpha_E$ must have drifted to α_H. The latter condition is clearly impossible if $(1 - w)\,\alpha_E = 0$ and $\alpha_H > 0$. Otherwise, these two steps are represented by

$$y_1\left(\alpha_B, c_D\right) = \Phi\left(\frac{0 - \alpha_B}{\sqrt{\alpha_B\left(1 - \alpha_B\right) c_D}}\right), \tag{5.16}$$

and

$$y_2\left(\alpha_F = 0 + \left[1 - w\right]\alpha_E, \alpha_H, c_H\right) = \frac{1}{\sqrt{\alpha_F\left(1 - \alpha_F\right) c_H}}\phi\left(\frac{\alpha_H - \alpha_F}{\sqrt{\alpha_F\left(1 - \alpha_F\right) c_H}}\right), \tag{5.17}$$

respectively when $0 < \alpha_H < 1$.

When $\alpha_H = 0$, the second step is represented by

$$y_2\left(\alpha_F = 0 + \left[1 - w\right]\alpha_E, c_H\right) = \Phi\left(\frac{0 - \alpha_F}{\sqrt{\alpha_F\left(1 - \alpha_F\right) c_H}}\right), \tag{5.18}$$

and when $\alpha_H = 1$ by

$$y_2\left(\alpha_F = 0 + \left[1 - w\right]\alpha_E, c_H\right) = 1 - \Phi\left(\frac{1 - \alpha_F}{\sqrt{\alpha_F\left(1 - \alpha_F\right) c_H}}\right), \tag{5.19}$$

or if $(1-w)\,\alpha_E = 0$ and $\alpha_H > 0$ (the case where $\alpha_D = 0$ is impossible) then $y_2\left(\alpha_F = 0 + [1-w]\,\alpha_E, c_H\right) = 0$.

By similar reasoning, the equivalent functions can be found to represent the case of $\alpha_D = 1$,

$$z_1\left(\alpha_B, c_D\right) = 1 - \Phi\left(\frac{1-\alpha_B}{\sqrt{\alpha_B\left(1-\alpha_B\right)c_D}}\right), \tag{5.20}$$

and when $0 < \alpha_H < 1$,

$$z_2\left(\alpha_F = w + [1-w]\,\alpha_E, \alpha_H, c_H\right) = \frac{1}{\sqrt{\alpha_F\left(1-\alpha_F\right)c_H}}\phi\left(\frac{\alpha_H - \alpha_F}{\sqrt{\alpha_F\left(1-\alpha_F\right)c_H}}\right), \tag{5.21}$$

or when $\alpha_H = 0$,

$$z_2\left(\alpha_F = w + [1-w]\,\alpha_E, c_H\right) = \Phi\left(\frac{0-\alpha_F}{\sqrt{\alpha_F\left(1-\alpha_F\right)c_H}}\right), \tag{5.22}$$

or when $\alpha_H = 1$

$$z_2\left(\alpha_F = w + [1-w]\,\alpha_E, c_H\right) = 1 - \Phi\left(\frac{1-\alpha_F}{\sqrt{\alpha_F\left(1-\alpha_F\right)c_H}}\right). \tag{5.23}$$

Again, this is impossible when $w + (1-w)\,\alpha_E = 1$ and $\alpha_H < 1$,

so $z_2\left(\alpha_F = w + [1-w]\,\alpha_E, c_H\right) = 0$ in that case.

Next, there is the possibility that α_D is in the interval $(0,1)$. The probability density of the drift from α_B to α_D is represented by

$$v\left(\alpha_B, c_D, r\right) = \frac{1}{\sqrt{\alpha_B\left(1-\alpha_B\right)c_D}}\phi\left(\frac{r-\alpha_B}{\sqrt{\alpha_B\left(1-\alpha_B\right)c_D}}\right). \tag{5.24}$$

The drift from $\alpha_F = w\alpha_D + (1-w)\alpha_E$ to α_H is represented by the probability

density

$$u\left(\alpha_H, c_H, r\right) = \frac{1}{\sqrt{r\left(1-r\right)c_H}}\phi\left(\frac{\alpha_H - r}{\sqrt{r\left(1-r\right)c_H}}\right), \tag{5.25}$$

when $0 < \alpha_H < 1$. If $\alpha_H = 0$ then

$$u\left(\alpha_H, c_H, r\right) = \Phi\left(\frac{0 - r}{\sqrt{r\left(1-r\right)c_H}}\right), \tag{5.26}$$

or when $\alpha_H = 1$

$$u\left(\alpha_H, c_H, r\right) = 1 - \Phi\left(\frac{1 - r}{\sqrt{r\left(1-r\right)c_H}}\right). \tag{5.27}$$

The maximum value that α_F can take is $\alpha_{F_{\mathcal{U}}} = w + (1-w)\alpha_E$, and the minimum it can take is $\alpha_{F_{\mathcal{L}}} = (1-w)\alpha_E$.

The probability of α_D being 0 is then the expression for the drift c_D carrying α_D to 0, divided by sum of the expressions for all the possibilities,

$$Pr\left(\alpha_D = 0\right) = \frac{y_1 y_2}{y_1 y_2 + z_1 z_2 + \int_{\alpha_{F_{\mathcal{L}}}}^{\alpha_{F_{\mathcal{U}}}} v\left(\alpha_B, c_D, r\right) u\left(\alpha_H, c_H, r\right) dr}. \tag{5.28}$$

A draw from Uniform(0,1) can be taken and if it is lower than this value, then $\alpha_D = 0$. Otherwise, this possibility is eliminated and the case of $\alpha_D = 1$ is considered in light of this, which has probability

$$Pr\left(\alpha_D = 1\right) = \frac{z_1 z_2}{z_1 z_2 + \int_{\alpha_{F_{\mathcal{L}}}}^{\alpha_{F_{\mathcal{U}}}} v\left(\alpha_B, c_D, r\right) u\left(\alpha_H, c_H, r\right) dr}. \tag{5.29}$$

Another draw from Uniform(0,1) is taken and if it is lower than this new value, then $\alpha_D = 1$. Otherwise, with these two possibilities eliminated, α_D takes a value in $(0, 1)$ drawn by the Gibbs sampler as usual. The same process can be found and followed for α_E by symmetry.

5.3.2 Determination of Candidate Subpopulations for Modelling as an Admixture

A reasonable question to ask is how can it be determined that a subpopulation should be modelled as admixed and of which subpopulations should it be an admixture? If it is known, as in the case of the HapMap data, what the subpopulations represent then knowledge of world history can be used to determine which subpopulations are likely to need to be modelled as admixture events. For example, there is a Mexican subpopulation in the HapMap dataset. There were native Americans in Mexico, the best known being the Aztec and Mayan civilisations. Europeans from Spain colonised the area in the early 16th century and as they were mostly men, took native wives and concubines and produced children. Martínez-Cortés et al. (2012) found the Y chromosome (male lineage) ancestry of modern Mexicans to be over 60% European, while Kumar et al. (2011) found that their maternal ancestry through mitochondrial DNA was 85%-90% native American. A knowledge of world history would lead to a view that present-day Mexicans are descended from an admixture of native Americans and Spanish Europeans. Native Americans are, in turn descended from people who crossed the Bering Straight from east Asia when sea levels were lower and it formed a land bridge (Elias et al., 1996). So it would seem reasonable to model Mexicans as an admixture of a drifted version of an old east Asian subpopulation such as the ancestor of modern day Japanese and Chinese subpopulations and a more recent ancestor of a west European subpopulation such as maybe CEU, the European subpopulation, or the ancestor of both CEU and the TSI Tuscan subpopulation. Similarly, the Afro-Americans (ASW) subpopulation could reasonably be expected to be an admixed subpopulation. Afro-Americans are descended from slaves taken from mostly West Africa and transported to America to work on plantations and for domestic service. There is also the possibility of more recent admixture, since cultural taboos about mixed race relationships and legal prohibitions fell away in the last few decades of the 20th century. These would lead the population of modern day Afro-Americans

to be mostly descended from a west African subpopulation like the Yoruba from Nigeria (YRI) but also to have an element of European ancestry, which again could be CEU, TSI or an ancester of both.

In a more general situation, it may not be the case that the history of the subpopulation is known from other sources. How could the need for modelling a subpopulation as an admixed population be identified then? One way would be to examine a post predictive checking table such as that in table 4.3. There it can be seen that ASW, has a large number, 8 out of 10, of predictive p-values below 0.025. A low p-value in that case indicates that ASW is more closely related to the subpopulation to which it is being compared than the model that led to that post predictive check allows. There ASW was placed in a branch of the phylogenetic tree with African subpopulations, which makes sense but how can it be simultaneously kept close to those subpopulations in the tree but also moved nearer to the non-African subpopulations without simultaneously moving the other African subpopulations? It can if it were modelled as an admixture of an African subpopulation with a non-African subpopulation. Its lowest p-values are 0 for the two European subpopulations, CEU and TSI, and the Gujarati subpopulation GIH. This suggests that one of these population's ancestors or their common ancestors would be good candidates to be one of the two populations contributing to the admixture. The other contributing subpopulation would be African. Of the African subpopulations, the Yoruba from Nigeria, YRI, also has a low p-value of 0.003 indicating that it should be more closely related to the Afro-Americans than in that model so taking the other contributing subpopulation to be its ancestor is worth trying. The other subpopulation with a lot of low predictive p-values in table 4.3 is the Mexican one, MEX, which has low p-values with ASW and the three east Asian subpopulations, CHB, CHD and JPT. For that model the Mexicans were placed on a branch of the tree that included the two European subpopulations, TSI and CEU. Assuming that modelling the Afro-American subpopulation as an admixture as described above eliminates that low p-value, that suggests that

the Mexicans are more closely related to the east Asian subpopulations than that model allows, so an admixture involving an ancestor of the European subpopulations and an ancestor of the east Asian subpopulations would be a promising candidate as an admixture. The process could proceed by running a model with the Afro-American admixture, examining the resulting post predictive check table to make sure that the Afro-American subpopulation is now modelled adequately. This table may still suggest that the Mexican subpopulation needs to be modelled by the sort of admixture that has just been described. The next step would be to run a model with both these admixtures and to examine the resulting post predictive check table to consider whether further admixtures are required to adequately model the data. This process does, however, have the obvious downside that it is iterative and involves some trial and error of running models that may well take some days to accumulate a sufficient number of iterations of the Gibbs sampler to provide a sufficiently representative posterior distribution. In practice with the HapMap data, 100,000 MCMC iterations were used, taking a little over a week in each case for 2,000 loci. 100,000 is the number of iterations found, partly by accident, to be adequate in section 5.4.1 after an automated Windows shutdown at about this number.

5.3.3 Identifiability of Parameters Near Admixture Events

The model of admixture described thus far has a drawback. To understand what it is, consider figure 5.1 again. Consider what happens between the admixed population's two ancestor populations at B and C and the present-day admixed subpopulation at H. There are three periods of genetic drift, between B and D, and C and E, before the admixture event and between F and H after the admixture event. Imagine the case of an allele for which fixation is not a realistically likely outcome during these time periods. The possibility of fixation is being put to one side for now for simplicity, to make the problem easier to understand. The

proportion of an allele, α_D which is modelled as $\alpha_D \sim \mathrm{N}^{\mathrm{R}[0,1]}\left(\alpha_B, \alpha_B\left(1-\alpha_B\right)c_D\right)$ has rather complicated expressions for its first and second moments as described in appendix B. However, as discussed in section 4.5.5.2, over realistic values of c_D, the mean of α_D would be approximately α_B and the variance is approximately $\alpha_B\left(1-\alpha_B\right)c_D$ particularly where fixation is unlikely. Similarly, the mean of α_E would be approximately α_C and the variance approximately $\alpha_C\left(1-\alpha_C\right)c_E$ and for α_H the mean would be approximately α_F and the variance approximately $\alpha_F\left(1-\alpha_F\right)c_H$. Now, recall that $\alpha_F = w\alpha_D + (1-w)\alpha_E$. So the mean of α_H is approximately $w\alpha_D + (1-w)\alpha_E$ whose mean is in turn approximately $w\alpha_B + (1-w)\alpha_C$. This would provide an estimate for w given $\alpha_H, \alpha_B, \alpha_C$ since there would be one equation and one unknown. The variance of α_H is approximately $w\alpha_D + (1-w)\alpha_E\left(1 - w\alpha_D - (1-w)\alpha_E\right)c_H$. Putting this in terms of α_B and α_C instead of α_D and α_E will produce an expression that involves not only $\alpha_H, \alpha_B, \alpha_C$ but also c_H, c_D and c_E. So if these former 3 are known and w is identifiable, there will still not be a unique solution for c_H, c_D and c_E.

To look at this a different way, suppose the values for α_B and α_C were known with certainty for all loci to be 0.3 and 0.7, respectively, and suppose there was a lot of data for the admixed subpopulation at H so that the mean value of $\frac{x_H}{n_H}$ over all the loci was 0.6 but distributed such that it is extraordinarily unlikely that $\alpha_H = 0.6$ for all loci. The values of α_H would then be distributed around a value close to 0.6. It would then be reasonable from this to estimate that w has a most likely value of about 0.25, α_H being distributed in such a way as to be on average, three times further from α_B than α_C but this still leaves uncertainty about what the drift parameters should be. Clearly there has been some drift since if there was none, $\alpha_H = 0.6$ for all loci would be reasonably likely, which it is not. But how much of that drift took place between B and D, and C and E before the admixture event and how much between F and H after the admixture event? Even knowing α_B and α_C for all loci with certainty and having a lot of information about α_H at each locus from the data, there is still not enough

information to say. There could have been little time before the admixture event and all the α_F exactly 0.6 and the variation observed in α_H occurred due to drift after the admixture event, or the admixture event could have been very recent so that α_F and α_H would be nearly identical and all the variation occurred before the admixture event. In other words, there is an identifiability problem near the admixture event in relation to the drift parameters even though estimates of the admixture parameter are still reliable. The consequence of this is that individual drift parameters in the vicinity of admixture events have to be viewed with caution and their joint posterior distributions will reflect the uncertainty. The drift parameter after the admixture will be negatively correlated with those before it, since the more drift happened after the admixture, the less happened before it. As such, looking at overall drift in a lineage before and after the admixture event may be more reliable than the individual parameters. It would be possible to get around this problem by imposing some additional constraint by making an additional assumption. For example, in the case of the Mexicans, it might be possible to argue that no drift took place for the Europeans on the pre-admixture branch, because the journey took a relatively short amount of time. Alternatively, it could be assumed that admixture was recent and therefore no drift has taken place since admixture. However, these assumptions might be reasonable in specific cases, but they would not be reasonable in all cases, and to make them hard features of the model would involve a loss of generality. If they were imposed inappropriately, they would also lead to problems with interpretability. For these reasons, it has been decided to leave the model as it is and accept that identifiability is a problem in relation to genetic drift parameters near an admixture, which will be manifest in their posterior distributions.

Despite this issue, it is one of the useful features of Bayesian Hierarchical modelling that strong priors can be used to mitigate this problem where outside knowledge is available that allows the experimenter to believe that some values of c_H, c_D and c_E (or even w) are more likely and others less credible by adjusting the hyperpa-

rameters in the priors for these parameters or even changing the prior family. The approach used here preserves that flexibility.

So far in the discussion of identifiability, the effect of fixation has been ignored. However, this model does allow the possibility of alleles becoming fixed. How does fixation affect the discussion? The two situations where the allele has become fixed at both B and C, that is where $\alpha_B = \alpha_C = 0$ or $\alpha_B = \alpha_C = 1$ are uninformative about drift between B and D, C and E, and between F and H. In these cases the allele remains fixed regardless of how much genetic drift there has been. They are also uninformative about the admixture parameter, w. Regardless of the value of w in these cases, $\alpha_F = \alpha_B = \alpha_C$.

Next, consider the case where $\alpha_B = 1$ and $\alpha_C = 0$. In this case, regardless of the genetic drift between B and D or C and E, $\alpha_D = 1$ and $\alpha_E = 0$ so it is uninformative about these drifts. However, $\alpha_F = w$ and so α_H has approximate mean w and approximate variance $c_H w\,[1 - w]$. This provides clearer information about c_H than is available in any of the scenarios discussed so far. If this scenario were common, it would be possible that these situations would give useful information about c_H and that this would inform the situations where there is no fixation effect and thus alleviate the identifiability problem. Unfortunately, it is not likely to happen very often in practice that alleles will be fixed in opposite states on either side of an admixture event in this way. The rarity of this situation means that it is unlikely to help much: we should expect weak identifiability at best. The situation where $\alpha_B = 0$ and $\alpha_C = 1$ is similar to this except that $1 - w$ would appear in place of w.

The other possibility is that an allele is fixed on one side of the admixture but not on the other. Suppose $\alpha_C = 0$ but $0 < \alpha_B < 1$. Other situations where the allele on only one side of the admixture is fixed are analogous to this by symmetry. There is no information about the drift between C and E and $\alpha_E = 0$. Now α_D has approximate mean α_B and approximate variance $c_D \alpha_B\,[1 - \alpha_B]$. and α_F has approximate

mean $w\alpha_B$ and approximate variance, $wc_D\alpha_B\left[1-\alpha_B\right]$. α_H has approximate mean α_F whose mean is in turn approximately $w\alpha_B$. This can provide a good estimate of w. The approximate variance of α_H is $wc_D\alpha_B\left[1-\alpha_B\right]+wc_H\alpha_B\left[1-w\alpha_B\right]$. However, this does not allow c_D and c_H to be identified separately. The expression for the variance of α_H shows that for a particular variance of α_H, if c_D is larger then c_H is smaller and vice versa. c_D and c_H should be expected to be related in this way. In other words, there is still a lot of uncertainty about how much drift has occured between B and D and between F and H but less uncertainty about how much has occured between B and H. The situation is entirely uninformative about drift between C and E.

These issues can be illustrated using simulated data. Data were simulated for 1,000 loci, with sample sizes of 200 in each population in three simulations, comparable to HAPMAP data for a medium-sized chromosome. The three simulations differed in having 10%, 20% and 30% of the true values for α_B and α_C being either 0 or 1 (fixation) corresponding to increasing chance of the locus reaching fixations for different alleles on either side of an admixture event but also increasing the chance of the uninformative case of the locus having reached fixation for the same allele on both sides of the admixture event. The model was used on the datasets with the assumption that the true values for α_B and α_C were known in each case. To do this, they were held at their true values for each of the MCMC iterations. This was to enable the effect on the drift and admixture parameters of more and more alleles having reached fixation to be seen more clearly. The results of doing this are shown in tables 5.1 to 5.3.

	95% HPD Interval			
	Lower Bound	Upper Bound	Width	True Value
c_D	0.0001	0.0236	0.0235	0.03
c_E	0.0014	0.0290	0.0276	0.03
c_H	0.0305	0.0450	0.0145	0.03
w	0.2323	0.2603	0.0280	0.25

Table 5.1: Results of Using the Model on Simulated Data with α_B and α_C Held at Their True Values and 10% of These Having Reached Fixation.

Data were simulated for 1000 loci. α_B and α_C were assumed known and held fixed in the inference, 10% of which were either 0 or 1 representing fixation having been reached. The table shows the resulting 95% HPD intervals from using the model on such data for the three drift parameters around the admixture event and the 95% HPD for the admixture parameter, w.

	95% HPD Interval			
	Lower Bound	Upper Bound	Width	True Value
c_D	0.0003	0.0319	0.0316	0.03
c_E	0.0029	0.0360	0.0331	0.03
c_H	0.0280	0.0411	0.0131	0.03
w	0.2440	0.2645	0.0205	0.25

Table 5.2: Results of Using the Model on Simulated Data with α_B and α_C Held at Their True Values and 20% of These Having Reached Fixation.

Data were simulated for 1000 loci. α_B and α_C were assumed known and held fixed in the inference, 20% of which were either 0 or 1 representing fixation having been reached. The table shows the resulting 95% HPD intervals from using the model on such data for the three drift parameters around the admixture event and the 95% HPD for the admixture parameter, w.

	95% HPD Interval			
	Lower Bound	Upper Bound	Width	True Value
c_D	0.0002	0.0412	0.0410	0.03
c_E	0.0038	0.0313	0.0275	0.03
c_H	0.0267	0.0381	0.0113	0.03
w	0.2409	0.2591	0.0182	0.25

Table 5.3: Results of Using the Model on Simulated Data with α_B and α_C held at Their True Values and 30% of These Having Reached Fixation

Data were Simulated for 1000 loci. α_B and α_C were assumed known and held fixed in the inference, 30% of which were either 0 or 1 representing fixation having been reached. The table shows the resulting 95% HPD intervals from using the model on such data for the three drift parameters around the admixture event and the 95% HPD for the admixture parameter, w.

Highest Probability Density (HPD) intervals were calculated using the *boa* package

in R (Smith, 2007) which uses the algorithm described in Chen and Shao (1999). The width of the intervals for c_D and c_E over the three tables tends to increase, as the proportion of loci at fixation increases, as expected due to there being more cases that are uninformative about these parameters, even though the width of the interval for c_E in table 5.3 does buck this trend. The intervals are still very wide regardless, reflecting the uncertainty about individual periods of drift around the admixture event. Nonetheless, in table 5.1, the intervals do not quite contain the true values. They underestimate the true value. The interval for c_H in that table correspondingly overestimates its true value, again only narrowly failing to contain it. To an extent this is expected. When the drift parameters c_D and c_E, the drifts before the admixture event, are lower than the true values, the parameter c_H is usually correspondingly higher than its true value and vice versa. This is consistent with there being less uncertainty about the drift overall, through (before plus after) the admixture event, than there is for each separate period immediately before or after it. The interesting thing here is how the widths of the HPD for the c_H parameters narrow as the proportion of α_B and α_C that are at fixation increases. The discussion above showed that cases where there are opposite fixations on either side of the admixture event should provide useful information about c_H. As these cases become more common, having more information about c_H is reflected in its HPD interval narrowing. The HPD intervals for the admixture parameter w are relatively tight around the true value showing that there is much less uncertainty about it than there is for the individual drift parameters around the admixture event. The admixture parameter w is far less affected by the non-identifiability issue as was expected from the preceding discussion.

Another simulation was carried out with the structure shown in figure 5.3. 1,000 loci were simulated in samples from each of 3 populations plus an outgroup each of size 500. The larger than usual sample size was intended to reduce the variance from this source because it is the admixture that is of interest here. In this simulation the true drift parameter along each edge was 0.1 and the true admixture

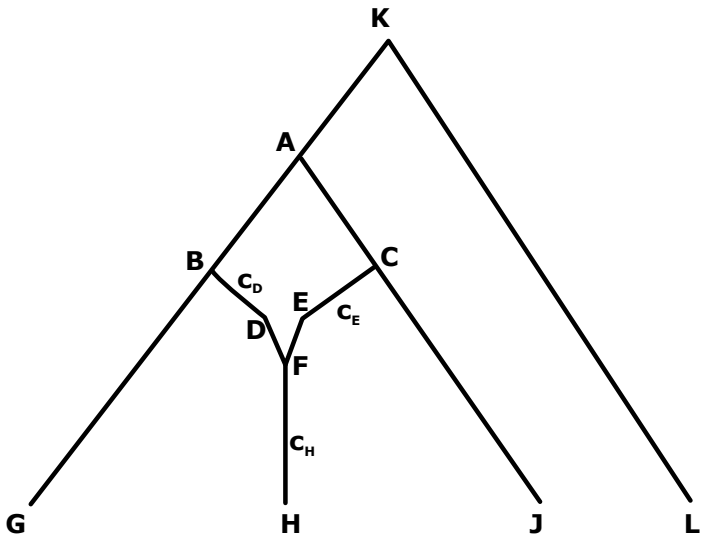

Figure 5.3: Simulation Model Used to Investigate Admixture Model Behaviour
A phylogenetic network of a simulation used to investigate the behaviour of drift parameters before and after an admixture event. The parameters of interest are c_H and c_E.

parameter was 0.5 and all alphas were drawn at each iteration by the Gibbs sampler as usual. The correlation between the drift parameters before and after the admixture is manifest in the bottom left plot in figure 5.4, where there is a clear ridge at an angle to the axes. There is a lot of uncertainty about each parameter: the 95% HPD interval for c_H was $(0.034, 0.120)$ with a median of 0.075 and for c_E was $(0.003, 0.298)$ with a median of 0.141. However taking the drift through the admixture, using the formula $c_{tot} = 1 - (1 - c_E)(1 - c_H)$ (see Appendix A) at each iteration, the HPD for c_{tot} was $(0.106, 0.348)$, narrower than that of c_E alone and the median was 0.204, closer to the true value of 0.19 than the point estimates of each individual parameter.

Looking at the other plots in figure 5.4, the top left one shows no evidence of a correlation between the drift after the admixture event, c_H, and the value of the admixture parameter, w. The other plots involving the admixture parameter do show some evidence that unusually large values of a drift parameter feeding into an admixture at an iteration, i.e., c_D or c_E are associated with an admixture parameter indicating a lower contribution to the admixture from that subpopulation. It makes sense that if a subpopulation has allele proportions that have drifted by

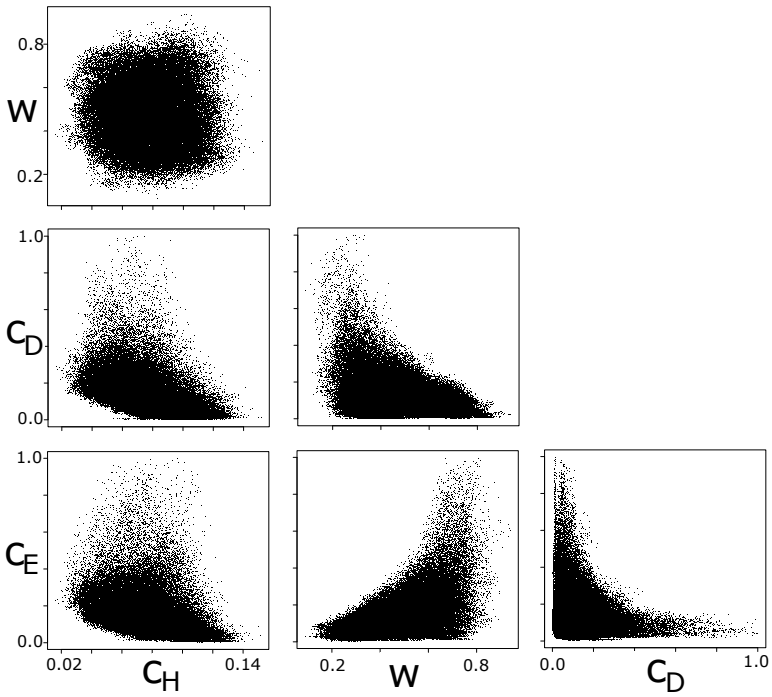

Figure 5.4: Pairwise Scatterplots of Drift Parameters c_D, c_E, c_H and Admixture Parameter w at Each Iteration.

an unrealistic amount that the model will compensate by not allowing that subpopulation to contribute so much to the admixture. There is a similar relationship between c_D and c_H, reflecting the symmetry of the model. Finally, there is an apparent inverse relationship between c_D and c_E in the bottom right plot with very high levels of drift in one parameter being associated with moderate or low levels in the other. As noted above, if one of the drifts has become unrealistic, the admixture parameter is likely to allow it a relatively small contribution to the admixture and so the other drift parameter is more likely to be realistic. While one parameter becoming unrealistically high is a possible solution if it contributes little to the admixture, the other must remain within a realistic range for the state of the model at that iteration to be reasonably probable. This again, shows that looking at the marginal posterior distributions for each drift parameter would suggest more uncertainty about these values than there is if considering pairs of them.

In conclusion, although the case where there is fixation to different alleles on either

side of an admixture event can help to alleviate the problem of identifiability in the drift parameters around the event, that case is unlikely to be sufficiently common in practical situations to help. Outside knowledge could be used to apply additional constraints on particular drift parameters to ameliorate the identifiability issue, for example, if it is known from other historical sources that one or more of the three periods of drift around the admixture event is reasonably modelled with a c equal to 0. Here, instead, no such assumptions are made. This has the advantage of keeping the model as general as possible but has the downside that there will be considerable uncertainty in marginal estimates of the drift parameters for the three periods of drift adjacent to the admixture event. Point estimates in particular should be treated with extreme caution. The uncertainty will be reflected in the posterior distributions for these parameters. Estimates of overall drift through (both before and after) the admixture event should be more reliable than the drift parameters individually. The posterior distribution of the admixture parameter, w, and therefore the estimates of the proportions of the genome that the admixed population inherits from its two parent populations does not suffer from this problem to anything like the same extent.

5.4 Application to the HapMap Dataset

To illustrate how the admixture models in this chapter are represented figure 5.5 shows a four subpopulation model with an admixture. It involves only Han Chinese in Beijing (CHB), Mexicans (MEX), Italians in Tuscany (TSI) and Lhosa in Kenya (LWK). The MEX are modelled as an admixture of the TSI and the CHB. The Mexicans can be thought of as an admixture of European colonists of America, the Spanish conquistadors, and the Native Americans that were already living in Mexico before European colonisation. These Native Americans will have descended from the people that arrived in America by crossing what is now the Bering Strait from east Asia during the period when there was a land bridge at

that location, that is to say, at the time when there was no water in the strait (Elias et al., 1996). These people in turn would have had a common ancestor in East Asia with the Han Chinese, making the Chinese the best of the available subpopulations to represent the native American component in the ancestry of the Mexicans.

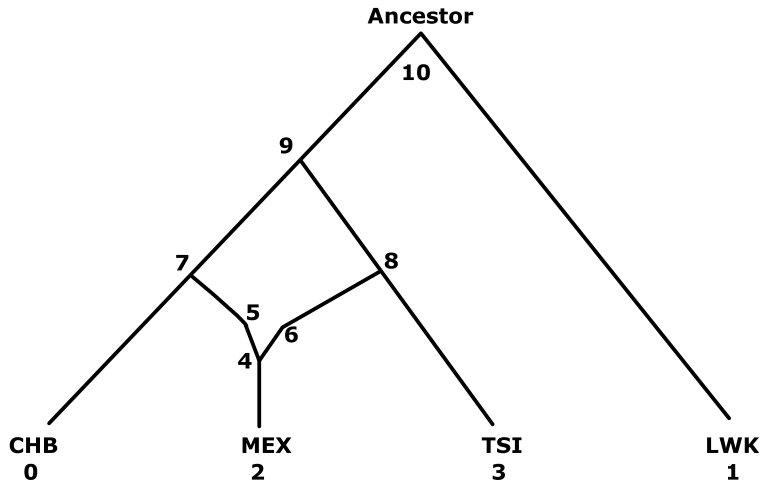

Figure 5.5: Four Subpopulation Model
A model of four present-day subpopulations featuring an admixture event for the MEX subpopulation. The edges represent periods of genetic drift. The common ancestral population is at the top and is the root of this phylogenetic network.

The network in figure 5.5 shows the present-day subpopulations at the bottom with these leaves numbered from 0 to 3. All the ancestral subpopulations at bifurcation points and around the admixture event are represented by higher numbers so that 4 is the ancestor of the Mexicans just after the admixture event. Nodes 5 and 6 are the two ancestral populations of the Mexicans, the descendants of East Asians and Europeans respectively, just prior to the admixture event. Node 7 is the common ancestor of 5 and the present-day Han Chinese at the point where the two subpopulations diverged. Node 8 is the common ancestor of 6 and modern-day Tuscans. Node 9 is the common ancestor of 7 and 8, and node 10 is the common ancestor of all the other subpopulations in this model. Table 5.4 shows the posterior median estimates and 95% credible intervals for the drift along each of the edges in the figure based on 40,000 posterior samples. Note that w represents the

proportion of the resulting admixed subpopulation's genetic information that has been inherited from the lower numbered of the two contributing subpopulations. So the point estimate for w_4 of 0.4474 means that 44.7% of the admixed population's genomes come from subpopulation 5, the one descended from the East Asians, and 55.3% comes from subpopulation 6, the population of European descent. The 95% credible interval for w_4 ranges from 38.4% to 50.2%, so the European contribution ranges from 48.8% to 61.6%.

Table 5.5 shows the post predictive checking table for the model. All of the values are within the $[0.025, 0.975]$ interval so none are particularly high or low, suggesting that the model represents the relationships between the four subpopulations reasonably well, at least in terms of the F_{ST} statistic.

Table 5.4: Parameter Estimates for the Model in Figure 5.5

	Bounds for 95% HPD Interval		
Parameter	lower	upper	Median
c_0	0.0726	0.1089	0.0911
c_1	0.0796	0.1080	0.0933
c_2	0.0001	0.0011	0.0005
c_3	0.0168	0.0338	0.0248
w_4	0.3837	0.5024	0.4474
c_5	0.0378	0.1122	0.0723
c_6	0.0001	0.0021	0.0008
c_7	0.0762	0.1142	0.0936
c_8	0.0201	0.0508	0.0353
c_9	0.0916	0.1255	0.1083

The table shows the resulting 95% HPD intervals for the drift parameters, $c_i (i = 0, ..., 3, 5, ..., 9)$, and the admixture parameter, w_4. The subscripts for the drift parameters refer to the node in the figure where the drift ends. For the admixture parameter, it refers to the node at which the admixture takes place.

p-value	CHB	LWK	MEX	TSI
CHB	X	0.9086	0.6952	0.8861
LWK	0.9086	X	0.8548	0.9697
MEX	0.6952	0.8548	X	0.2559
TSI	0.8861	0.9697	0.2559	X

Table 5.5: Predictive p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.5.
Values near 0 indicate that the subpopulations are more closely related in the data than in the model. Numbers close to 1 indicate the opposite.

Moving to the full complement of the HapMap dataset, a range of models will now be examined and compared to those produced in the previous chapters. Table 5.6 provides a summary of the models including the WAIC for each model, the number of very low or very high predictive p-values and whether an implausibly high drift parameter is present, which may be suggestive of a misspecification. WAIC was used for model comparison because of the convenient comparative ease with which it can be calculated from the posterior distribution chains. The first batch of models in the middle of that table are based on adding admixtures to the tree suggested by the neighbour joining algorithm. The second batch in the lower part of that table are based on adding admixtures to an alternative tree structure.

5.4.1 Models Based on the Neighbour Joining Algorithm Tree

In the previous chapter, a purely tree-like model without any admixture events was considered and when its associated post predictive checking table (table 4.3) was considered was found to be inadequate to describe the data. Examining that table revealed that the Afro-American subpopulation (ASW) needed to be more closely related to the European subpopulations while still needing to be like the African subpopulations. This strongly suggested an admixture relationship. Nigeria is on the coast of Africa from which most Africans were involuntarily migrated

Model Figure	Admixtures	Implausibly High drift Parameters	Number of Predictive p-values which are <0.025	>0.975	WAIC
3.12	-	-	18	34	138,874
4.7	-	c_{14}	14	12	129,353
5.6	ASW	c_{15}	5	6	129,183
5.7	ASW, MEX	c_{14}	1 (MKK/TSI)	4	129,106
5.11	ASW, MEX, MKK	c_{18}	0	3	129,123
5.12	MEX	c_{12}	9	9	129,262
5.13	-	c_{12}	12	12	129,364
5.14	GIH	c_{12}	10	11	129,293
5.15	GIH, MEX	c_{12}	10	8	129,306
5.16	GIH, MEX, ASW	c_{12}	1 (MKK/TSI)	4	129,141
5.17	GIH, MEX, ASW, MKK	c_{12}	0	1 (CEU/YRI)	129,142
5.18	GIH, ASW, MKK	c_{12}	0	4	129,124
5.19	MEX, ASW, MKK	-	0	4	129,156
5.20	GIH, MEX, MKK	c_{12}	9	8	129,390
5.21	MEX, ASW	-	1 (MKK/TSI)	7	129,162

Table 5.6: Summary Table of Admixture Models

Models in the top section are from previous chapters. Models in the middle section are based on the structure suggested by the Neighbour Joining algorithm. Models in the bottom section are based on the structure suggested by TreeMix.

to America, so the YRI subpopulation is the most likely candidate to be most closely related to the African ancestor population from which Afro-Americans are descended, while the European ancestors of African Americans could have come from many parts of Europe and so both European subpopulations could be descended from it. Such an admixture model is shown in figure 5.6. Table D.1 shows the posterior medians and 95% Highest Probability Density (HPD) interval based on 100,000 posterior samples (of which the first 10,000 were discarded as burn-in) for the drift and admixture parameters for that model. As will be explained shortly, this was found to be an adequate number of iterations by accident after an automated Windows shutdown at around that number. Of interest is that it estimates the proportion of the Afro-American's genetic heritage that is European to be between 18.9% and 21.1%. This might sound like a higher estimate than might be expected. However, previous studies such as that of Bryc et al. (2015) produce estimates using different datasets and different techniques that are similar, if anything, a little higher (24%). That this estimate is similar to those found by previous work is encouraging. Of the estimates of genetic drift, c_{15} at between 0.130 and 0.153 seems surprisingly large and raises suspicions that the model is misspecified somewhere. One possibility is that the branch to the Gujarati (GIH) at node 16 could be misplaced. The post predictive check table (table D.2) associated with the model still shows the Mexicans as being more closely related to the East Asians than this model allows. This model has a WAIC of 129,183 which compares favourably to the model from the previous chapter which had a WAIC of 129,353 and the simpler model from Chapter 3 which had a WAIC of 138,874.

The next model includes an admixture model for the Mexicans as well as the Afro-Americans in response to the low predictive p-values in the relationship between them and the three East Asian subpopulations. It is shown in figure 5.7. The estimates and 95% HPD intervals for the parameters are given in table D.3 based on 102,000 samples. (It was intended to run the sampler for more iterations, however the process was interrupted by an automatic Windows operating system

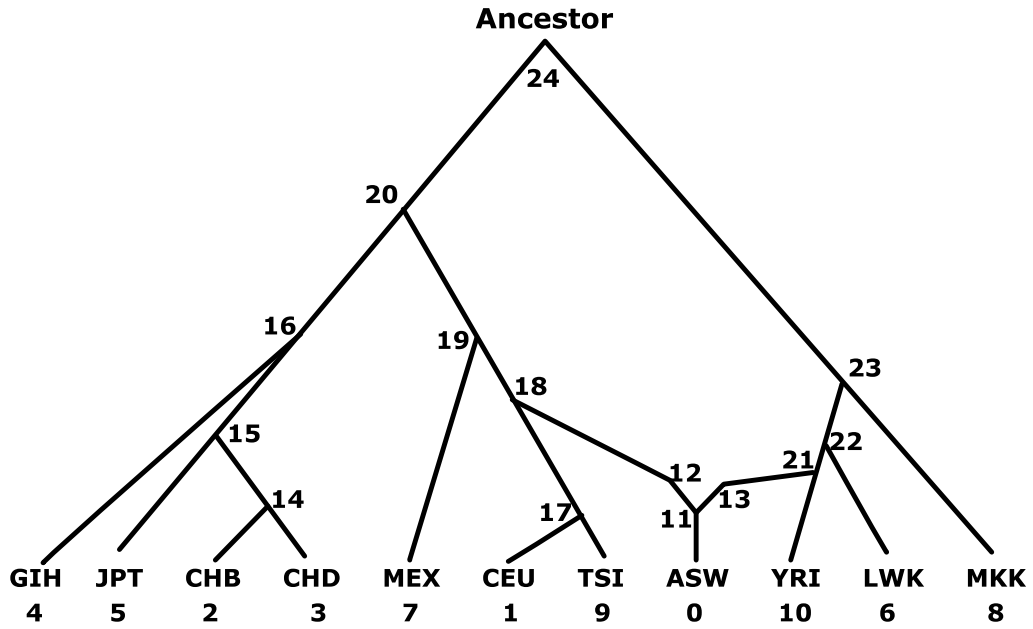

Figure 5.6: Model with Afro-American Admixture
A model of all eleven present-day subpopulations in the HapMap dataset featuring an admixture event for the Afro-American (ASW) subpopulation.

shutdown. At that stage there were 102,000 samples from the joint posterior distribution saved to disk. Examination of the posterior distributions suggested this was an adequate number and models after this batch were run for the similar, but rounder number of 100,000 iterations.) This model has a large estimate of the drift leading up to the Mexican admixture with c_{14} being between 0.144 and 0.274. As noted before, the drift estimates around any admixture event should be treated with some caution. It could be argued that if the migrations across the Bering Strait involved only a small population this could lead to this edge having a large genetic drift parameter. Nevertheless, it still seems more likely that the model is misspecified in some way. The post predictive checking table (table D.4) has only one very small value, for the Maasai (MKK) and Tuscan (TSI) pair, suggesting that these are more closely related than the model allows. There are also a few high values for the Central Europeans (CEU) with the Denver Chinese (CHD), Lhosa (LWK) and Yoruba (YRI) so there are still a few issues with this model. The WAIC for this model is 129,106, making it the best model so far.

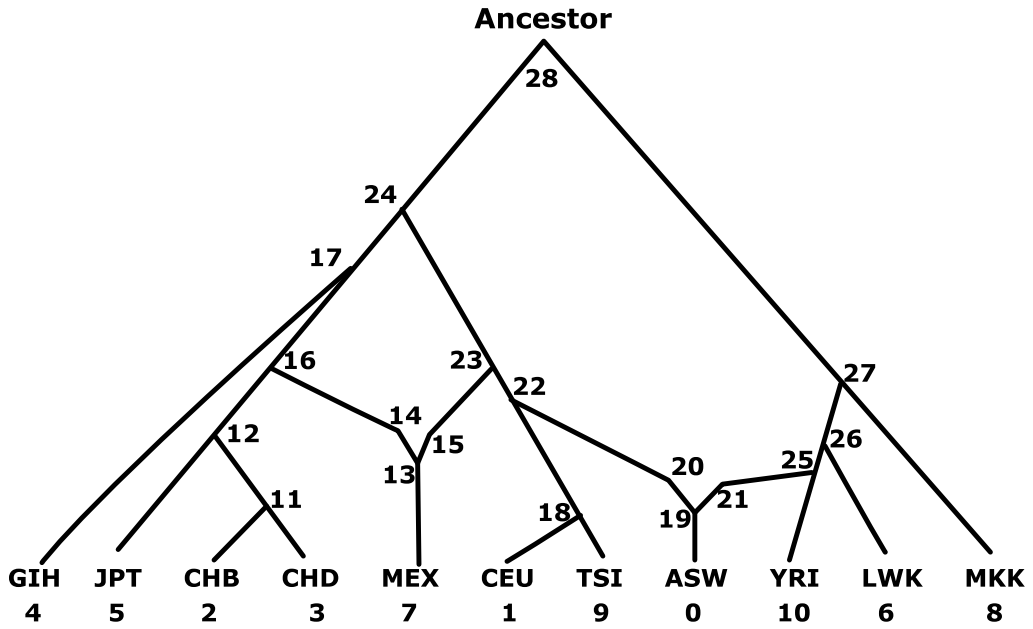

Figure 5.7: Model With Admixture events for ASW and MEX
A model of all eleven present-day subpopulations in the HapMap dataset featuring admixture events for the Afro-American (ASW) and Mexican (MEX) subpopulations.

The high drift parameter leading to the Mexican admixture could suggest the model is misspecified somewhere near that edge. In the models so far, the terminal edge for the Gujaratis (GIH) has branched from the Asian side just before the branch towards the Mexican admixture. How does removing the Gujaratis affect the size of the drift parameter leading to the Mexican admixture? The model in figure 5.8 is intended to answer this. As can be seen from the parameter estimates for that model in table D.5, the drift parameter for the branch leading to the admixture, c_{12}, is much reduced to between 0.046 and 0.117, suggesting that the Gujarati may be misplaced in the model in some way. The proportion of the Mexican genomes deriving from their Asian ancestry, w_{10}, is estimated to be between 39.2% and 48.4% which is a little higher, but still broadly consistent with the previous model. The exclusion of the Gujarati leaves the proportion of European ancestry among Afro-Americans, w_{11}, at between 19.1% and 21.4%.

So, if the Gujarati are wrongly placed in the model, where would be better? A number of alternatives were tested. Firstly, what if the branch going towards the

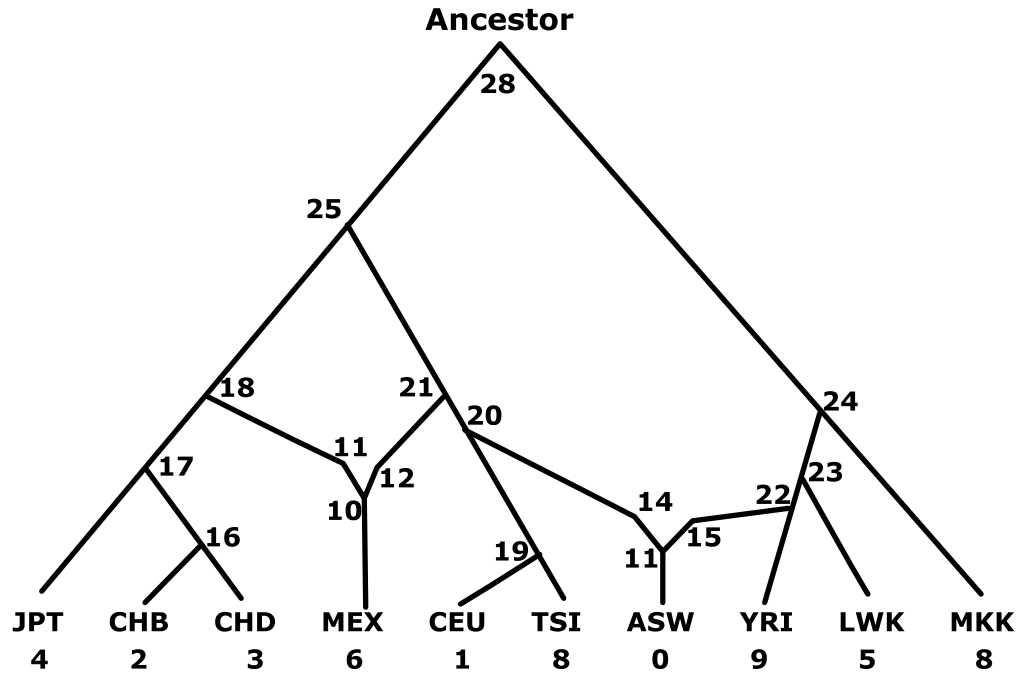

Figure 5.8: Model with Admixture for MEX and ASW without the Gujarati (GIH)

Mexican admixture happens before the branch to the Gujaratis? This is the model in figure 5.9. Table D.6 shows that the proportion of the Mexican genome that has European ancestry, $1 - w_{14}$, falls sharply to between 31.9% to 43.9%, almost 20% lower than in the model without Gujaratis. The amount of drift between the Gujarati branch at 13 in the figure and the branch where the Japanese branch off, c_{12}, is rather high at between 0.135 and 0.159. Finally, the post predictive p-values in table D.7 for both the Mexicans (MEX) with the Beijing Chinese (CHB) and Japanese (JPT) are very low even with the admixture. Overall, this model is even less plausible than those considered earlier.

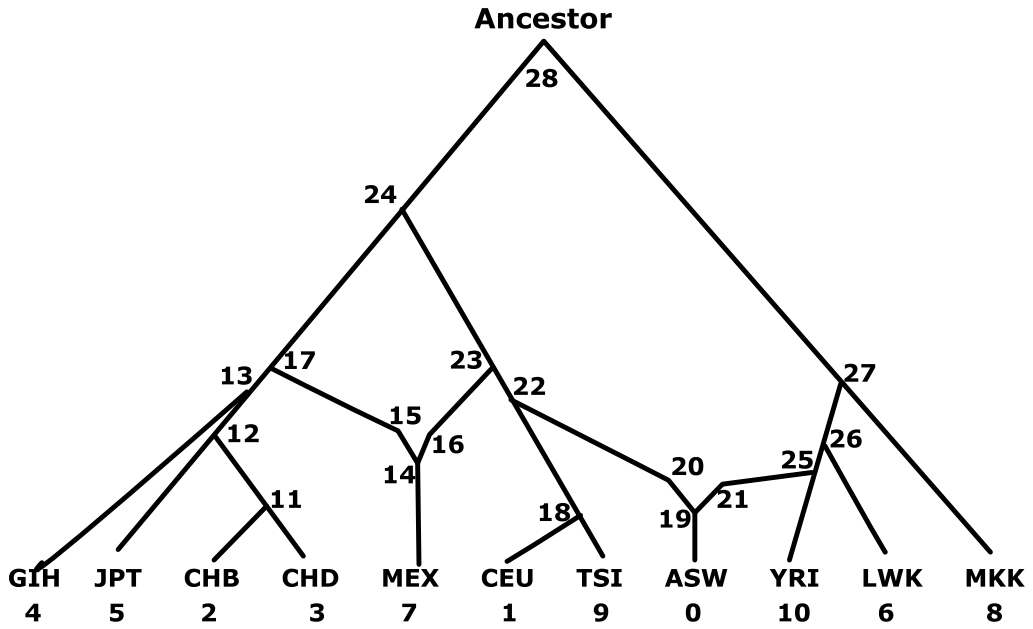

Figure 5.9: Model with Admixture for MEX and ASW and with GIH Placed Nearer East Asians

What if the branch to the Gujaratis were instead moved above the European/Asian ancestor (figure 5.10)? Examining the drift and admixture parameters in table D.8, the Asian drift parameter before the Mexican admixture, c_{12}, has a 95% HPD interval from 0.096 to 0.198. While the lower end of the interval is not incredibly huge, it is still high. Looking at the predictive check table (table D.9) in this case, the only very small value is for the pair of Maasai (MKK) and Tuscans (TSI). There is also a small value for the pairing of Mexicans and Maasai (MEX and MKK). While this does seem a plausible candidate model, its WAIC of 129,184 is rather larger than the 129,106 for the model with the Gujarati in their original position.

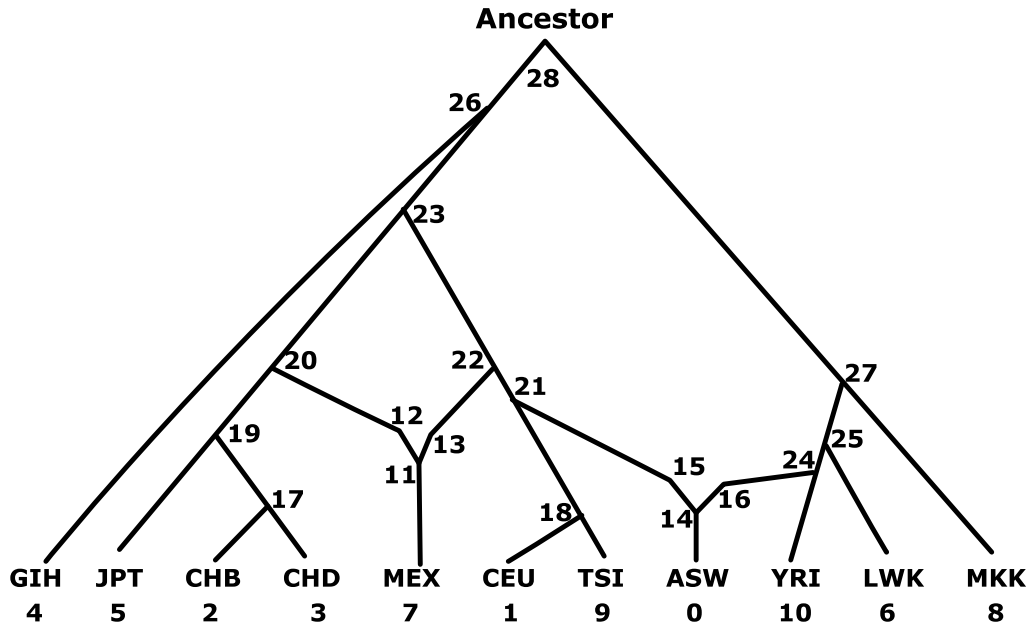

Figure 5.10: Model with Gujarati branching before the European/East Asian Ancestor

Returning the Gujarati to their original position, it is a curious feature of the better models so far that the post predictive p-values are low for the pairing of Maasai (MKK) and Tuscans (TSI) meaning that they are more closely related than this model reflects. The central Europeans (CEU) are closely related to the Tuscans, but there is no correspondingly low value for their pairing with the Maasai. This suggests there is a specifically South European relationship to the Maasai. It is true that Italy was a colonial power in the area of East Africa near Ethiopia and Somalia (then known as Abyssinia and Italian Somaliland) (Oliver and Fage, 1970). However, this seems too recent to create such a close genetic link between the two subpopulations. There is another theory that the Maasai are descended from Roman soldiers. Its proponents point to the traditional footwear, weapons and red cloak of the Maasai saying that they resemble designs from ancient Rome (Saruni, 2016). However, no support for this idea could be found among serious academic historians. More likely, both populations could be related to a third subpopulation that is absent from the HapMap dataset, such as a Near Eastern or North African one. It might also be tempting to dismiss the p-value as spurious and put it down to random chance. It is, however, a feature of some East African populations

that has been noted by others such as Pickrell et al. (2014), who suggest a back migration from Western Eurasia occured into East Africa and admixtures with African populations went all the way to Southern Africa. Much earlier work by Cruciani et al. (2002) suggests a similar back migration into sub-Saharan Africa from Asia. These could support both groups being related to a third Eurasian group but that would make it curious that the Central Europeans are not so closely related to the Maasai when the Tuscans are. Llorente et al. (2015, 2016) suggest Eurasian DNA in modern East African populations could be as much as 25%. So, could the Maasai be modelled as an admixture of Tuscans and other Africans? This is reflected in the model shown in figure 5.11. The 95% credible intervals for the resulting drift and admixture paramers are shown in table D.10. This still has the problem that the drift on one of the branches preceding the admixture for the Mexicans, c_{18}, is unrealistically high at between 0.150 and 0.267, but there are no such problems on the branches around the other two admixtures. This suggests that the Maasai have an ancestry that is between 19.0% and 23.1% Tuscan-like and between 76.9% and 81.0% sub-Saharan African. The former is similar to the proportion of European ancestry in Afro-Americans. Table D.11, of the pairwise post predictive p-values, has no very low values in it. There are a few very high values. The pairings of CEU with CHD and with YRI, as well as that of CHD with the Tuscans TSI are rather high suggesting that they are not as closely related as this model suggests but there is no obvious way to modify the model to reflect that without disturbing the other relationships within it. It has a WAIC of 129,123 which is higher than the 129,106 for the earlier model without the admixture for the Maasai. The difference in WAIC is not large but it does suggest that the extra complexity of this model with three admixtures is not justified by the improvement in the way the model represents the data.

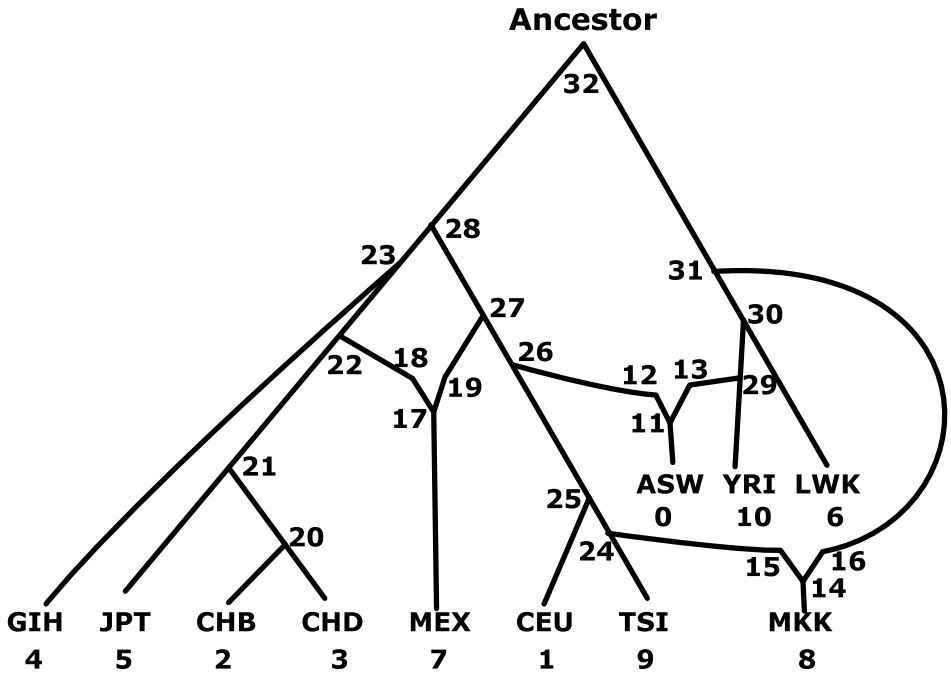

Figure 5.11: Model with Mexican, Afro-American and Maasai Admixtures

A simpler model with just a Mexican admixture was fitted as shown in figure 5.12. The parameter values for this model are shown in table D.12. The drift parameter to the Mexican admixture on the Asian side, c_{12}, is still high at between 0.159 and 0.281. There are a number of low predictive p-values (table D.13), particularly involving the Afro-Americans (ASW). This model has a high WAIC of 129,262 so does not represent the data as well as many of the other models considered so far.

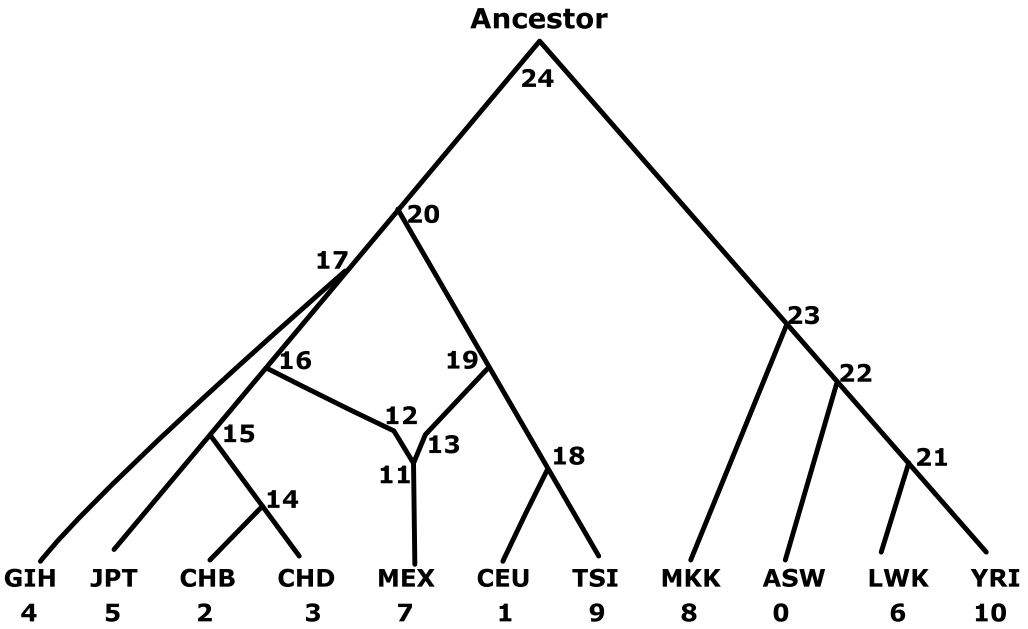

Figure 5.12: Model with Admixture for just Mexicans

5.4.2 Models Based on the TreeMix Tree

So far the models have all had a problem with at least one suspiciously high drift parameter suggesting a misspecification and the position of GIH on the network being a likely cause. Next the admixture adding process is restarted from a different tree model. First the case with no admixtures is considered. This tree is a model suggested by the method of Pickrell and Pritchard (2012) implemented in software called TreeMix which will be discussed in more depth below. The model structure is shown in figure 5.13. It differs from the Neighbour Joining tree in the last chapter (figure 4.7) in the position of GIH on the European branch and MEX on the Asian branch. These have swapped position compared to the Neighbour Joining tree. Given the problems that have been experienced thus far by the positioning of the Gujarati, this alternative tree is worth considering. The drift parameter values for this model are shown in table D.14. The drift parameter, c_{12}, is high at between 0.134 and 0.156 suggesting that there may still be something misspecified near the East Asian branch. The obvious explanation is

that the Mexicans are still modelled as being purely on the Asian branch when they should be admixed with Europeans. This is confirmed by the low predictive p-value (table D.15) for the pair of MEX with CEU. There are also low values for ASW with all non-African subpopulations, strongly suggesting an admixture, and for the pairing of TSI with MKK which has been noted before. The WAIC for the model is 129,364 which is not better than the 129,353 for the model with no admixtures based on the neighbour joining tree.

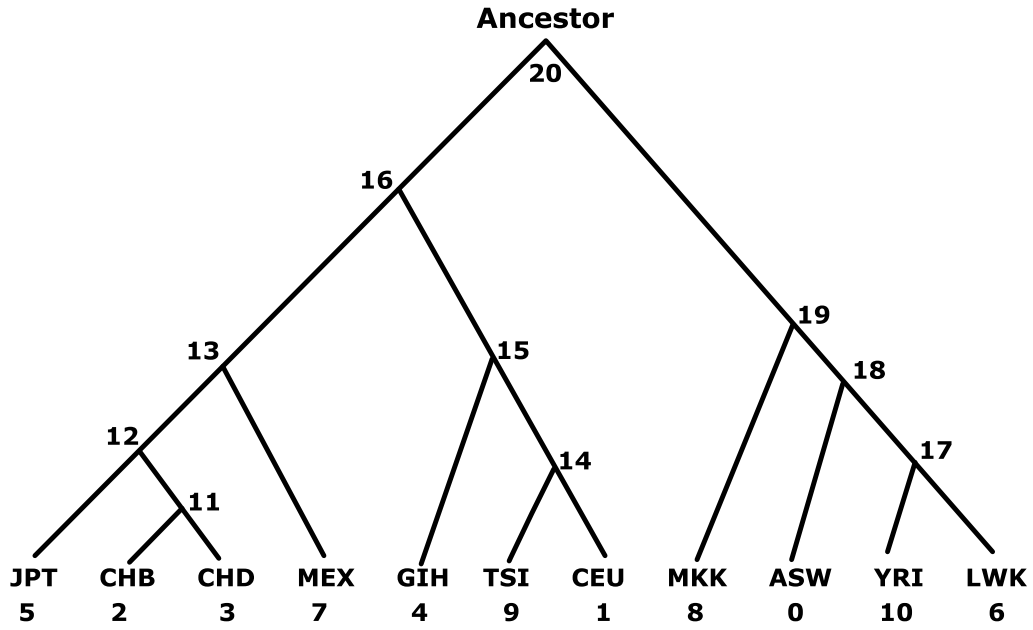

Figure 5.13: Model suggested by TreeMix with no Admixtures
A model of all eleven present day subpopulations in the HapMap dataset as suggested by TreeMix.

Interestingly, the first admixture that TreeMix suggests relates not to the Mexicans or Afro-Americans but is one for the Gujaratis. It suggests an admixture involving the Europeans and the Chinese. This is perhaps plausible given that Gujerat was historically on the trade routes running between Arabia and Indo-China (Sharma, 2014). This leads to the model structure shown in figure 5.14. The drift and admixture parameters for this model are given in table D.16. The model suggests that the Gujarati have between 21.2% and 26.7% ancestry with Chinese (and between 73.3% and 78.8% with Europeans). However, the drift parameter leading from the Chinese branch to the admixture, c_{12}, is very large at between 0.166 and

0.437 and is difficult to justify. As might be expected, the predictive p-values (table D.17) are very small for several pairings involving the Afro-Americans (ASW) as well as the pairing of Maasai (MKK) with Tuscans (TSI). The WAIC for this model was 129,293 and so was an improvement of only 71 on the previous model.

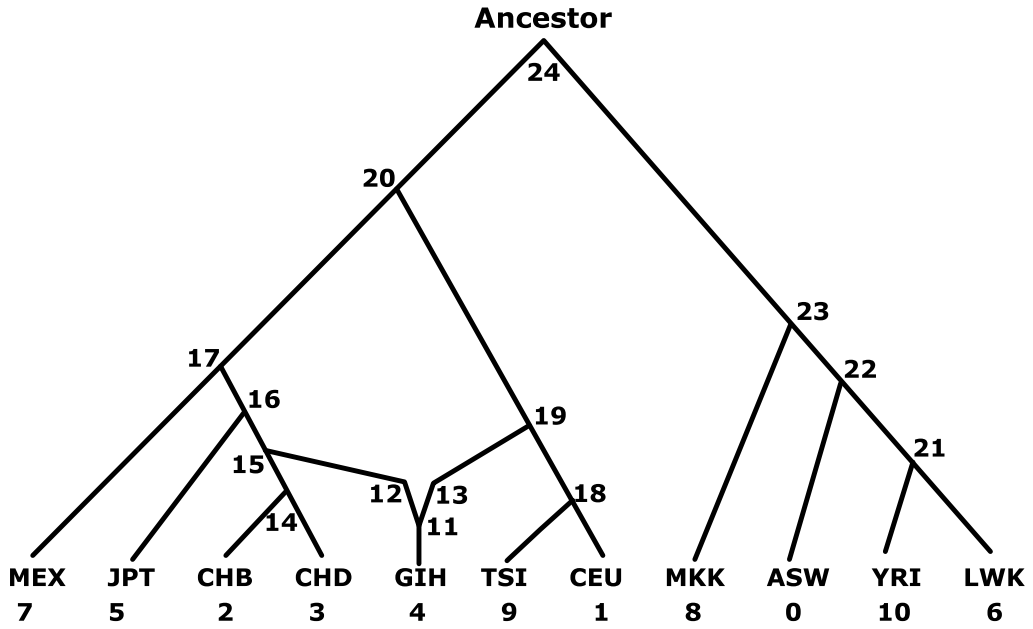

Figure 5.14: Model with an Admixture for the Gujarati Subpopulation
A model of all eleven present day subpopulations in the HapMap dataset featuring an admixtures for the Gujarati (GIH) subpopulation.

The next admixture that TreeMix suggests is not for the Afro-Americans as might be expected, but for the Mexicans. It connects the Mexicans with the Central European (CEU) branch only and so is slightly different from the way the Mexican admixture has been treated earlier. The model is shown in figure 5.15. The resulting admixture and drift parameters are shown in table D.18. The admixture parameter for the Mexicans is now between 50.9% and 58.6% Asian, so is now slightly more Asian than European, compared with the opposite in the preceding models (such as figure 5.7) but is still within the range of values found from other studies. Lisker et al. (1995) note that the European contribution to modern Mexican DNA has been estimated variously as between 34.8% and 70.8%. In trying to account for this wide variation, they suggest, by considering the places

and groups that the samples in each previous study were drawn from, that this may be due to samples being drawn from different social strata, with the lowest social strata having the highest levels of native American ancestry. There are now no unrealistically high drift parameters leading to or from the Mexican admixture. The drift parameter leading to the Gujerat admixture, c_{12}, is still high at between 0.105 and 0.339. This is, however, smaller than in the previous model and the bottom end of that range is not unreasonable. The predictive p-values (table D.19) are still low for some pairs involving the Afro-Americans (ASW) as well as for the pairing of the Maasai (MKK) and Tuscans (TSI). The WAIC of 129,306 is not an improvement of over the previous model and nowhere near the best of those examined so far.

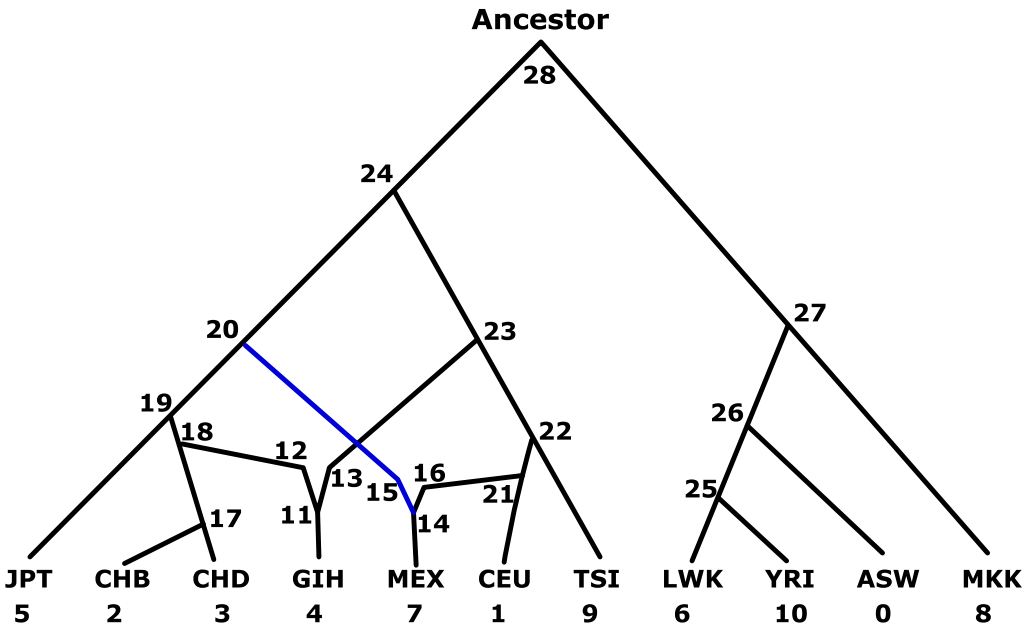

Figure 5.15: Model with Admixtures for the Gujarati and Mexicans
A model of all eleven present day subpopulations in the HapMap dataset featuring an admixtures for the Gujarati GIH and Mexican MEX subpopulations. The blue line in this figure and the ones which follow is a period of drift just like the black lines but can be thought of as passing behind the black lines that cross it without touching them.

TreeMix next suggests adding an admixture for the Afro-Americans with connections to the European branch and the Yoruban branch. This is similar to the Afro-American admixture considered earlier. This model is shown in figure 5.16.

The table of drift and admixture parameters for this model is in table D.20. The proportion of Afro-American genomes that is of European heritage is between 18.7% and 20.9%, which is in line with that seen in previous models. Parameters in the non-African part of the tree have not changed much from the previous model. Those in the African part of the tree are plausible. The predictive p-values in table D.21 are much improved from the previous model with only the only very low value being for the pairing of the Maasai (MKK) and the Tuscans (TSI). The WAIC for this model was 129,141 which is a clear improvement over the previous model but still 35 more than the lowest seen so far.

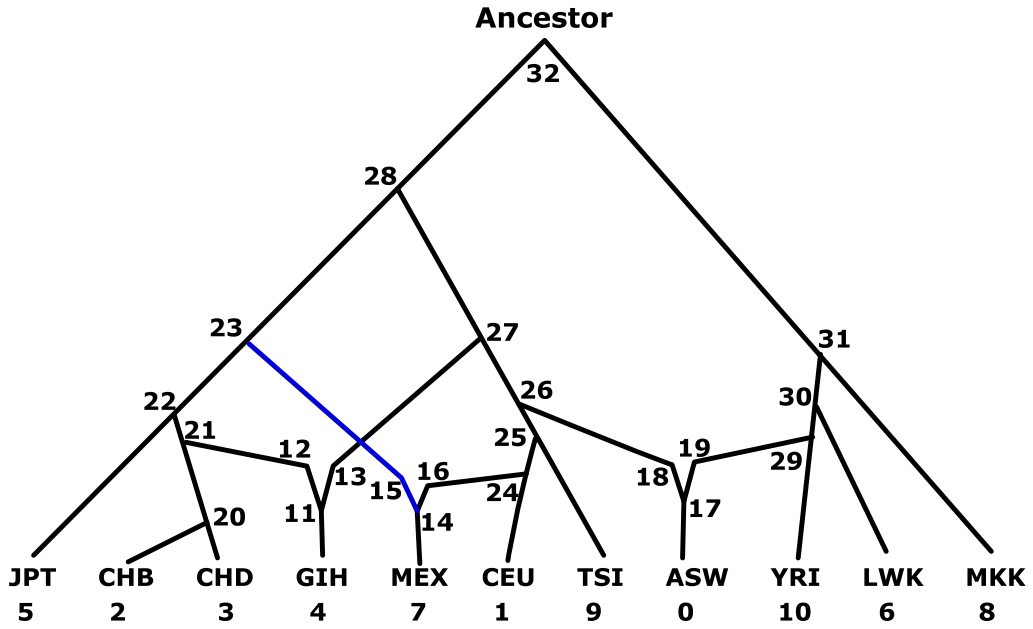

Figure 5.16: Model with Admixtures for the Afro-Americans, Mexicans and Gujarati
A model of all eleven present day subpopulations in the HapMap dataset featuring an admixtures for the Gujarati GIH, Afro-American ASW and Mexican MEX subpopulations.

The fourth admixture suggested by TreeMix is for the Maasai, mixing Tuscans and the African branch. This model is shown in figure 5.17. The drift and admixture parameters are in table D.22. The (African) admixture parameter for the Maasai of between 19.4% and 23.2% is similar to that obtained earlier (e.g., figure 5.11), an encouraging level of consistency. The drift parameter leading from the Chinese branch to the Gujarati admixture, c_{12}, is still stubbornly high at between 0.100

and 0.332, but the lower end of that range could be reasonable. No predictive p-values (table D.23) are very low and the only very high values are for the pairing of the Central Europeans (CEU) and Yoruba (YRI), making it the best model so far in terms of post predictive checks. The WAIC has however risen by only 1 compared to the previous model to 129,142, leaving it as a matter of judgement whether the extra complexity justifies its improved representation of the data.

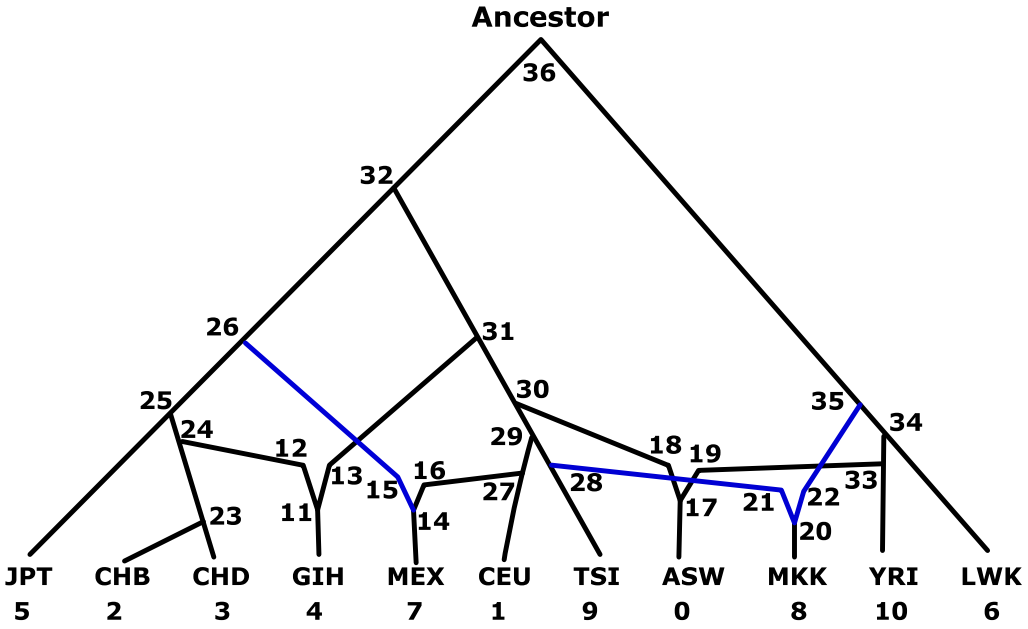

Figure 5.17: Model with Admixtures for the Gujarati, Afro-Americans, Mexicans and Maasai

A model of all eleven present day subpopulations in the HapMap dataset featuring an admixtures for the Gujarati GIH, Maasai MKK, Afro-American ASW and Mexican MEX subpopulations.

If, as the WAIC for the previous model suggests, four admixtures is too much complexity, do any of the other models with three admixtures have a lower WAIC? The model without the Maasai admixture has already been described. What if the Mexican admixture were removed? This is the model shown in figure 5.18. The drift and admixture parameters are in table D.24. The drift between the Mexican branch and the East Asian cluster, c_{22}, has grown uncomfortably to between 0.140 and 0.164 which suggests that this treatment of the Mexicans may be a model misspecification. The drift from the Chinese to the Gujarati admixture event, c_{12}, has also grown to between 0.169 and 0.415 which is moving back to

being unrealistically large. The predictive p-values (table D.25), however, are surprisingly good, with no very low values, although some high ones remain. The WAIC for the model is 129,124, which is 17 better than the model with the Mexican admixture but without the Maasai admixture but only slightly so.

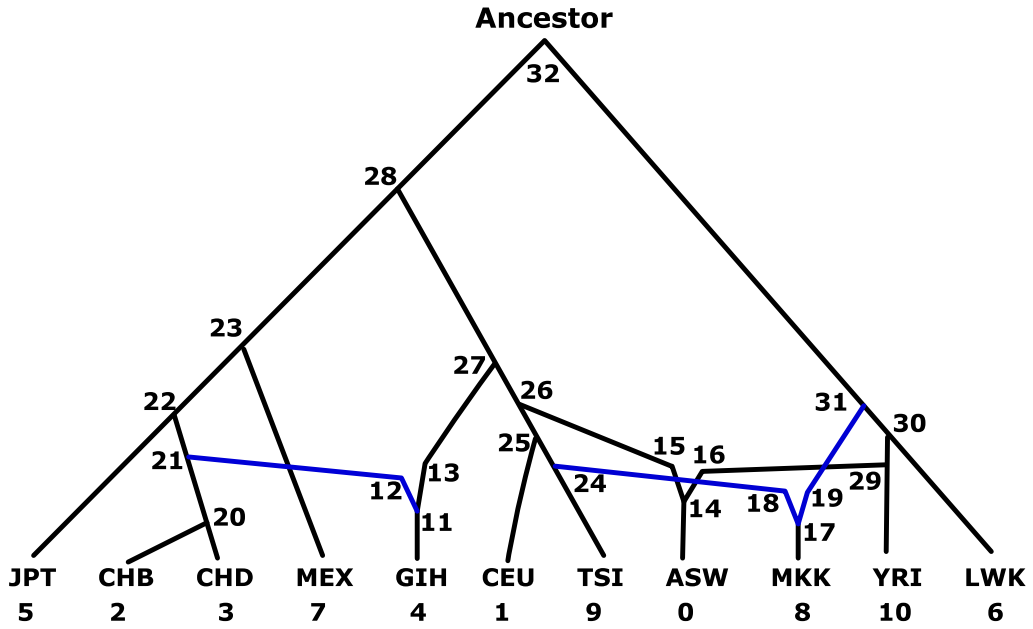

Figure 5.18: Model with Admixtures for the Gujarati, Afro-Americans and Maasai
A model of all eleven present day subpopulations in the HapMap dataset featuring an admixtures for the Gujarati GIH, Afro-American ASW and Maasai MKK subpopulations.

Next the model without the Gujarati admixure was considered. This admixture was the first of these admixtures to be suggested by TreeMix. This model considers excluding it and is shown in figure 5.19. The Gujarati are now branching off at node 27 from a European branch of the tree rather than an Asian branch. The drift and admixture parameters for this model are in table D.26. There are no incredibly large drift parameter values for this model. This is the first model examined that has this property, suggesting that a satisfactory specification may have been achieved or be close. The drift parameter from the Asian branch down to the Mexican admixture, c_{12}, which was a problem for models with other positions of the Gujarati branch is now between 0.048 and 0.090. The admixture parameters are similar to those seen in earlier models for their respective admixtures. The

predictive p-values (table D.27) for this model were also very encouraging. There are no very low values in that table and only four very high values. These were for the pairings of the Denver Chinese (CHD) with the two European subpopulations (CEU and TSI) and for the Central Europeans and Gujarati (CEU and GIH) with the Yoruba (YRI). Taken together, this model looks very encouraging. The problem is that the WAIC is 129,156, a little higher than some of the models considered so far, the lowest WAIC of which was 129,106 (from figure 5.7). However since there was a suspicion of misspecification in those models, this model seems worthy of consideration.

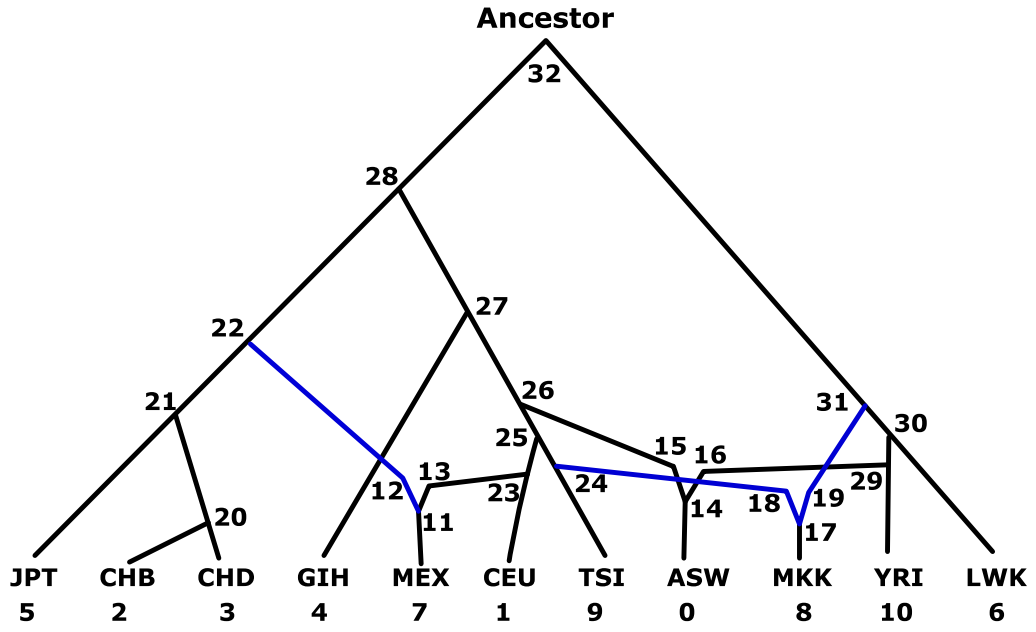

Figure 5.19: Model with Admixtures for the Maasai, Mexicans and Afro-Americans
A model of all eleven present day subpopulations in the HapMap dataset featuring an admixtures for the Maasai MKK, Afro-American ASW and Mexican MEX subpopulations.

For completeness, the Afro-American admixture can be removed. This model is shown in figure 5.20. The parameter values for the model are shown in table D.28. The problem of the large drift parameter from the Chinese to the Gujarati admixture, c_{12}, has returned, it being between 0.090 and 0.341. The lower end of that range might, nonetheless, be reasonable. What rules this model out is consideration of the predictive p-values (table D.29). The Afro-Americans (ASW)

have very low values for all subpopulations except the two Kenyan ones (MKK) and (LWK). Removing the admixture for the Afro-Americans has damaged the way the model represents their genetic relationship to eight of the other subpopulations.

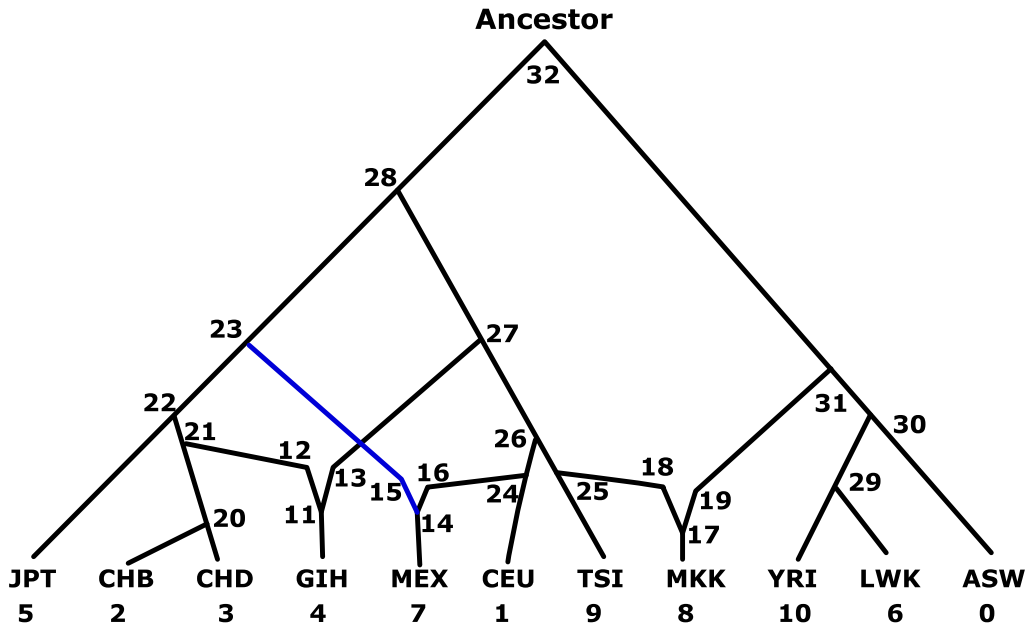

Figure 5.20: Model with admixtures for the Gujarati, Mexicans and Maasai
A model of all eleven present day subpopulations in the HapMap dataset featuring an admixtures for the Gujarati GIH, Maasai MKK and Mexican MEX subpopulations.

At this stage the model of figure 5.19 looked quite promising. It has three admixtures for ASW, MEX and MKK. Experience of the process thus far has shown that removing admixtures for ASW or MEX leads to models that do not represent the data well enough. Those admixtures are imporant if the dataset is to be adequately represented by a model. But could a simpler model still be good if the admixture for MKK was removed? This model is shown in figure 5.21. There are still no incredibly high drift parameter values (table D.30), the WAIC increases slightly to 129,162, an increase of only 6 compared to the model of figure 5.19 but there is now a low predictive p-value for the MKK and TSI pairing and an additional 4 high values (7 compared to 3 for figure 5.19) suggesting that the earlier model in figure 5.19 was a superior model of the data.

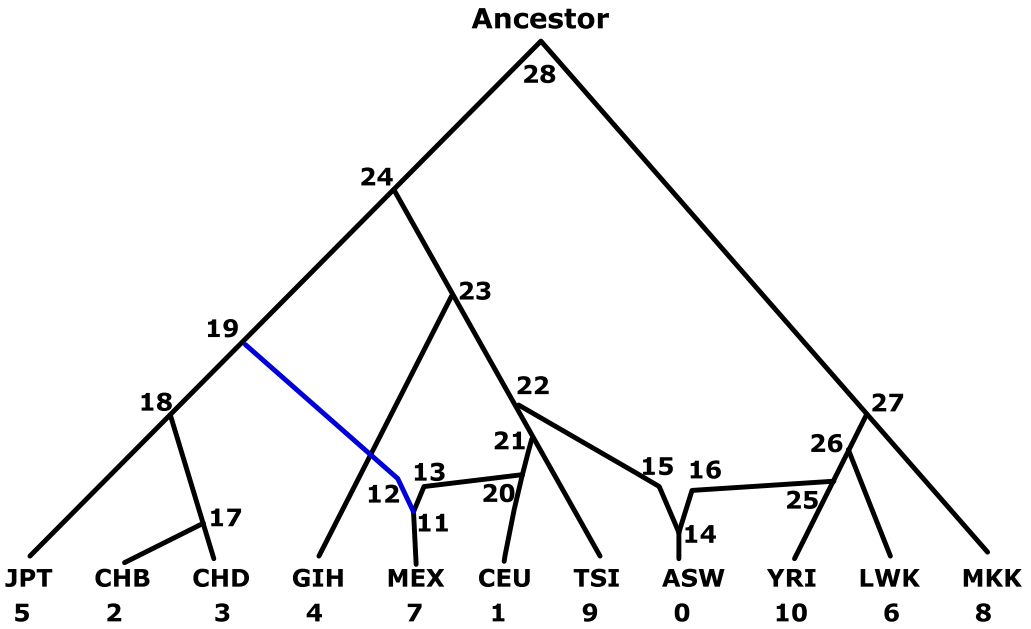

Figure 5.21: Model with admixtures for the Mexicans and Maasai
A model of all eleven present day subpopulations in the HapMap dataset featuring an admixtures for the Gujarati GIH, Maasai MKK and Mexican MEX subpopulations.

5.5 Comparison of Proposed Models

It could be argued that the best model is the one with the lowest WAIC. However, see the discussion in section 2.8. The model with the lowest WAIC would be the one shown in figure 5.7, with a WAIC of 129,106. However in table D.3, the model has a large estimate of the drift leading to the Mexican admixture (c_{14} in that table) being between 0.144 and 0.274. Even at the lower end, that is high. All of the models based on the Neighbour Joining tree had the problem of having a larger than credible drift parameter. The problem however went away when the Gujarati were removed from the dataset, strongly suggesting that the models were misspecified in the way they treat that subpopulation. The TreeMix-based models also had a similar problem until the admixture involving the Gujarati, the first admixture that TreeMix suggests, was removed. While these models have WAIC values higher than 129,106 they are not much higher. To dismiss them without considering their merits could be criticised as being overly mechanistic.

In particular, the model of figure 5.19 has a WAIC of 129,156, only 50 points higher. The post predictive checks in table D.27 for that model do not suggest that any further admixtures are required. However, removing admixtures, such as in the model of figure 5.21, results in a model that is a poorer representation of the data and no improvement in terms of WAIC. It can be argued that to select the model of figure 5.19 over the one with the lowest WAIC (figure 5.7) moves away from the objectivity of using an information criterion into making subjective judgements about the models. Reasonable arguments can be advanced for either of these models. In this case, it is judged here that the model in figure 5.19 with the more plausible parameter values and better post predictive behaviour is preferred (despite the slightly higher WAIC).

The posterior traces for all 70,080 parameters of the selected model were divided into 5 equal parts, after discarding 10,000 iterations for burn-in. Gelman's R was calculated for these to ensure the chain had converged. The results provided no reason to doubt that the model had converged for any of these parameters. Furthermore, to ensure that this finally selected model's chain had indeed converged properly, four other MCMC chains were started, each chain with different starting values for the parameters, the first with α, π and w started from 0.5 and c started from 0.1, the second with α and π started from 0.3, w started from 0.5 and c started from 0.2, the third with α and π started from 0.7, w started from 0.5 and c started from 0.05, the fourth with α and π started from 0.2, w started from 0.2 and c started from 0.05 and the fifth with α and π started from 0.8, w started from 0.8 and c started from 0.2, these were run for 80,000 iterations, the first 10,000 in each case were discarded as burn-in, providing 70,000 samples from their posterior distributions each. Gelman's R statistic was calculated for the five groups consisting of these four chains and the fifth being iterations 10,001 to 80,000 of the original chain. Again, the R statistics were nowhere near giving any cause for concern about convergence for this model.

5.6 Comparison with TreeMix Model

5.6.1 Description of the TreeMix Model

Earlier, a software package called TreeMix was mentioned, which was announced in the paper of Pickrell and Pritchard (2012). Like the method developed earlier in this chapter, this also seeks to develop a bifurcating network model representing population splits and admixture events. However, unlike the approach developed here, it attempts to do so within a frequentist framework. In order to do that it has to make a number of additional assumptions that impact on the applicability of the model and the ease of interpretation of the results. In return for making these assumptions, Pickrell and Pritchard obtain a model that has the attraction of being much less compuationally intensive, producing output within a handful of minutes as opposed to the many hours that the model described above takes. Their model will be examined critically in this section.

If the frequency of an allele at a particular locus in the ancestral population A is π_A, in a population B descended from population A the frequency of that allele α_B under a similar model of genetic drift as described by Nicholson et al. (2002) can be written as

$$\alpha_B \sim N\left(\pi_A, \pi_A\left(1-\pi_A\right)c_B\right). \tag{5.30}$$

This is similar to the same way that drift has been modelled in this and the preceding chapter. However, it has a key difference. An ordinary Normal distribution has been used instead of a Normal distribution rectified at 0 and 1 as used by Nicholson et al. (2002) and in the new models in this thesis. Pickrell and Pritchard explicitly do not model the boundary effects at 0 and 1 and so do not model fixation. This means that their model cannot be expected to be accurate in modelling drift where there are alleles near, or which may have reached, these boundaries in the present-day subpopulations. This can happen when some alleles were already near that boundary in the ancestral population or where there has

been appreciable genetic drift separating the present-day subpopulations from the ancestral population. This already restricts the applicability of the model. It will also be seen later to have a potential impact on interpretability. The model of drift can be rewritten by separating the mean and variance.

$$\alpha_B = \pi_A + \epsilon_B, \tag{5.31}$$

where

$$\epsilon_B \sim N\left(0, \pi_A\left(1-\pi_A\right)c_B\right). \tag{5.32}$$

Similarly, the frequency of the allele in a population C, α_C, that is in turn descended from B and therefore a grand-descendant of A can be described as

$$\alpha_C = \alpha_B + \epsilon_C \tag{5.33}$$

where

$$\epsilon_C \sim N\left(0, \alpha_B\left(1-\alpha_B\right)c_C\right). \tag{5.34}$$

Pritchard and Pickrell then make the additional simplifying assumption that the overall amount of genetic drift between all the populations involved in the model is small. Effectively, this restricts the model to only being applicable to data sets where the present-day subpopulations are already very closely related. This again restricts the applicability of the model. The models developed so far in this thesis make no such assumptions. These two conditions do, however, allow Pritchard and Pickrell to assume that the genetic drift between populations B and C is independent of that between A and B and that $\alpha_B\left(1-\alpha_B\right)$ is approximately the same as $\pi_A\left(1-\pi_A\right)$. Then the variance of α_C is approximately $Var\left(\epsilon_B\right) + Var\left(\epsilon_C\right)$, which is in turn approximately $\pi_A\left(1-\pi_A\right)\left(c_B + c_C\right)$. Drift parameters in series are then simply additive rather than the slightly more complicated relationship derived in appendix A which allows for larger drift parameters.

Next they consider the effects of bifurcations in the phylogenetic tree. Suppose

that population B has a second offspring population, D, in addition to C and that A has a second offspring population, E, in addition to B, so that C, D and E are the present day subpopulations, B is the common ancestor of C and D and A is the ancestor of B and E (5.22).

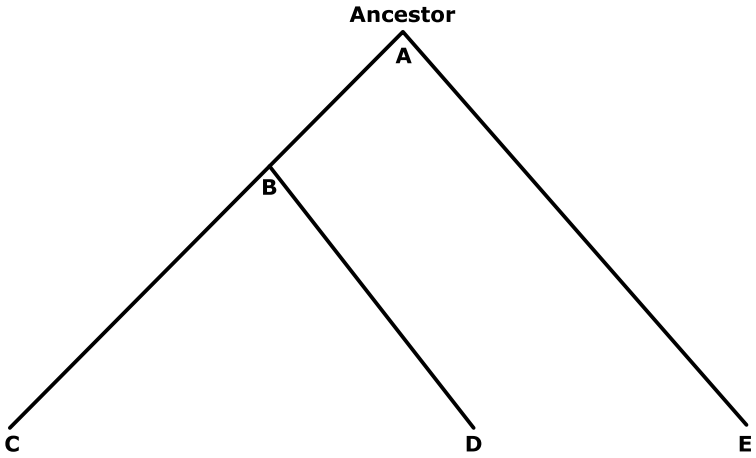

Figure 5.22: Example Phylogenetic Tree

The covariance of α_C and α_D is just the drift parameter for the period of drift that C and D share in common before the bifurcation at B, c_B, multiplied by $\pi_A(1-\pi_A)$. Neither C nor D shared any period of drift in common with E after the ancestral population A, so the covariances of α_C and α_E and of α_D and α_E are both 0. This pattern is followed in a general tree, the covariance of an allele frequency in two present-day subpopulations is the sum of the drift parameters along any periods of drift they have in common multiplied by $\pi_A(1-\pi_A)$, or equivalently, the variance of the allele frequency in their most recent common ancestor, unless that ancestor is the root, in which case it is 0. By building up a covariance matrix, $\mathbf{V}$, based on the phylogenetic relationships between the present-day subpopulations in this way, the allele frequencies in these present-day subpopulations can be modelled as a Multivariate Normal distribution where all the means are the ancestral frequency, π_A, $MVN(\boldsymbol{\pi}_\mathbf{A}, \mathbf{V})$ where $\boldsymbol{\pi}_\mathbf{A} = [\pi_A\, \pi_A \ldots \pi_A]^T$.

Admixtures can be added to this framework in a similar way to the model developed earlier in this chapter but with an important difference. In the Pickrell and

Pritchard model, a population, H, that is an admixture of two populations F and G that are ancestral to it, has an allele frequency $\alpha_H = w\alpha_F + (1-w)(\alpha_G + \epsilon_H)$. Like the models developed earlier, this has an admixture parameter, w, and the allele frequency of the admixed population is a linear combination of the two parent populations' frequencies. There is also an extra term $(1-w)\epsilon_H$. There is a reason for this. Recall that in the models developed earlier, there were three periods of drift allowed around an admixture event. There were two periods of drift before the admixture event, one from each of the two parent populations and a third period of drift after the admixture event. However, this led to non-identifiability. In a Bayesian hierarchical model this can lead to problems with sluggish mixing and high uncertainty about marginal parameter values in the posterior distribution. In a frequentist setting, it is a bigger problem. There is no single point of maximum likelihood. Instead, there is typically a maximum likelihood "ridge" of points that the models cannot distinguish between, preventing it from estimating the parameters. For that reason, Pickrell and Pritchard have to impose additional assumptions. They assume that there is only drift near an admixture event in one of these three directions. They assume no drift after the admixture event and also that there is no drift between the parent population with the lower admixture parameter weight and the admixture event. They do, however allow drift between the heavier-weighted population and the admixture event, leading to the $(1-w)\epsilon_H$ term above. This breaks the symmetry before the admixture event, and restricts w to being less than $\frac{1}{2}$. The edge in the graph just before the admixture, to which w is applied is termed the migratory edge. While this is really just a case of a choice of labelling, this terminology risks being misinterpreted. The word 'migratory' implies that it is that parent population that moved in order to meet the other parent population and create an admixture. Since it has the lower weight, this may well usually be the case but it need not necessarily be so. For example, most of the models featured earlier in this chapter with admixture events for the Mexicans, gave the Europeans contributing to the admixture slightly higher weights than the East Asians, even though the native Americans as descendants of the Asians

were already in Mexico and it was the Europeans who were migrating there. The assumption may also explain why the TreeMix model was only observed to suggest admixture events at the leaves of the tree. The model developed earlier in this chapter is more general and can accommodate admixtures earlier in the tree or, indeed, have two or more two parent population admixtures in series to represent an admixture with three or more parent populations.

The entries for an admixed population in the covariance matrix are built in a similar way to those for the non-admixed present-day subpopulations. Periods of drift that are common to a population and the path to the migratory edge of the admixed population are weighted by w. Periods of drift that are common to a population and the path to the other edge leading to the admixed population are weighted by $1-w$. These may occur for the same population, in which case, the two terms are added together. A population that shares no period of drift in common with either path from the overall ancestor to the admixed population has a covariance with it of 0. The resulting combination of drift parameters are multiplied by $\pi_A(1-\pi_A)$ and enter into the covariance matrix $\mathbf{V}$.

The problem with using the resulting covariance matrix $\mathbf{V}$ is that the values of the proportions of alleles in the ancestral populations are not known. So an expectation based covariance matrix, $\mathbf{W}$ with ($\mathcal{I}$,$\mathcal{J}$)th element

$$W_{\mathcal{I}\mathcal{J}} = E\left[\left(\frac{x_{\mathcal{I}}}{n_{\mathcal{I}}} - \hat{\mu}\right)\left(\frac{x_{\mathcal{J}}}{n_{\mathcal{J}}} - \hat{\mu}\right)\right], \tag{5.35}$$

where

$$\hat{\mu} = \frac{1}{J}\sum_{\mathcal{I}=1}^{J}\frac{x_{\mathcal{I}}}{n_{\mathcal{I}}} \tag{5.36}$$

and J is the number of populations, is considered instead. $x_{\mathcal{I}}$ and $n_{\mathcal{I}}$ are the allele counts and sample sizes for each population as before. This can be shown to be

related to $V_{\mathcal{IJ}}$ by

$$W_{\mathcal{IJ}} = V_{\mathcal{IJ}} - \frac{1}{J}\sum_{\mathcal{A}=1}^{J} V_{\mathcal{AI}} - \frac{1}{J}\sum_{\mathcal{A}=1}^{J} V_{\mathcal{AJ}} + \frac{1}{J^2}\sum_{\mathcal{A}=1}^{J}\sum_{\mathcal{B}=1}^{J} V_{\mathcal{AB}} \tag{5.37}$$

In practice this matrix $\mathbf{W}$ is estimated from the data to produce a sample covariance matrix $\hat{\mathbf{W}}$, using

$$\hat{W}_{\mathcal{IJ}} = \frac{1}{L}\sum_{i=1}^{L}\left[\left(\frac{x_{i\mathcal{I}}}{n_{i\mathcal{I}}} - \hat{\mu}_i\right)\left(\frac{x_{i\mathcal{J}}}{n_{i\mathcal{J}}} - \hat{\mu}_i\right)\right] \tag{5.38}$$

where, as usual, i indexes loci and L is the total number of loci.

$$\hat{\mu}_i = \frac{1}{J}\sum_{\mathcal{I}=1}^{J}\frac{x_{i\mathcal{I}}}{n_{i\mathcal{I}}} \tag{5.39}$$

To deal with linkage disequilibrium, the sample is divided into equal-sized blocks of loci so that there is no linkage disequilibrium between two loci in different blocks. $\hat{W}_{\mathcal{IJ}}$ is then calculated within each block as described above. The mean over all the blocks is used in the overall estimated covariance matrix, $\bar{\hat{W}}_{\mathcal{IJ}}$. So if $\hat{W}_{\mathcal{HIJ}}$ is the entry for subpopulations $\mathcal{I}$ and $\mathcal{J}$ for the $\mathcal{H}$ th block out of $\mathcal{P}$ blocks, $\bar{\hat{W}}_{\mathcal{IJ}} = \frac{1}{\mathcal{P}}\sum_{\mathcal{H}=1}^{\mathcal{P}} \hat{W}_{\mathcal{HIJ}}$. This does allow Pickrell and Pritchard to make use of data on loci that are in linkage disequilibrium. Each block is assumed to be independent of each other. However, it is likely that adjacent blocks will contain loci that are in linkage disequilibrium. This is in contrast to the approach in the models that have been developed earlier in this thesis where only loci that are seperated enough from each other to be reasonably assumed independent are analysed. There is a trade-off between the robustness of the independence assumption and making fuller use of available data.

Taking samples introduces an additional source of variance or noise into the analysis, so each $\bar{\hat{W}}_{\mathcal{IJ}}$ can be thought of as being approximately normally distributed around a true $\hat{W}_{\mathcal{IJ}[true]}$ with a variance $\sigma^2_{\mathcal{IJ}}$ to express the variability across blocks.

This variability can be estimated from the data from

$$\hat{\sigma}_{\mathcal{IJ}} = \sqrt{\frac{\sum_{\mathcal{H}=1}^{\mathcal{P}} \left(\hat{W}_{\mathcal{HIJ}} - \bar{\hat{W}}_{\mathcal{IJ}} \right)^2}{\mathcal{P}\left(\mathcal{P}-1\right)}} \tag{5.40}$$

The $\mathcal{P}-1$ in the denominator comes from the definition of variance and the $\mathcal{P}$ comes from $\hat{\sigma}_{\mathcal{IJ}}$ being the error of a mean or standard error. The point of all this is to obtain a likelihood for the data for a given graph. Each graph, G, will have a particular covariance matrix, $\mathbf{V}$ associated with it and corresponding $\mathbf{W}$. The composite likelihood for $\bar{\hat{\mathbf{W}}}$ is the product of the probability density for each pair of subpopulations, $\mathcal{I}$ and $\mathcal{J}$.

$$L\left(\bar{\hat{\mathbf{W}}}|\mathbf{W}\right) = \prod_{\mathcal{I}=1}^{J}\prod_{\mathcal{J}=\mathcal{I}}^{J} N\left(\bar{\hat{W}}_{\mathcal{IJ}}|W_{\mathcal{IJ}}(G,c), \hat{\sigma}^2_{\mathcal{IJ}}\right) \tag{5.41}$$

For diagnostic purposes for a given graph G, a matrix of residuals, $\mathbf{R}$ can be calculated from $\mathbf{R} = \bar{\hat{\mathbf{W}}} - \mathbf{W}\left(c\right)$. These residuals can be used to calculate the proportion of the variance in $\bar{\hat{\mathbf{W}}}$, which has been calculated from the data, that is explained by $\mathbf{W}$, which depends on the choice of graph. This approximate proportion of the relatedness that is reflected in the model, $\mathcal{F}$, is defined by

$$\mathcal{F} = 1 - \frac{\sum_{\mathcal{I}=1}^{J}\sum_{\mathcal{J}=\mathcal{I}+1}^{J}\left(R_{\mathcal{IJ}} - \bar{R}\right)^2}{\sum_{\mathcal{I}=1}^{J}\sum_{\mathcal{J}=\mathcal{I}+1}^{J}\left(\bar{\hat{W}}_{\mathcal{IJ}} - \overline{\bar{\hat{W}}}\right)^2}, \tag{5.42}$$

where

$$\bar{R} = 1 - \frac{2}{J(J-1)}\sum_{\mathcal{I}=1}^{J}\sum_{\mathcal{J}=\mathcal{I}+1}^{J} R_{\mathcal{IJ}}, \tag{5.43}$$

and

$$\overline{\bar{\hat{W}}} = 1 - \frac{2}{J(J-1)}\sum_{\mathcal{I}=1}^{J}\sum_{\mathcal{J}=\mathcal{I}+1}^{J} \bar{\hat{W}}_{\mathcal{IJ}}. \tag{5.44}$$

But how does TreeMix go about choosing which graph G, to analyse? For any

unrooted bifurcating tree graph with J present-day subpopulations, there are $(2J-5)!!$ possible graphs (Penny et al., 2007). For $J = 5$, say, that is only 15 possible trees and it is feasible to test all the possible graphs and find the one with the highest (composite) likelihood by exhaustion or brute force. Such an approach would even be feasible for the Bayesian hierarchical models described above. However, the number of graphs very quickly becomes huge with increasing J. For the 11 subpopulations of the HapMap dataset, the number of possible unrooted bifurcating graphs is 34,459,425. If it took only 1 second to compute the likelihood for each graph, it would still take nearly 400 days to find the optimal one by an exhaustive search. Clearly for larger numbers of subpopulations, it is not feasible even in the framework of Pritchard and Pickrell's relatively fast frequentist model, to consider every possible graph. Instead a greedy algorithm can be used such as that of Felsenstein (1981).

To within graph isomorphism, there is only one possible unrooted tree for 3 subpopulations, A, B and C (figure 5.23). To understand what graph isomorphism means, imagine the tree is made of rubber in 3D. It can be stretched, bent, flipped over, rotated and its edges can even be twisted without breaking it, but it must not have any parts cut and/or re-attached to another part of the tree. If one unrooted tree can be made to look exactly like another by any combination of these permitted operations then the two trees are said to be graph isomorphic. Effectively, they are just two different ways of drawing the same unrooted tree.

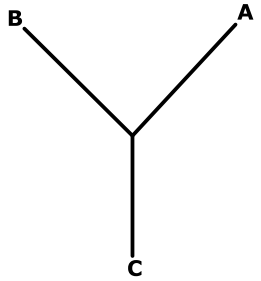

Figure 5.23: The only unrooted tree for 3 subpopulations.

Figure 5.24 shows the three such possible unrooted trees for 4 subpopulations

labelled A, B, C and D. These are effectively the same trees as would have been produced by adding the edge leading to D to each of the three edges in the unrooted tree in figure 5.23. In general an unrooted bifurcating tree with J subpopulations has $2J - 3$ edges. So, for each of these three graphs there are $2 \times 4 - 3 = 5$ edges to which a fifth subpopulation, E could be attached leading to $3 \times 5 = 15$ possible trees for 5 subpopulations. By building trees up in this way, it can be readily seen where the expression $(2J - 5)!!$ for the possible number of trees with J subpopulations comes from.

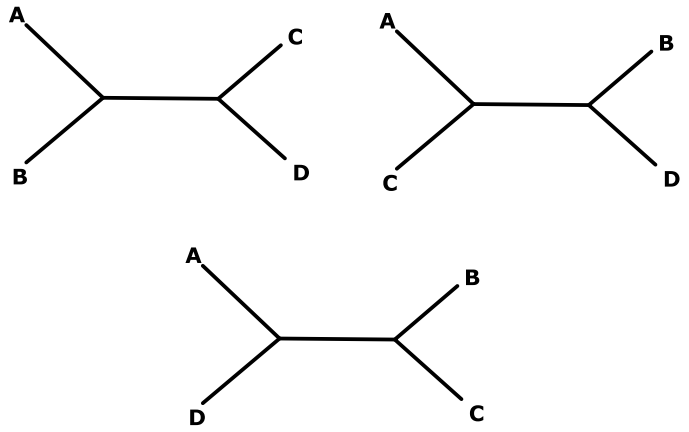

Figure 5.24: The three unrooted trees for 4 subpopulations.
All three unrooted trees of three subpopulations A, B,C and D. All other ways of drawing the unrooted trees can be made to look identical to one of these trees by stretching, squashing, flipping, twisting, rotating or bending them.

So how does this help to solve the problem of reducing the number of possible unrooted trees that need to have their likelihood evaluated? The procedure starts by taking three of the subpopulations and arranging them in their only possible tree. The choice of which three is arbitrary and can just be the order in which the subpopulations have been indexed. A fourth subpopulation is chosen. This can be added, as has been seen, in three different ways. Three is not a large number so the likelihood of each of these three resulting trees can be calculated. The one with the largest likelihood is accepted and moves forward to the next step. A fifth subpopulation is chosen, which can be added to the tree in five different ways. Five is still not a large number, so it is reasonable to calculate the likelihood for all of these. Again, the one with the largest likelihood is accepted and moves forward

to the next step. It would be possible to keep adding subpopulations in this way until they have all been added. However, this would have the drawback that the resulting unrooted tree could be dependent on which three subpopulations had been chosen to start the process and the order in which the subpopulations were added. Trying all the possible orders of subpopulations would just lead back to having to test a very large number of trees, defeating the point of trying to find an algorithm to reduce the number of trees that have to be tested. Early versions of this algorithm advocated trying a small number of possible orders to see how robust the resulting tree was to the choice of order. Later, an additional step was added between adding subpopulations that evaluated "local rearrangements" of the tree, so that before adding a sixth (or subsequent) subpopulation, the likelihood of a number of these local rearrangements of the tree would be evaluated before the additional subpopulation is added.

So what are these local rearrangements? One local rearrangement method, called Nearest Neighbour Interchange, involves looking at rearrangements of the tree around internal edges. Every such tree with J subpopulations will have $J-3$ internal edges. Internal edges are edges with no present-day subpopulation labels at either end, or equivalently in the case of these unrooted bifurcating trees, an edge that is connected to exactly four other edges. So, in figure 5.23, the tree has no internal edges and in figure 5.24, each tree has one internal edge. At each internal edge of a tree, the four edges connecting to it can be disconnected and reconnected to it in exactly three different ways up to graph isomorphism. There are a total of three ways for exactly the same reason that there are only three different trees with four subpopulations. So, at any internal edge, the four edges connecting to it can be reconnected in three different ways, the original way and two others. The likelihood for the graphs resulting from these two other ways of connecting to the internal edge can be evaluated and compared to the likelihood for the original tree. The tree with the largest likelihood is selected and the next internal edge is examined in the same way. Since there are only $J-3$ internal

edges and 2 new trees to evaluate at each edge then, on each cycle through the internal edges, only $2(J-3)$ tree likelihoods are evaluated. For 11 subpopulations that is only 16 trees so the numbers are very manageable. This is repeated until such a cycle through all the internal edges reveals no trees were more likely than the tree that had been the selected one at the beginning of the cycle. It is at that point that the next subpopulation is added to the tree. The process of adding subpopulations and doing local rearrangements continues until there are no more subpopulations to add.

But to construct the matrix $\mathbf{V}$ requires a graph that is rooted and these are unrooted trees. The user then must choose where the root should go. The user names a subpopulation and the root always goes on the edge nearest that subpopulation. This subpopulation must also be one of the first three subpopulations that start the process with a three subpopulation tree. This does restrict the number of possible rooted trees. The way Pritchard and Pickrell advocate getting round this is to have an outgroup among the subpopulations that is not as related to the other subpopulations as they are to each other. This makes it obvious that the root belongs on the edge leading to the outgroup. This has a downside however. One of the assumptions of this model is that there is not much drift along any edge. Advocating the use of a less related outgroup seems inconsistant with that assumption but manifestly some way of locating the root is needed.

So how are the admixtures chosen and the migration edges added? That part of the process uses the residual matrix, $\mathbf{R}$, although the specifics are somewhat sketchy. The user defines how many migrations there should be. Suppose they specify that there should be $\mathcal{M}$ migrations. If $\mathcal{M} = 0$ no migration edges are added and the process ends. Otherwise, the $\mathcal{M}$ pairs of populations with the highest entries in $\mathbf{R}$ are found. Migration edges between edges and nodes near or at these population pairs are tried and the one that most increases the likelihood is chosen. There is another round of the "local rearrangements" part of the algorithm described above before repeating the migration edge selection procedure for the

second edge, unless of course, $\mathcal{M} = 1$ when the process finishes. The process of adding migration edges and performing local rearrangements continues until $\mathcal{M}$ edges have been added and the last round of local rearrangements have taken place.

This process is a greedy algorithm which arrives at some locally optimal graph. It is not guaranteed to find the graph with the globally maximal component likelihood in the way an exhaustive approach would. It does however, cut down greatly on the number of trees whose likelihood needs to be evaluated and renders the whole process practical enough to take place in minutes even for large numbers of subpopulations.

5.6.2 Comparison of Output for the Two Models

Data were simulated for a simple tree of four ficticious Celtic tribes, Aon, Dhà, Trì and Ceithir (figure 5.25).

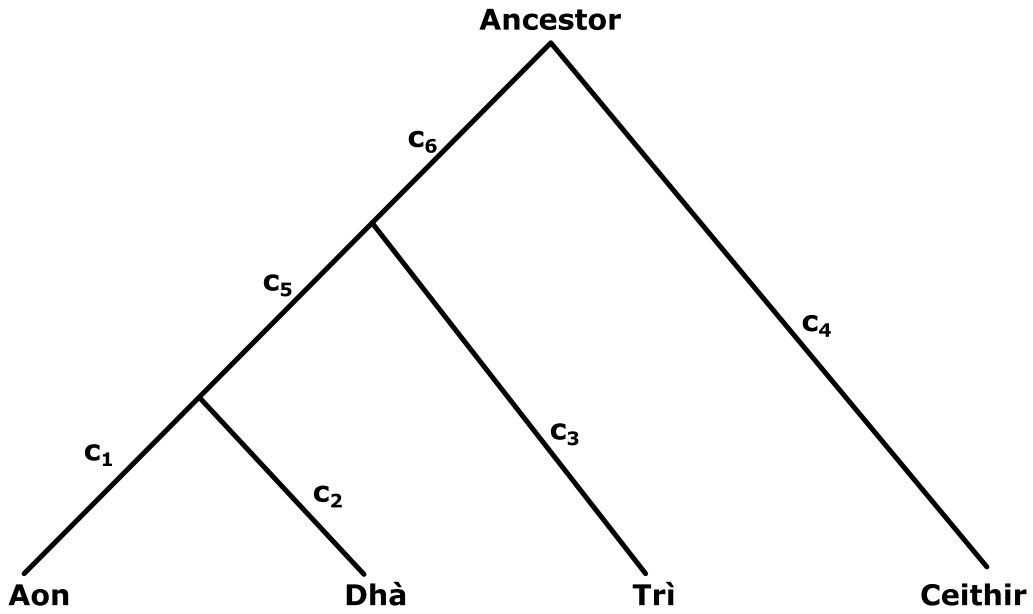

Figure 5.25: Phylogenetic Tree for Simulated Data for Four Subpopulations

The drift parameters along each edge were set at 0.05. The ancestral distribution of the allele frequencies was Uniform(0,1). There were no admixture events

simulated. Since all the 1000 loci (similar to the number in the dataset for a medium HAPMAP chromome) were simulated to be independent, there was no need for blocking. The root was set to its true position in running TreeMix. Figure 5.26 shows the TreeMix output graph for this data. TreeMix retrieves the correct structure. However, the drift parameters on the scale below the graph are more of a problem. If they were drift parameters (cs) and retrieved correctly, the nodes would be about 0.05 apart. They are nearer 0.006 to 0.009 apart. The label on the scale is misleading. It is a scale for $c\pi_A(1-\pi_A)$ rather than for just c. The scale measures variances rather than the drift parameters themselves. The models developed in this thesis, in contrast, do not seek the correct structure themselves but do give posterior distributions for the drift parameters, c_j, from which point estimates and measures of uncertainty about the parameters themselves can be derived.

To show how the interpretation of the TreeMix drift variances could be difficult, data with the same drift values but from a different distribution of ancestral allele frequencies were produced and analysed using TreeMix. The case where $\pi \sim$Beta(0.5,0.5), a u-shaped distribution, was used is shown in figure 5.27. Although, the graph has been drawn differently, it is still graph isomorphic to the correct structure. However, it can be seen that the positions of the nodes along the "Drift parameter" axis are shifted to the left. This appears to suggest that the estimates of drift are less but it is only the estimates of $c_j\pi_A(1-\pi_A)$ that have been reduced.

The same was done with $\pi \sim$Beta(10,10) a very n-shaped distribution of ancestral allele frequencies (5.28). Once the change in scale of the "Drift parameter" axis has been taken into account, the graph can be seen to have been stretched to the right. The value of $\mathrm{E}[\pi_A(1-\pi_A)]$ is $\frac{1}{8}$ for Beta$(0.5,0.5)$, $\frac{1}{6}$ for Beta$(1,1)$ and $\frac{5}{21}$ for Beta$(10,10)$. Comparison of the scales of figures 5.26-5.28 shows them to differ in scale in proportion to these values. To the user who may be unaware of the detailed internal workings of TreeMix, the drift parameters appear to be different

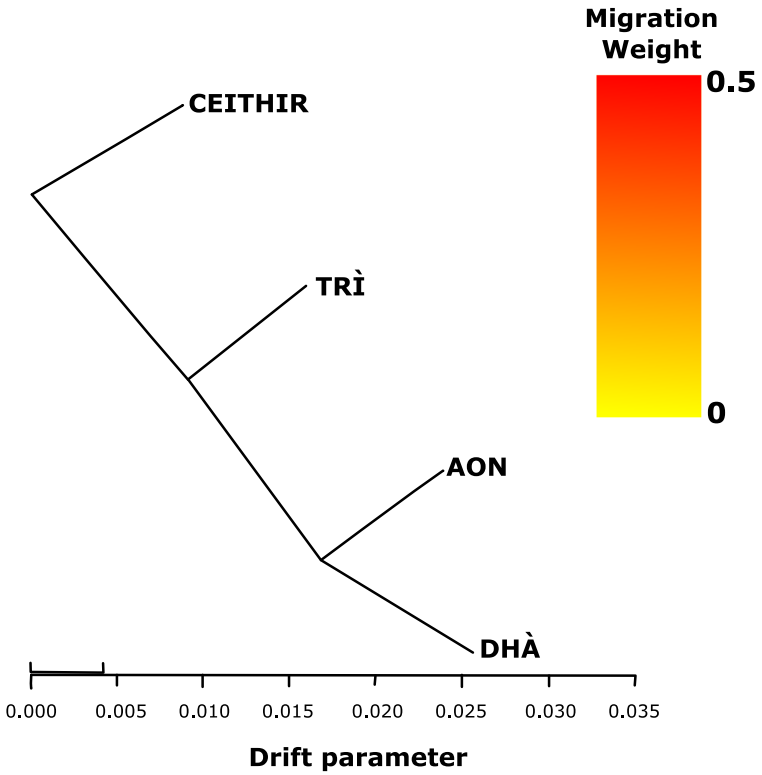

Figure 5.26: TreeMix Output from Analysis of Data Simulated with True Ancestral Allele Frequencies Drawn From Beta(1,1)

Output from TreeMix from analysing data simulated from the model in figure 5.25. The true drift parameters $c_1 ... c_6$ were all set to 0.05 and the ancestral allele frequencies π_A drawn from Beta(1,1). 1000 independent loci were simulated. No admixture events are inferred so the migration weight scale is irrelevant.

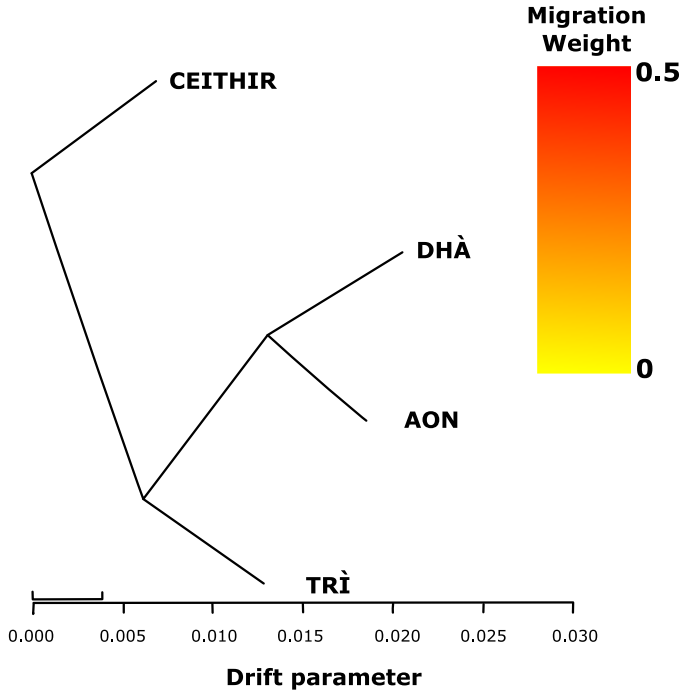

Figure 5.27: TreeMix Output from Analysis of Data Simulated with True Ancestral Allele Frequencies Drawn From Beta(0.5,0.5)

Output from TreeMix from analysing data simulated from the model in figure 5.25. The true drift parameters $c_1 ... c_6$ were all set to 0.05 and the ancestral allele frequencies π_A drawn from Beta(0.5,0.5). 1000 independent loci were simulated.

in each case but in fact are all the same. Even if the fact that these are really only variances is known, it is still hard to discern what the real drift parameters are. It does, however, show the relative size of the drift parameters because they are all multiplied by the same factor.

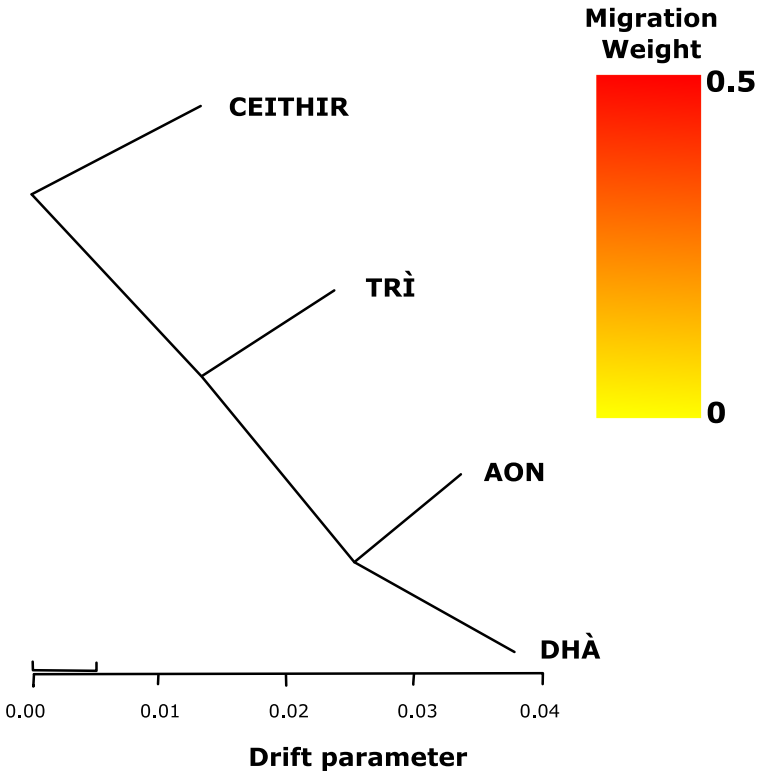

Figure 5.28: TreeMix Output from Analysis of Data Simulated with True Ancestral Allele Frequencies Drawn From Beta(10,10)
Output from TreeMix from analysing data simulated from the model in figure 5.25. The true drift parameters $c_1 ... c_6$ were all set to 0.05 and the ancestral allele frequencies π_A drawn from Beta(10,10). 1000 independent loci were simulated.

This dependence of the TreeMix "Drift parameter" on the distribution of π_A has other unfortunate consequences. It can become sensitive to irrelevant data. 1000 additional loci where the counts for all four populations were all 0 were added to the dataset analysed in figure 5.26. By far the most likely reason for an observation of 0 counts in all four populations is that the ancestral frequency is 0. If that was the case then these loci will contribute no information about the drift that has taken place because any level of drift would have the same outcome. The 1000 additional loci with 0 counts contain no (or at least little) information about drift

and so are uninformative which should not impact on any output. However, in the case of TreeMix it causes a problem. TreeMix does not take fixation into account. It will instead take the same data as being evidence of little or no drift (5.29). As might be expected, while the structure has still been retrieved, the "Drift parameter" estimates have reduced to about half their earlier values. Thus the estimates have been affected by the additional irrelevant data.

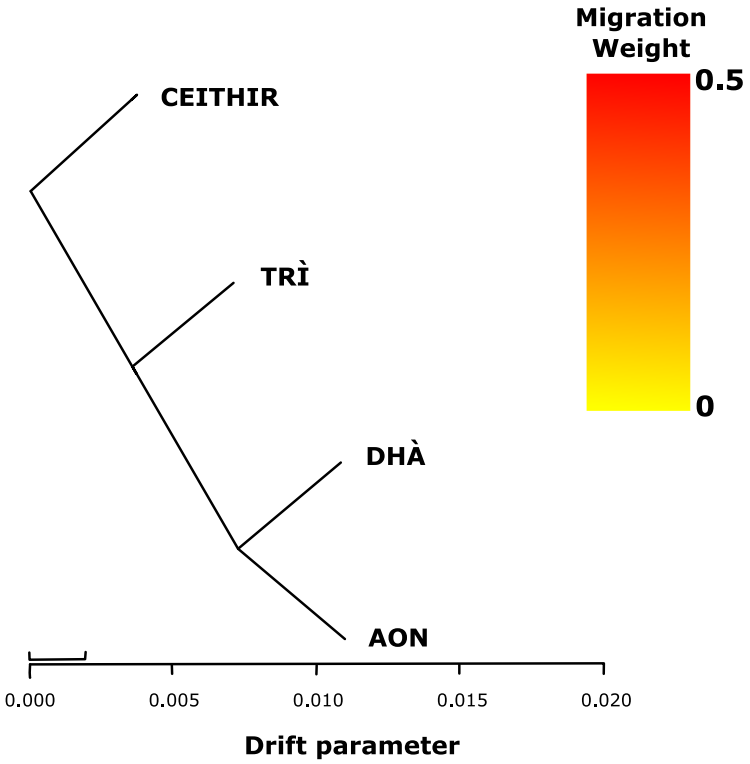

Figure 5.29: TreeMix Output from Analysis of Data Simulated with Half True Ancestral Allele Frequencies Drawn From Beta(1,1) and Half Set at 0.
Output from TreeMix from analysing data simulated from the model in figure 5.25. The true drift parameters $c_1 ... c_6$ were all set to 0.05 and the ancestral allele frequencies π_A drawn from Beta(1,1). 1000 independent loci were simulated. 1000 more loci were added with ancestral allele frequencies of 0.

This might be thought not to be a problem. In a real dataset, such loci with all 0 counts could be screened out. The problem is that in a setting with a more complicated structure and a larger number of subpopulations, the situation of the four Celtic tribes in these simulations could be a subtree of a much larger phylogenetic tree, analogous to the four African subpopulations in figure 4.7. There could be non-zero counts in other subpopulations and zeros for these four and these

zeros would have the effect of reducing the estimate of drift in their areas. It could be possible to weed out all the loci in a dataset containing any counts where there is any subpopulation where the count was 0 or the same as the sample size but that would mean discarding information that would convey useful information about drift elsewhere in the tree in order to satisfy the assumptions of the model and still not have drift parameter estimates that are easily interpreted anyway.

5.6.2.1 Use of An Outgroup to Strengthen Identifiability Near the Root

To show that the model developed in Chapter 4, deals more appropriately with irrelevant information, table 5.7 displays the 95% HPD range of the estimates of the drift parameters for the simulated dataset without the 1,000 loci with zero counts added. Subpopulation sizes were all 200 (100 individuals) similar to those in the HAPMAP dataset. Table 5.8 shows the same information with these 1,000 zero count loci added. When the 1,000 extra loci are added, the drift parameters for periods of drift that are not adjacent to the ancestral population, are almost unchanged. They are only changed for the two periods of drift c_4 and c_6 that are either side of the ancestral population. This has happened because the prior on π is now misspecified. This can, nonetheless, be easily overcome by use of an outgroup. The outgroup does not have to contain additional data. The allele counts of the outgroup used in this case, (figure 5.30) were created from taking an unweighted mean of the counts from the four subpopulation counts that had already been simulated. The estimates for c_4 and c_6 (table 5.9) are now much closer to those for the original dataset (table 5.7). The drifts c_7 and c_8 are artificial and can be ignored. When using an outgroup with the model in this way, it may be better to use a more bell-shaped distribution for the prior on π. This is because periods of drift make the distribution of the αs more u-shaped. The extra period of drift at c_7 should be taken into account. If there is a particular distribution of α expected after that drift, (in this case it was known to be Uniform(0,1) because

the data are simulated), a more bell-shaped one is really needed as the prior before c_7. The model's choice for the amount of drift will then help adjust to the more sypathetic distribution. For this reason it may be best to overestimate the bell-shapedness of the prior. For example, table 5.10 shows the results of using a rather extreme Beta(10,10) prior. This produces results, after drifts c_7 and c_8 are discarded, even closer to the original output (table 5.7). The true values of drift, 0.05, are now within the 95% intervals for all of c_1 to c_6. This approach shows how the whole problem described in section 4.5.6 of the results for drifts near the ancestral population being very sensitive to choice of the prior on the ancestral allele frequency can be overcome by use of an outgroup and a larger a on the prior for π, making it more bell-shaped. In this way, the model copes well with the irrelevant information when estimating drift parameters and with minor modification can cope even better, whereas TreeMix results are adversely affected.

Table 5.7: Parameter Estimates of the Model in Figure 5.25

	95% HPD Interval Bounds		
variable	lower	upper	median
c_1	0.0425	0.0589	0.0504
c_2	0.0395	0.0556	0.0473
c_3	0.0389	0.0556	0.0471
c_4	0.0444	0.0668	0.0554
c_5	0.0354	0.0538	0.0442
c_6	0.0319	0.0525	0.0417

Table of parameter estimates obtained for data simulated according to the model in figure 5.25. The true drift parameters $c_1 ... c_6$ were all set to 0.05 and the ancestral allele frequencies π_A drawn from Beta(1,1). 1000 independent loci were simulated. The table shows the resulting 95% HPD intervals from using the model on such data for the drift parameters, c.

Table 5.8: Parameter Estimates of the Model in Figure 5.25 with 1,000 Additional Uninformative Loci Added

	95% HPD Interval Bounds		
variable	lower	upper	median
c_1	0.0430	0.0598	0.0511
c_2	0.0381	0.0543	0.0461
c_3	0.0380	0.0598	0.0484
c_4	0.9248	1.0000	0.9793
c_5	0.0336	0.0557	0.0445
c_6	0.9025	1.0000	0.9697

Table of parameter estimates obtained for data simulated according to the model in figure 5.25. The true drift parameters $c_1 ... c_6$ were all set to 0.05 and the ancestral allele frequencies π_A drawn from Beta(1,1). 1,000 independent loci were simulated. Then 1,000 additional loci were simulated with π_A set at 0. The table shows the resulting 95% HPD intervals from using the model on such data for the drift parameters, c.

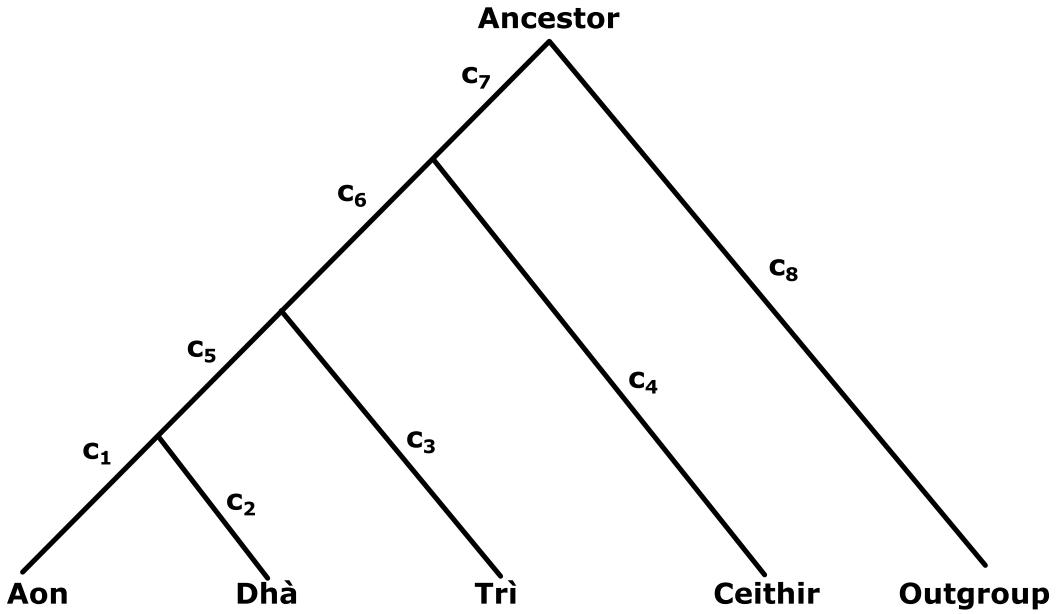

Figure 5.30: Phylogenetic Tree for Simulated Data for Four Subpopulations and an Outgroup

The outgroup is fabricated by taking the unweighted means of the counts in the four simulated subpopulations and rounding to the nearest integer.

Table 5.9: Parameter Estimates of the Model in Figure 5.30 with 1,000 Additional Uninformative Loci Added with a Beta(1,1) Prior on π.

	95% HPD Interval Bounds		
variable	lower	upper	median
c_1	0.0423	0.0586	0.0503
c_2	0.0397	0.0555	0.0472
c_3	0.0393	0.0562	0.0475
c_4	0.0481	0.0717	0.0596
c_5	0.0344	0.0529	0.0434
c_6	0.0272	0.0471	0.0370
c_7	0.7296	0.9752	0.8468
c_8	0.7371	0.9800	0.8535

Table of parameter estimates obtained for data simulated according to the model in figure 5.30. The true drift parameters c_1...c_7 were all set to 0.05 and the ancestral allele frequencies π_A drawn from Beta(1,1). 1,000 independent loci were simulated. Then 1,000 additional loci were simulated with π_A set at 0. The table shows the resulting 95% HPD intervals from using the model on such data for the drift parameters, c. The Prior on π for the analysis was Beta(1,1). The outgroup was fabricated by taking the unweighted means of the counts in the four simulated subpopulations and rounding to the nearest integer.

Table 5.10: Parameter Estimates Table of the Model in Figure 5.30 with 1,000 Additional Uninformative Loci Added after 100,000 Iterations With a Beta(10,10) Prior on π.

	95% HPD Interval Bounds		
variable	lower	upper	median
c_1	0.0424	0.0587	0.0502
c_2	0.0394	0.0551	0.0471
c_3	0.0388	0.0555	0.0469
c_4	0.0462	0.0679	0.0566
c_5	0.0352	0.0538	0.0440
c_6	0.0333	0.0534	0.0430
c_7	0.9908	1.0000	0.9978
c_8	0.9913	1.0000	0.9980

Table of parameter estimates obtained for data simulated according to the model in figure 5.30. The true drift parameters c_1...c_7 were all set to 0.05 and the ancestral allele frequencies π_A drawn from Beta(1,1). 1,000 independent loci were simulated. Then 1,000 additional loci were simulated with π_A set at 0. The table shows the resulting 95% HPD intervals from using the model on such data for the drift parameters, c. The Prior on π for the analysis was Beta(10,10). The outgroup was fabricated by taking the unweighted means of the counts in the four simulated subpopulations and rounding to the nearest integer.

5.6.3 Choices of Phylogenetic Tree in TreeMix

In reality, it will be unusual for a dataset to meet all the assumptions of the TreeMix model. To clean a dataset of data that compromise the assumptions would involve discarding much potentially useful information. Nevertheless, the model is still useful if, despite its assumptions, it is a sufficiently good approximation to reality to answer the questions of interest. All statistical models are, after all, to a greater or lesser extent, approximations to reality. In testing both with simulated and the HapMap data, TreeMix has suggested plausible phylogenetic trees in most cases.

So what did TreeMix do with the HapMap dataset for Chromosome 2? If no root is specified, TreeMix attempts to root the tree near the Gujarati (GIH) as shown in figure 5.31. However, the tree is reasonably plausible after a relocation of the root to the point marked by the red dot. Then it becomes the tree of figure 5.13. This in turn is the tree of figure 4.7 discussed in the last chapter with the Gujarati and Mexicans swapped. Since these were the two subpopulations whose positions were the least certain, this tree is not implausible.

The nearest (in terms of distance along the tree) subpopulation to the red dot root position is the Maasai (MKK) so the nearest position of the root that can be specified to TreeMix is to place it there. Figure 5.32 shows what happens if that is done. Other than the position of the root, the tree topology is unchanged. The model performs surprisingly well at the task of choosing a plausible structure considering how many of its assumptions are being blatently violated by this dataset. But then again, the same can also be said for the much simpler Neighbour Joining algorithm. The "drift parameters" suggested by TreeMix still however are hard to interpret since they conflate the demographically interpretable cs with ancestral allele frequencies (which of course vary over loci).

The problem thus far, is that the tree root cannot be specified to be along any edge - it has to be a terminal edge, which is not appropriate here. One of the things

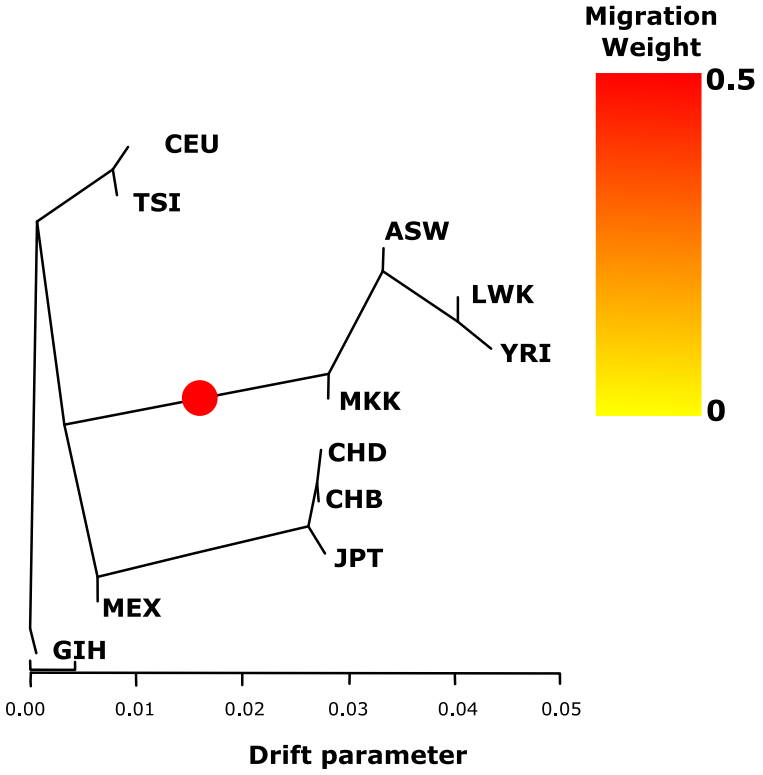

Figure 5.31: TreeMix Output for Chromosome 2 HapMap Data
TreeMix output from applying it to the Chromosome 2 HapMap dataset. TreeMix chose a root at the Gujarati (GIH). The Red dot marks an edge which is most consistant with where the root was placed in previous models. The number of admixtures was set to 0.

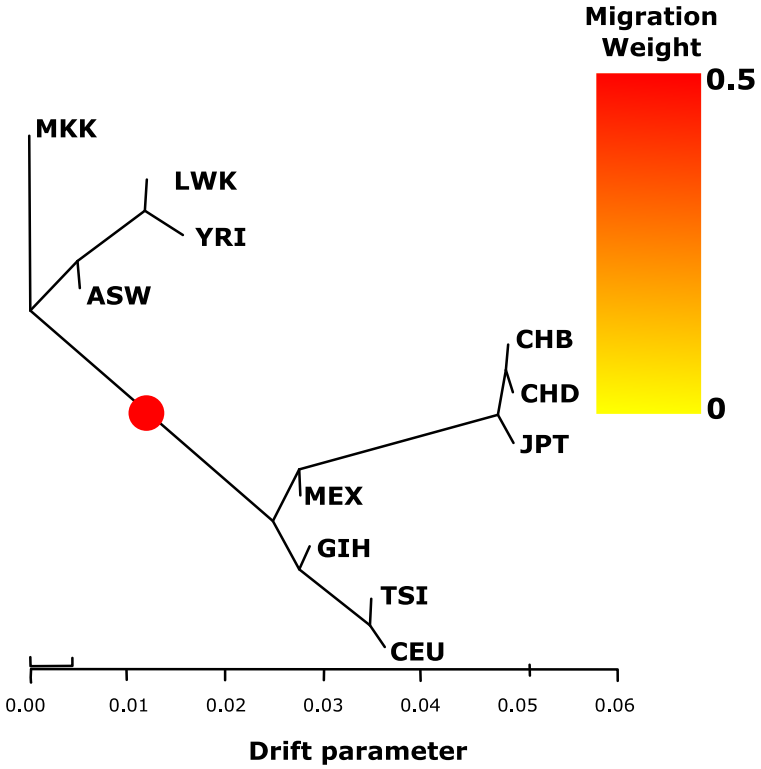

Figure 5.32: TreeMix Output for Chromosome 2 HapMap Data with Root Set on the MKK Edge

TreeMix output from applying it to the Chromosome 2 HapMap dataset. A root on the MKK edge was specified. The Red dot marks an edge which is most consistant with where the root was placed in previous models. The number of admixtures was set to 0.

that Pritchard and Pickrell suggest is to use an outgroup. Given the diversity of the subpopulations in the HapMap dataset ideally the outgroup would be some other species of human such as Neanderthal or Denisovan or perhaps a chimpanzee or one of the other Hominidae. However, no comparable data for these was readily available. An alternative approach was used. An easily calculated rough approximation to ficticious allele counts for the common ancestral population, could be calculated at each locus as an unweighted mean of the count data for all eleven subpopulations. These means, rounded to the nearest integer, could be included as a twelfth fictional subpopulation and since it should be vaguely similar to the common ancestral population, the root could be placed there. The results of adding this twelfth supopulation and analysing the resulting dataset using TreeMix were interesting but not in the way that was anticipated (figure 5.33). The addition of this fictional subpopulation has changed the suggested structure in a surprising and implausible way.

It understandably estimates that the outgroup has suffered little or no drift since the ancestral population, but now has all subpopulations other than the Yoruba (YRI), Afro-Americans (ASW) and Lhosa (LWK) as diverging at about the same time from this ancestral population. As has been noted before, the idea that the Han Chinese in Beijing and Han Chinese in Denver diverged from each other at much the same time as they diverged from the Maasai (MKK) is simply silly. Even if the idea of using the unweighted mean of the allele counts of the other subpopulations for an outgroup is itself a flawed idea, the addition of it as an extra subpopulation, should not radically alter the structure of the rest of the tree. It is difficult to see why adding one new subpopulation to the dataset should be disturbing the others' place in the phylogenetic tree to this extent.

As a point of interest, the same data were fed into the Neighbour Joining algorithm. The unrooted tree that results from doing that is shown in figure 5.34. The outgroup appears along an edge that is very reasonably the root of the tree, separating the African subpopulations from the non-African ones. This can be

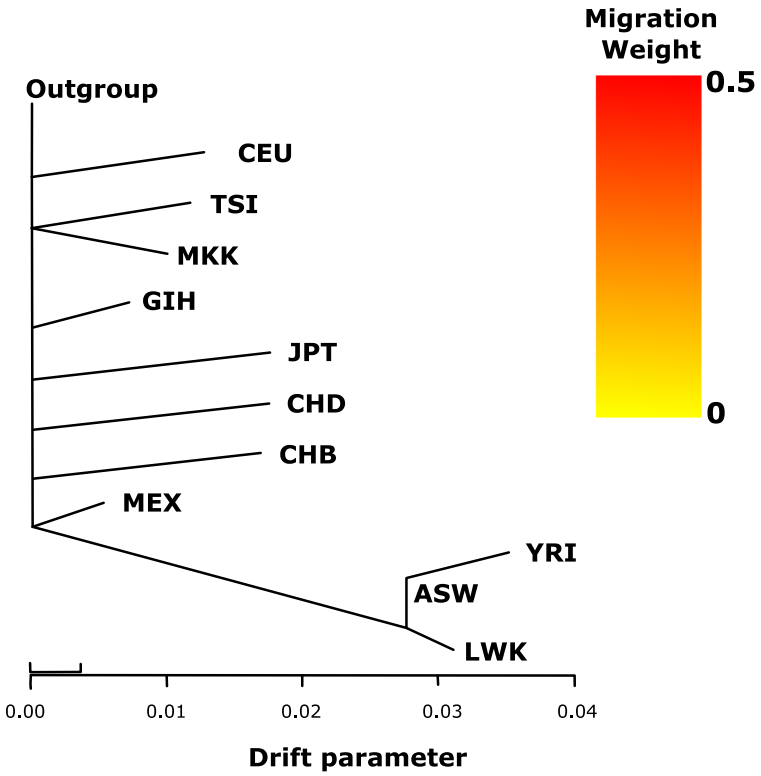

Figure 5.33: TreeMix Output for Chromosome 2 HapMap Data with Root Set Near An Artificial Outgroup

TreeMix output from applying it to the Chromosome 2 HapMap dataset. An artificial outgroup was added with its allele frequencies at each locus set to be the unweighted average of those of the 11 real subpopulations.

compared with figure 4.2. The only change is that the Gujarati (GIH) have moved from the Asian branch (the branch leading to CHD and CHB) to being along the European branch leading to CEU and TSI. This is still a plausible position for them and the Mexicans. It does not represent a huge disturbance to the entire tree. In this particular case, Neighbour Joining seems to have behaved in a much more consistent way than TreeMix when the additional fictional outgroup population was added to the dataset.

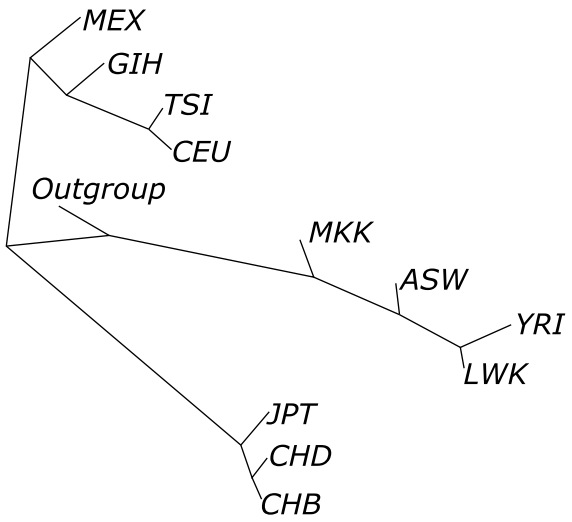

Figure 5.34: Neighbour Joining Tree for Chromosome 2 HapMap Data with An Artificial Outgroup
An artificial outgroup was added with its allele frequencies at each locus set to be the unweighted average of those of the 11 real subpopulations.

The point to be taken from this is that TreeMix usually suggests sensible structures, albeit after some manual adjustment to the position of the root. However, it does not always do so. The same can be said for the much simpler Neighbour Joining algorithm. To be fair, the HapMap dataset does violate many of the model's basic assumptions: it does not exclude loci that are likely close to fixation, and it does include subpopulations that are not so closely related and so potentially involves large periods of drift. Nonetheless, it would not be useful if datasets had to be put through contortions to fit the model. However, it does run quickly (as does the Neighbour Joining algorithm). The output of both these models can

be used as suggestions for further investigation using for example the approaches developed in Chapters 4 and 5.

5.6.4 Choices of Admixture Events in TreeMix

One thing that TreeMix can do that Neighbour Joining cannot is suggest admixture events, or as TreeMix calls them, migrations. To look at how well TreeMix handles admixture in a simple setting, data were simulated according to the model structure shown in figure 5.35. Here the tribe Dhà is an admixture of the two tribes Aon and Trì. Dhà will take $100w\%$ of its ancestry from Aon and the rest from Trì. Each of the 1000 simulated loci (similar to that found in the dataset for a medium HAPMAP chromosome) had its π drawn from Uniform(0,1). All drift parameters $c_1, ..., c_9$ are 0.05 unless stated otherwise.

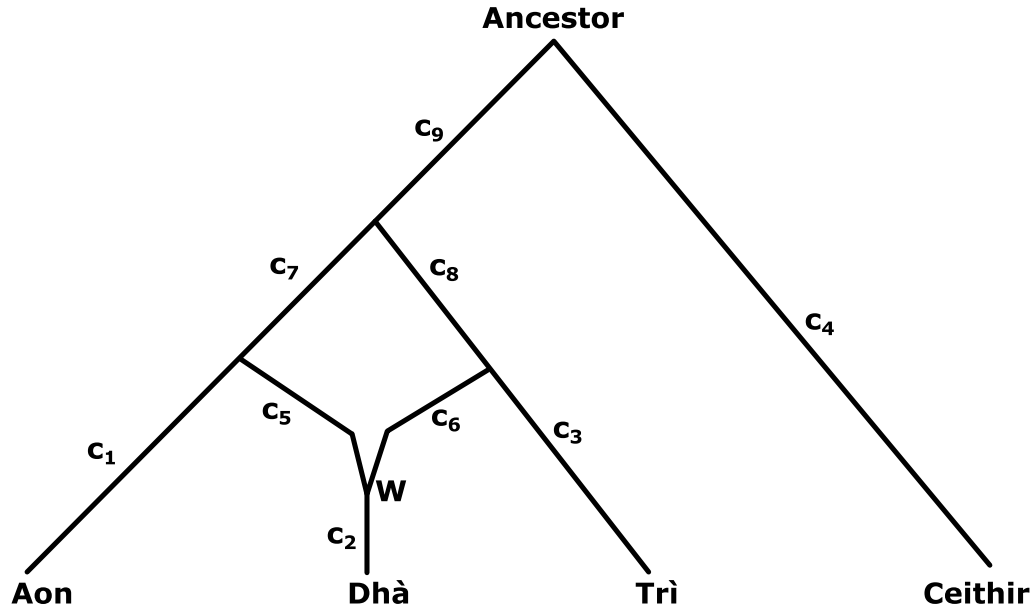

Figure 5.35: Phylogenetic Network for Simulated Data with Admixture for Dhà

A 50% admixture was used ($w = 0.5$) initially. To fit in with Pritchard and Pickrell's assumptions on the drift parameters near an admixture event as closely as possible, c_2 and c_5 were made very small (0.0001). The resulting simulated data

were analysed by TreeMix and it was asked to add one migration. The resulting structure is shown in figure 5.36. It produces the correct network but suggests a w of 0.397084 which is some way away from the true weight of 0.5.

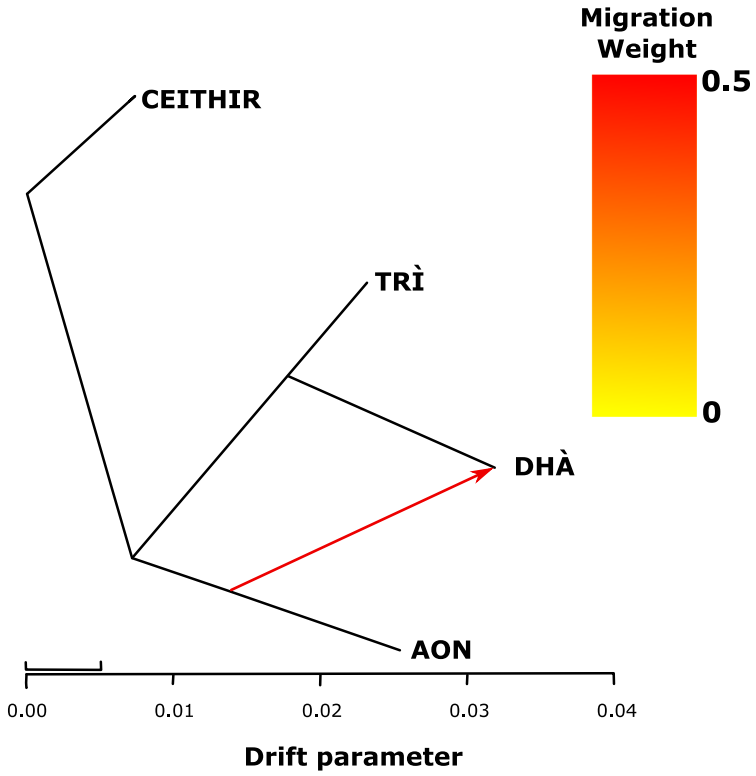

Figure 5.36: TreeMix Output For a 50% Admixture
TreeMix output for data simulated according to the model in figure 5.35. A 50% admixture was used ($w = 0.5$). To fit in with Pritchard and Pickrell's assumptions as closely as possible on the drift parameters near an admixture event, c_2 and c_5 were made very small (0.0001). The other true drift parameters were 0.05.

The same experiment was repeated with the only differences being that the data were simulated with $w = 0.75$, $c_5 = 0.05$ and $c_6 = 0.0001$. The results were much as expected as shown in figure 5.37. Again, the correct admixture is shown. The "migration" has flipped over to coming from Trì. However, it gives an admixture parameter of 0.397; the same as before. In this case, it translates to a w of $1 - 0.397 = 0.603$ which is still a long way from the true value of 0.75. TreeMix seems to pick only particular values for the admixture parameter. If a $w = 0.85$ admixture is used the result is the same.

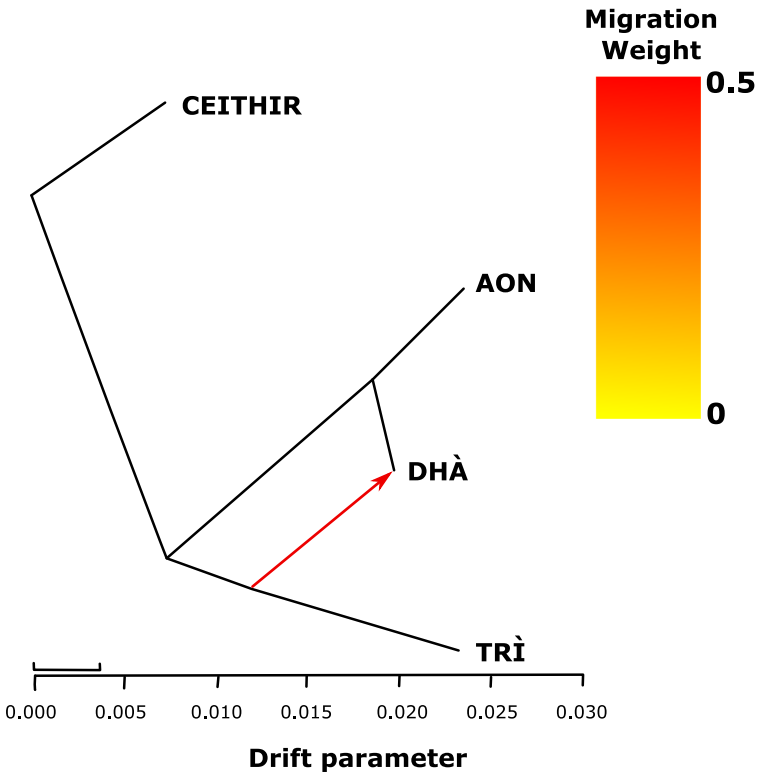

Figure 5.37: TreeMix Output For a 75% Admixture
TreeMix output for data simulated according to the model in figure 5.35. A 75% admixture was used ($w = 0.75$). To fit in with Pritchard and Pickrell's assumptions as closely as possible on the drift parameters near an admixture event, c_2 and c_6 were made very small (0.0001). The other true drift parameters were 0.05.

If the proportion is pushed even higher to an admixture with $w = 0.95$, a different value of 0.100 is returned for the admixture parameter. This corresponds to a w of $1 - 0.100 = 0.900$. In these and similar experiments, TreeMix did retrieve the correct structure and suggest a sensible migration that corresponded with the simulated admixture, but as well as the drift parameters being difficult to interpret, only particular values of the admixture parameter seem to be possible and these were not particularly close to the true value.

It might be wondered what would happen if the TreeMix model assumptions about two of the three periods of drift adjacent to an admixture being zero were violated to some extent and to what extent can they be bent without the model badly failing. To address this, all $c_1, ..., c_9$were set at 0.05 contrary to the TreeMix

assumption and w was set to 0.75 to simulate the data. The TreeMix results are shown in figure 5.38. As can be seen from that figure, a completely wrong migration is selected. The migration parameter for the selected migration was again 0.397.

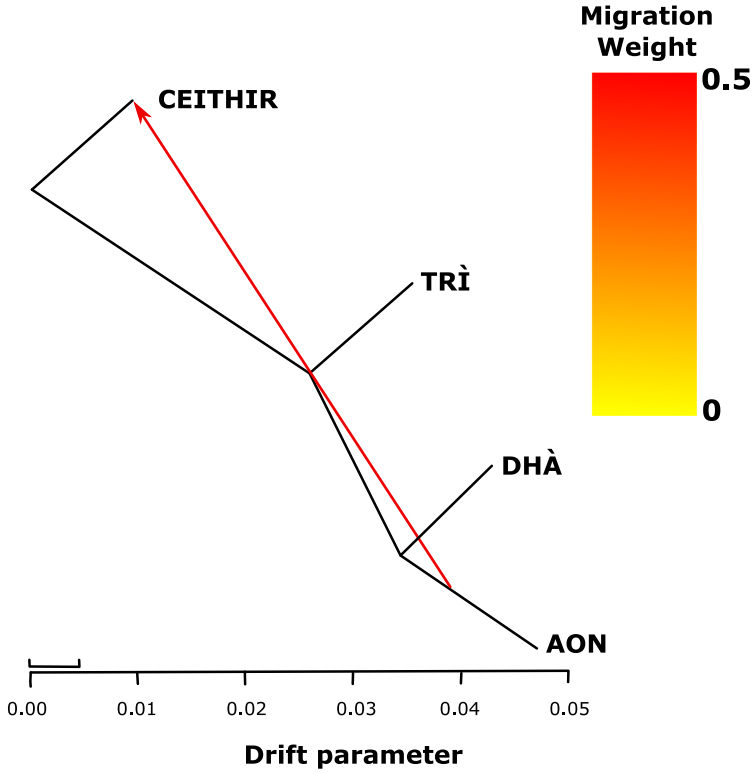

Figure 5.38: TreeMix Output for a 75% Admixture with Data Simulated That Does Not Conform to TreeMix Drift Assumptions Near an Admixture: All True Drift Parameters, Including Before and After Admixture set at 0.05

TreeMix output for data simulated according to the model in figure 5.35. A 75% admixture was used ($w = 0.75$). All true drift parameters were 0.05.

So clearly, there are situations in which TreeMix will suggest erroneous migrations when its model assumptions are not met. But how far do the assumptions need to be bent before TreeMix fails in this way? Keeping w at 0.75, if the drift marked c_2 in figure 5.35 is reduced to 0.0001, the model still chooses the same wrong migration. If c_2 is returned to 0.05 and c_6 is reduced to 0.0001, the model yet again chooses the same wrong migration, so if either drift parameter that TreeMix assumes to be 0 is too far from 0, TreeMix can fail. How far from 0 can they be before this happens? After some experimentation, it was found that when c_2 and c_6

were both 0.011, TreeMix produced the correct structure and suggested the correct migration, albeit with the wrong migration parameter of $1 - 0.09999 = 0.90001$ as shown in figure 5.39. If c_2 and c_6 were both 0.012 or higher, however, the wrong migration is selected as shown in figure 5.40. In this simple setting, it appears that if the true genetic drifts around an admixture are less than about 0.01, then the model can make sensible suggestions for the admixtures/migrations, but if they are larger than that, the inference becomes unreliable. In reality, there is no way to be sure which is the case by just using TreeMix, so the migrations suggested by TreeMix may be useful suggestions but should be treated with caution without investigating further in other ways. Like the drift parameters, it would be unwise to use the migration or admixture parameters suggested by TreeMix without also investigating these further, perhaps by using a more flexible model such as the one developed earlier in this chapter.

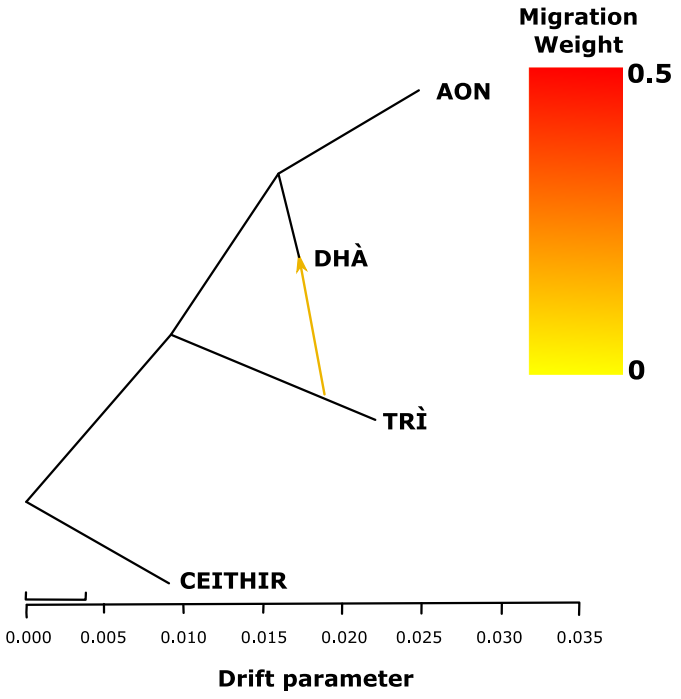

Figure 5.39: TreeMix Output for a 75% Admixture with Data Simulated That Does Not Conform to TreeMix Drift Assumptions Near Admixture: True Drift Parameters After, and the Migration Before Admixture Set at 0.011

TreeMix output for data simulated according to the model in figure 5.35. A 75% admixture was used ($w = 0.75$). To test the limits of Pritchard and Pickrell's assumptions the drift parameters near the admixture event, c_2 and c_6, were set to 0.011. The other true drift parameters were 0.05. This returns the correct structure.

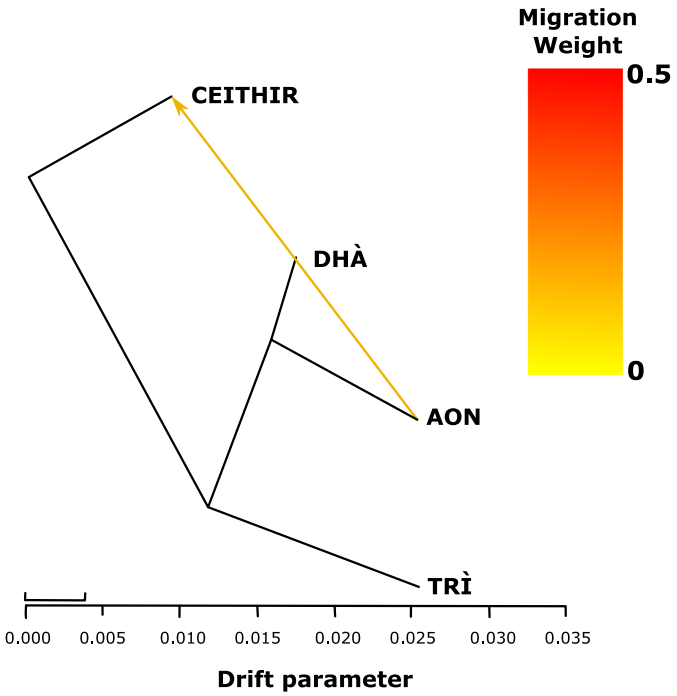

Figure 5.40: TreeMix Output for a 75% Admixture with Data Simulated That Does Not Conform to TreeMix Drift Assumptions Near Admixture: True Drift Parameters After, and the Migration Before Admixture Set at 0.012

TreeMix output for data simulated according to the model in figure 5.35. A 75% admixture was used ($w = 0.75$). To test the limits of Pritchard and Pickrell's assumptions the drift parameters near the admixture event, c_2 and c_6, were set to 0.012. The other true drift parameters were 0.05. This returns the wrong structure.

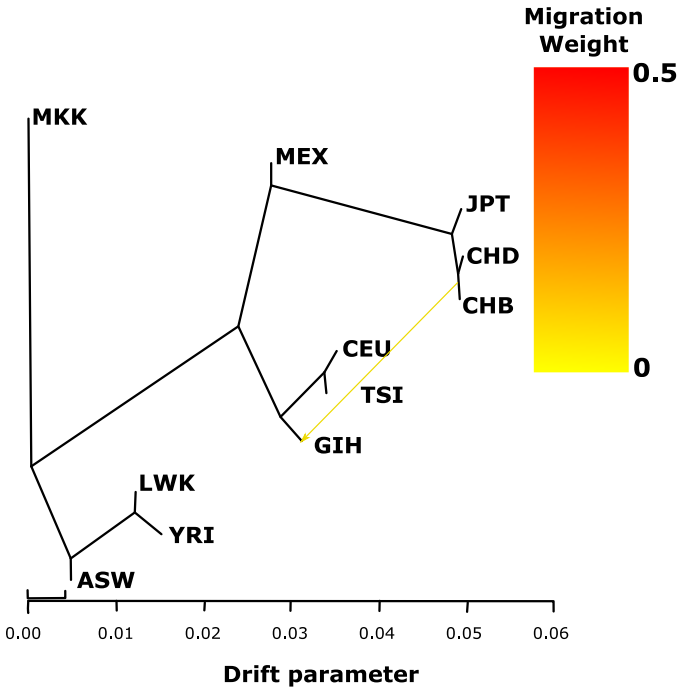

Figure 5.41: TreeMix Output for the HapMap Chromosome 2 Dataset with the Root Set Above the Maasai and with One Admixture Specified
Migration/admixture is displayed with a coloured arrow.

The HapMap dataset for Chromosome 2 was again fitted by TreeMix. This time TreeMix was instructed to suggest 1 migration or admixture, then it was asked for 2, then 3 and so on. The first migration it suggested was a bit unexpected, as shown in figure 5.41, it proposed that the Gujarati (GIH) could be modelled as an admixture of the Chinese (CHB and CHD) and the European branch ending in Tuscans (TSI) and Central Europeans (CEU) as shown, when a migrations leading to the Mexicans, Afro-Americans or even the Maasai might have been expected.

The next three migrations that TreeMix suggested were more expected. First the Mexicans were proposed as an admixture between Central Europeans and East Asians, then Afro-Americans as an admixture between Europeans and Nigerian Yoruba, then Maasai as an admixture involving Africans and Tuscans, as shown in figure 5.42. These three are all very reasonable migrations to suggest and, as seen earlier in the chapter, could be identified by other means. Note that the direction of the migration from Central Europeans to Mexicans chosen by

TreeMix does restrict the Central Europeans to contributing no more than 50% of the genetic information to the admixture. Earlier analysis by the more flexible model developed earlier in the chapter suggests that it is at least unclear whether Europeans contribute less than 50%. They may indeed have contributed slightly more.

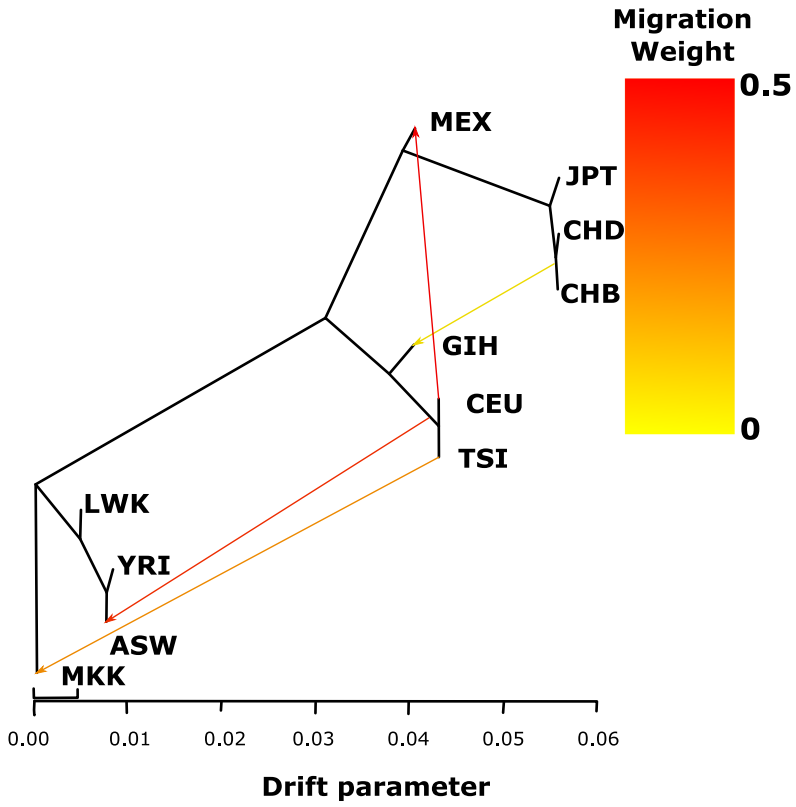

Figure 5.42: TreeMix Output for the HapMap Dataset with the Root Set Above the Maasai and Four Admixtures Specified

After this, the next two migrations, TreeMix suggests are, a migration from the Chinese to the Central Europeans (CEU) and one from Africans to the Lhosa in Kenya (LWK) as shown in figure 5.43. The former could be explained as a legacy of the Mongol invasion in the second half of the 13th century or an earlier migration.

The first four migrations that TreeMix suggested were the ones used earlier in the chapter and analysed using the model that was developed in the early part of this chapter. However, the first admixture that TreeMix suggested, involving the Gujarati, was the one which was eventually found to be the least important in terms

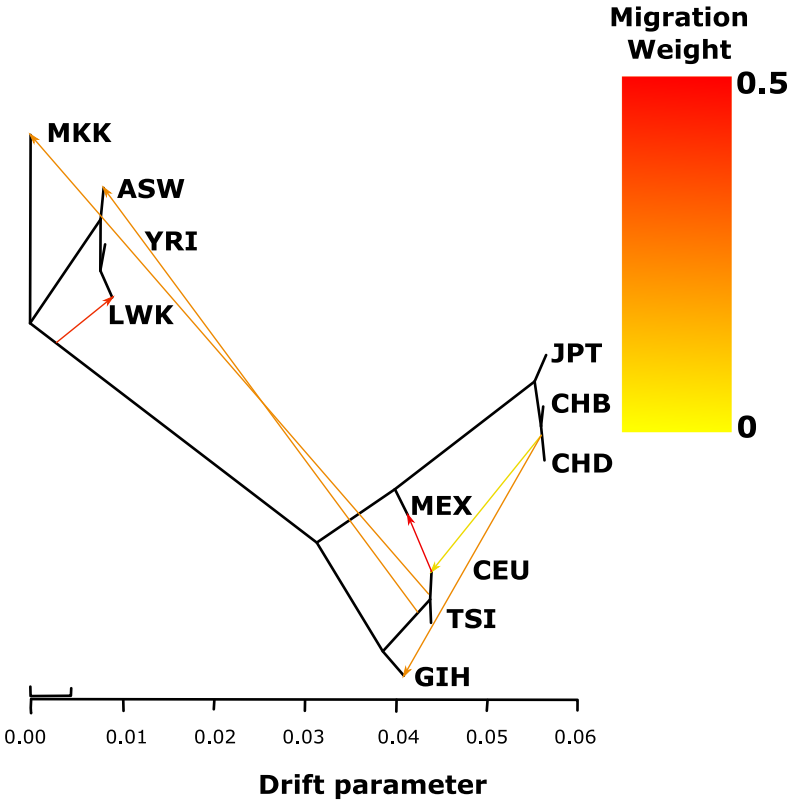

Figure 5.43: TreeMix Output for the HapMap Dataset with the Root Set Above the Maasai and Six Admixtures Specified

model fit and discarded after further analysis. It was, nevertheless, reasonably plausible a-priori and was worth looking at in the further depth that the more flexible Bayesian hierarchical model allowed. So, in practice, TreeMix made some useful suggestions in terms of migrations to investigate. However, migrations were not suggested in a believable order of importance, so it is worth asking TreeMix to produce more suggestions to analyse more deeply than are really expected to be used as in this case. The output itself cannot always be simply relied on, on its own, without further analysis.

5.6.5 Comparison of the Merits of the Two Approaches

TreeMix has the undeniable attraction that it runs very quickly and produces graphical output. It suggests trees and admixtures, only requiring the user to specify the tree root and number of admixtures required. In these respects, it does things that the more flexible model developed in the earlier part of this chapter does not even attempt to do. As is common for Bayesian hierarchical models, the latter takes many hours and sometimes days to run in order to obtain an adequate representation of the posterior distribution. A particular tree to be investigated must be specified as must the admixtures. These need to be suggested by other means or by examining the results (e.g., post predictive checks) of previous runs of analysis on other phylogenetic trees. The other thing that TreeMix does is allow the use of data that has not been thinned to ensure that the loci used are not in linkage disequilibrium and so can be modelled as independent. However, in exchange for these advantages, TreeMix makes a lot of assumptions that carry a price. Ignoring fixation, assuming that the allele frequency, α_j for all nodes j throughout the tree is approximately the same as that for the overall common ancestor, π_A, for the purposes of calculating variances and assuming all periods of genetic drift are small are quite restrictive assumptions that are not made in the model developed earlier in this chapter. As a result, TreeMix drift parameters are

dependent on the distribution of the π_A across loci in a non-transparent way that is not the case the model developed here. In fact, it has been shown earlier that when an outgroup is used with the more interpretable model developed here, the effect on the drift parameters near the root of the tree of misspecifying the prior on π_A almost vanishes. The consequence of the dependence on the distribution of the π_A across loci is that TreeMix drift parameters are much harder to interpret. The assumptions also mean that TreeMix is not robust when data that are uninformative about drift due to fixation are introduced. The phylogenetic trees it suggests have to be treated with caution since they can, for example, be radically altered simply by adding an extra subpopulation to the dataset as has been demonstrated above. Its suggestions do, as such, require critical investigation e.g., with a more flexible model such as the one developed here.

In the case of admixture, both models suffer in different ways from the problem of non-identifiability of drift parameters in the vicinity of an admixture event. TreeMix gets around the problem by applying hard constraints i.e., assuming there is drift only along one of the edges involved in the admixture. The Bayesian hierarchical model does not need to make such assumptions but pays a price in terms of slower mixing requiring more iterations of the Gibbs' sampler and therefore longer running time. It also leads to uncertainty about the values of drift immediately adjacent to the admixture e.g., as reflected in the diffuseness of their marginal posteriors, but it may be argued that this is an honest uncertainty that is preferable to making assumptions about some of these parameters that lead to a false level of certainty about the other parameters. Of course, cogent prior information about any of the drift parameters around an admixture can be reflected in informative priors for them. This can also be argued to make the output of the more flexible model more easily interpretable. The admixture parameter estimates from TreeMix have been observed to take one of a small number of particular values. In contrast, the output for the admixture parameters, w_j, from the model developed here is, as is always the case for Bayesian models, in the form of a posterior distri-

bution from which point estimates, measures of uncertainty and correlation with other parameters can be drawn. The w_j can take take all values in the range $(0, 1)$. In many applications of this type of model these parameters may be of particular interest and the more flexible model is able to give much fuller information about them. The admixture parameters do not suffer from the uncertainty issues arising from non-identifiability in the same way that the adjacent drift parameters do. These differences are summarised in table 5.11.

TreeMix is an interesting model the speed of which, like that of Neighbour Joining, can play a useful role in suggesting phylogenetic tree structures and additionally, potential admixtures for further investigation. The models it suggests are not always the best ones on further investigation and the parameters for drift and admixture in its output can be difficult to interpret and need to be treated with some caution. The more flexible models developed in this thesis dispense with many of the assumptions that TreeMix makes but take a long time to run and require phylogenetic trees and admixtures to consider to be set manually. Nevertheless, they do provide posterior distributions for the parameters from which more easily interpretable point estimates and measures of uncertainty can be derived. In particular, this is much more useful in situations where these parameters are of interest rather than the tree or network structure. The running time is not so much of an issue when the length of time it takes to gather sufficient reliable genetic data is taken into consideration. If analysing a genetic dataset is thought of as a multi-stage process the approaches become complimentary. Models like Neighbour Joining and TreeMix that run quickly can be used at an earlier stage to produce starting points and suggestions for analysis by more flexible models such as those developed in this thesis, that take considerably longer to run but which provide much more detailed and easily interpreted information about the model parameters, that properly quantify uncertainty.

Table 5.11: Comparison of Chapter 5 Model with TreeMix.

	Chapter 5 Model	TreeMix
Running time.	A few days.	A few minutes.
Suggests trees?	No, these need to be determined by other means.	Yes, but not always good ones.
Suggests admixtures?	No, these need to be determined by other means.	Yes, but not always good ones
Does data need to be thinned to take linkage disequilibrium into account?	Yes, assumes independence between loci.	No, but block sizes need to be set by the user. Assumes independence between blocks.
Takes fixation into account?	Yes.	No, completely ignores this issue.
Models allele frequencies for all subpopulations?	Yes.	No, assumes allele frequencies remain approximately the same as that in the ancestral population?
Are drift parameters interpretable in terms of time and effective population size?	Yes.	No, the drift parameters are dependent on the distribution of the ancestral population's allele frequencies and so are difficult to interpret.
Are admixture parameters interpretable?	Yes, allows the admixture parameters to take any value in $(0, 1)$.	Yes, but only allows the admixture parameter to be in $(0, 0.5)$ and even then only some particular values were observed.
Is output robust to extra uninformative data?	Yes, particularly if an outgroup is used.	No
How does it deal with the non-identifiability issue near the admixture?	Allows user to set strong informative priors on any, all or none of the drift and admixture parameters.	Has constraints hardcoded that assume no drift from one of the contributing subpopulations and no drift after the admixture.
Produces joint posterior distributions?	Yes, allows the uncertainty of the parameters and their joint relationships to be explored and allows the user to set their own favoured estimates of location and spread.	No, making estimation of uncertainty of the parameters difficult.

5.6.6 EIGENSTRAT

EIGENSTRAT (Price, 2017) was mentioned previously in section 1.5 and is described by Patterson et al. (2006) and Price et al. (2006). The first of these papers describes a simulation that was done to find out what pattern the principal components of an admixed population made compared to its two parent populations. It found that, if the first two principal components clustered the subjects from the two parent populations, those of the admixed population derived from these two populations formed a pattern that stretched between these two clusters. One way to support the admixtures detected in this chapter could be to use the principal components analysis of EIGENSTRAT (from within the SMARTPCA package) on the HAPMAP data for chromosome 2, find the first two principal components, plot the data projected onto those components and find out if the candidate admixed subpopulation does indeed form a pattern between the two proposed parent populations which themselves appear as clusters.

EIGENSTRAT uses data on individuals within subpopulations whereas the methods described in this chapter use aggregated data at the subpopulation level. The same procedure was used as described in section 3.4.1 with the exception that the data was not aggregated on subpopulation level and was instead processed into .map and .ped files which are formats that SMARTPCA can use as inputs.

The first ten principal components had eigenvalues 99.49, 45.11, 6.89, 6.03, 3.10, 3.03, 2.90, 2.89, 2.86, 2.80. The total of all the eigenvalues was 959. The two largest of these were much larger than all the others and together account for 15.1% of the total variation. Although this is not the majority of the variation in a large and complex dataset, using more than two eigenvectors would make visualisation more difficult. It would require 167 eigenvectors to take account of 50% of the variation, which is clearly an impractically large number for the purposes here. Using two eigenvectors should give an adequate impression of the data for the purpose of finding out if subpopulations can reasonably be modelled as admixtures. It does,

nonetheless, have to be borne in mind that it is far from the full picture.

A full plot of the data for all 11 subpopulations projected onto these two principal components is shown in figure 5.44. Many of the features are as expected. The two European subpopulations, CEU and TSI cluster near each other. The three East Asian subpopulations, JPT, CHB and CHD, also cluster near each other. The African subpopulations, ASW, MKK, LWK and YRI, are also very close to each other. While LWK and YRI appear as clusters (but not as close together, reflecting the greater genetic diversity in Africa), the patterns for MKK and ASW are more elongated. GIH and MEX are near each other but MEX has an elongated pattern rather than a cluster, while GIH could be argued to be closer to being like a cluster than MEX.

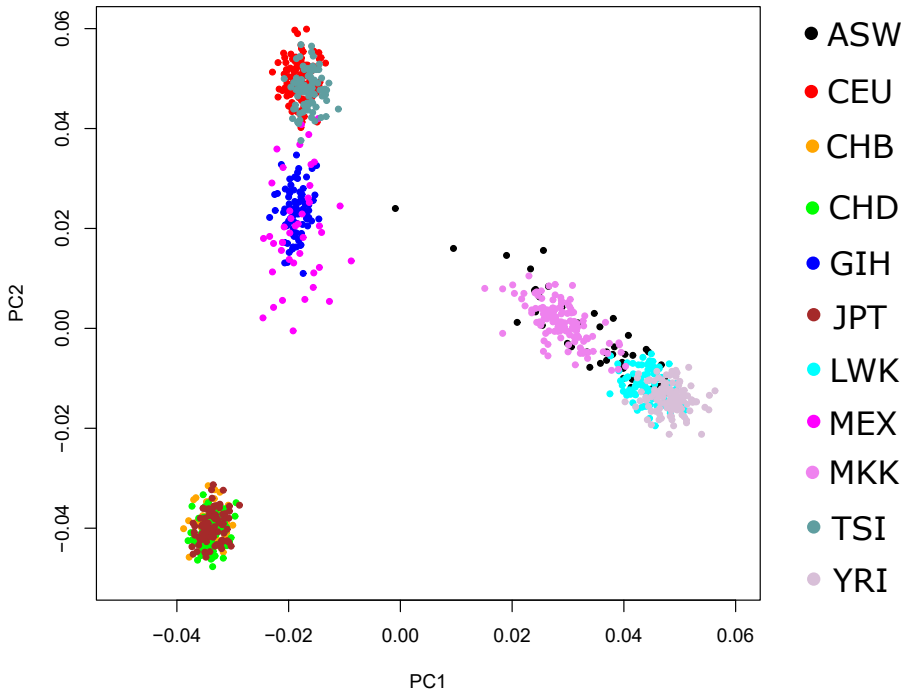

Figure 5.44: Plot of HAPMAP data for All 11 Subpopulations on First Two Principal Components

Looking at the candidate admixtures individually, figure 5.45 shows the Afro-American, (ASW) subpopulation plotted with the Yoruba (YRI) cluster and the two European clusters (CEU and TSI). ASW forms a pattern strung out away from the YRI cluster and towards the CEU and TSI clusters. This is a pattern consistent

with ASW being an admixture of ancestral YRI and European populations. That the pattern is nearer the YRI cluster is expected if YRI contributes more of the ASW genome on average than the European subpopulations do. This is consistent with the findings earlier in this chapter.

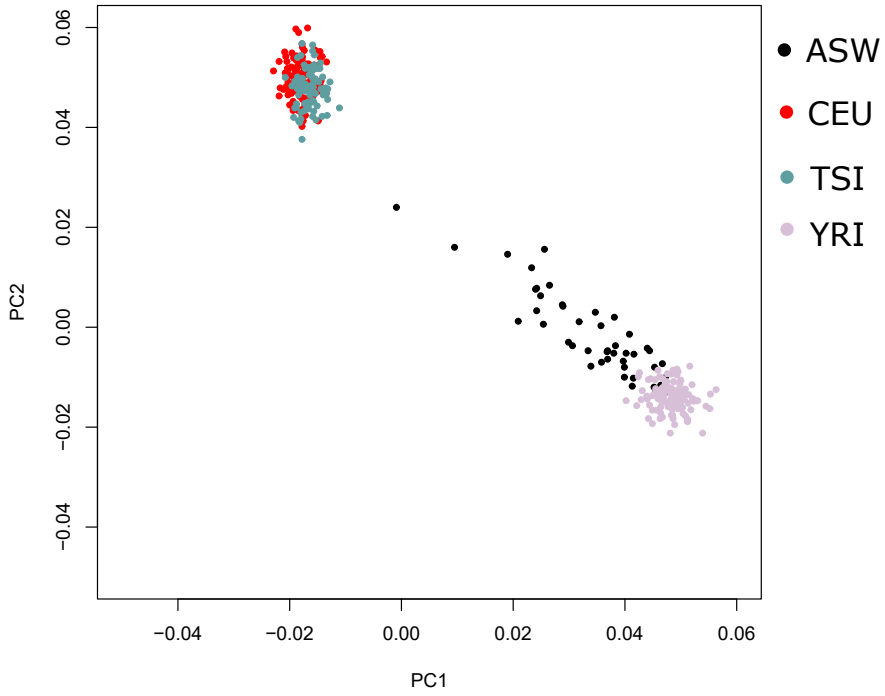

Figure 5.45: Plot of HAPMAP data for Afro Americans (ASW), Europeans (CEU and TSI) and Yoruba (YRI) Subpopulations on First Two Principal Components.

Turning next to the Mexicans, who are plotted in figure 5.46. The European (CEU) subpopualation is also shown, as are the three East Asian subpopulations (CHB,CHD and JPT). The pattern for MEX is strung out away from CEU and towards the direction of the East Asian subpopulations but not quite directly in their direction. The method used in this subsection assumes that the admixture happened sufficiently recently and sufficiently quickly that drift plays little significant part. There is plenty of reason to expect some drift between the ancestors of the Aztecs leaving East Asia and when they encountered Europeans which could easily explain this pattern. Again this pattern is consistent with the findings earlier in the chapter.

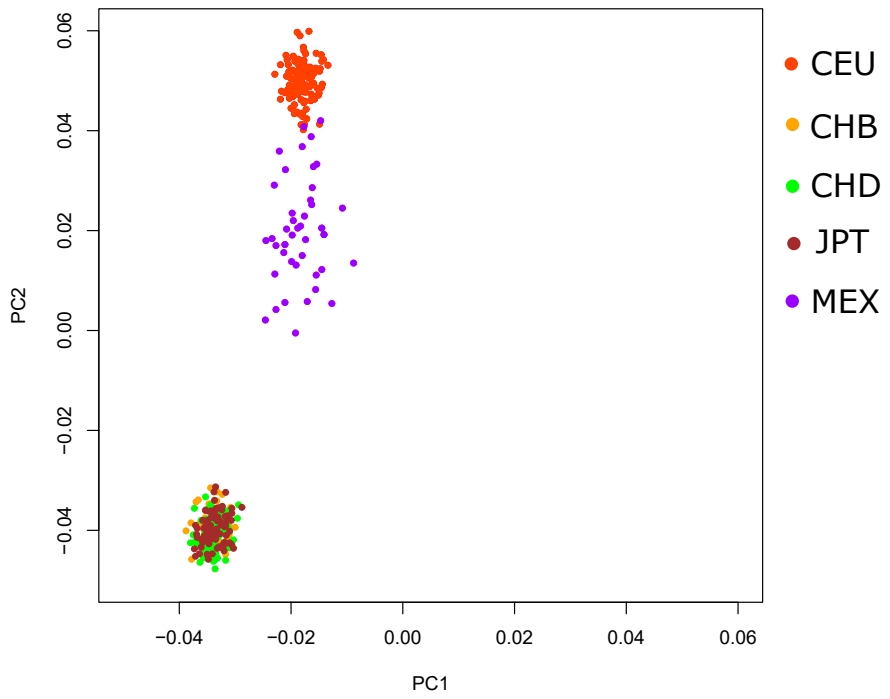

Figure 5.46: Plot of HAPMAP data for Central European (CEU), East Asian (CHB, CHD and JPT) and Mexican (MEX) Subpopulations on First Two Principal Components

Next, the Maasai, (MKK) are plotted along with two African clusters, (LWK and YRI) and the European Tuscans (TSI) in figure 5.47. The Maasai pattern is a bit tighter than the two earlier admixtured populations, that is closer to that of a cluster, but is still elongated between the two other African clusters and the TSI cluster, providing a hint of an admixture here but it is not as clear as the previous two cases. It is, nonetheless, not inconsistent with the findings earlier in this chapter.

The final candidate admixture is for GIH. This is plotted along with the European and East Asian subpopulations in figure 5.48. This time things are unclear. The GIH pattern can be interpreted as a cluster in its own right. It is not as elongated as the MEX pattern was but it could also be interpreted as an admixture between Europeans and the Asians if some significant drift has taken place or between the Europeans and an unsampled subpopulation. This difficulty in interpretation mirrors the difficulty found in placing GIH in the phylogenetic network earlier

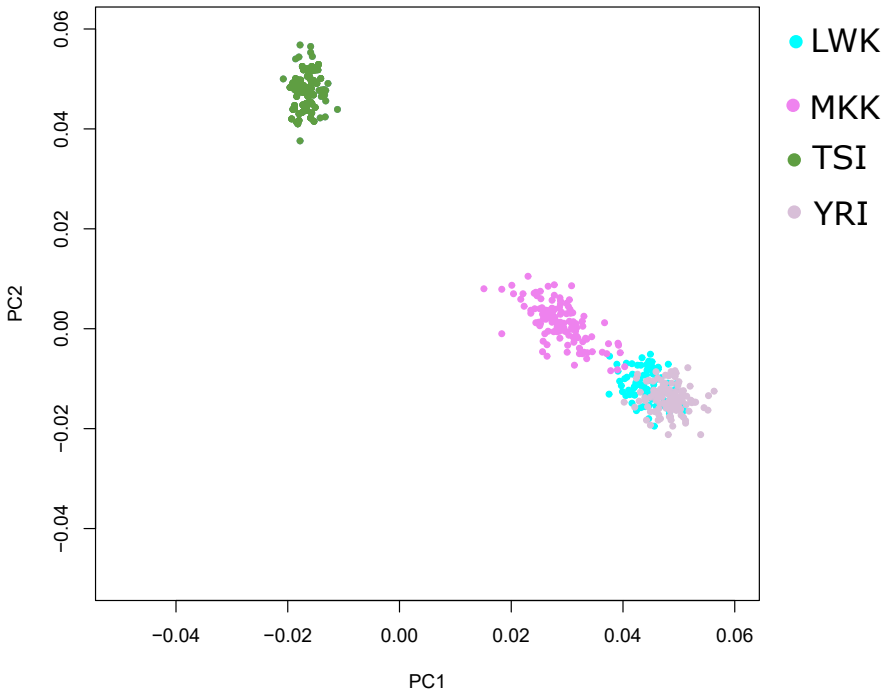

Figure 5.47: Plot of HAPMAP data for Tuscan (TSI), Lhosa (LWK), Maasai (MKK) and Yoruba (YRI) Subpopulations on First Two Principal Components

in the chapter. The finally chosen model does not treat them as an admixed population and there is no strong evidence against this interpretation provided by these plots.

The PCA approach of EIGENSTRAT, used here, produces plots that are consistent with the decisions on candidate admixture subpopulations that appeared earlier in this chapter and were also suggested by TreeMix.

5.7 Conclusions

This chapter developed the model from chapter 4 to include the possibility of admixture events. The increased complexity of the model and in particular, the problem of non-identifiability of the drift parameters that are adjacent to an admixture event in the hierarchy (and the concomitant posterior correlation of the parameters) necessitate the resulting model being run for many more iterations

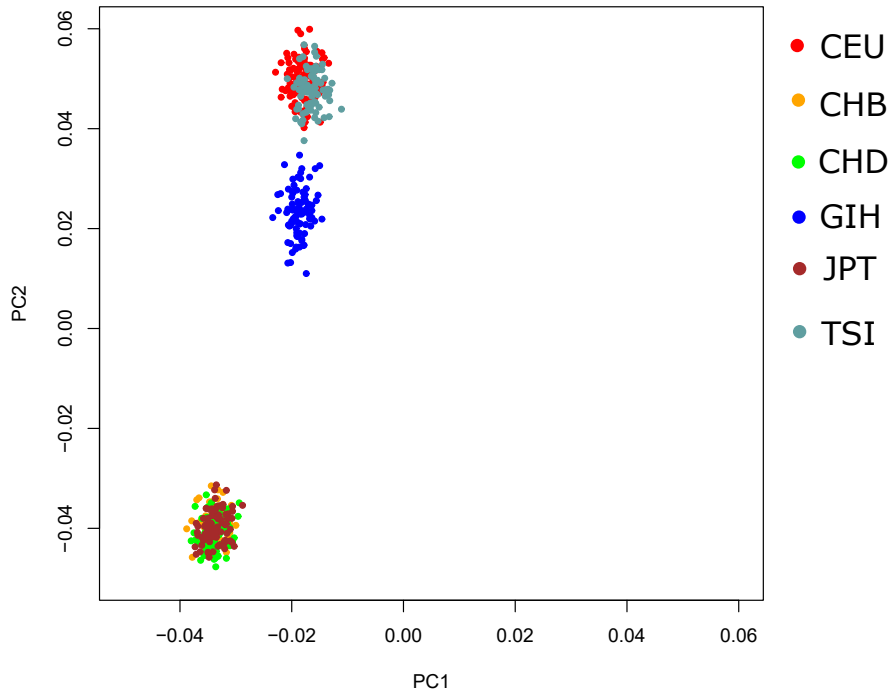

Figure 5.48: Plot of HAPMAP data for East Asians (CHB, CHD and JPT), Europeans (CEU and TSI) and Gujarati (GIH) Subpopulations on First Two Principal Components

than the one in the previous chapter, in practice, 100,000 iterations were used for the one chromosome of the HapMap data in this chapter, rather than the 20,000 iterations for the models in Chapter 4. This greatly increases the processor time required. Nevertheless, models including admixture do provide a much improved explanation of the observed data with WAIC values that are much lower despite the increased complexity. When applied to the HapMap data, it was judged that the model in figure 5.19 provided the most plausible model of the data despite not having the lowest WAIC of all the models considered. Models with lower WAIC had at least one implausibly large drift parameter.

The issue identified in the previous chapter that the model parameters for periods of drift adjacent to the ancestral population, were sensitive to the choice of prior on π, the allele frequency in the ancestral population, was found to be almost entirely mitigated by introducing an outgroup that did not even have to be made up from real data. It could instead be made up from an unweighted mean of

the observed frequencies in the data, with counts and sample size rounded to the nearest integer.

An existing model called TreeMix that appears to start from a similar standpoint to the models developed in this thesis but then introduces additional assumptions in order to take a more traditional likelihood-based approach was examined in detail. It has the attraction of running very quickly and suggests trees and admixtures where the models developed here need the trees and admixtures to be suggested by external means. However, it ignores the issue of fixation and assumes that the allele frequency, α_j for all nodes j throughout the tree is approximately the same as that of the common ancestor. It also assumes there is no drift along two of the three edges adjacent to an admixture. All of these are assumptions the model developed in this thesis does not make. TreeMix's drift parameters are based on variance and so depend on the distribution π_A in a way that makes them much harder to interpret than the ones for the model developed in this thesis. The phylogenetic trees and admixtures TreeMix suggests have to be treated with caution since they can, for example, be radically altered simply by adding an extra subpopulation to the dataset. Its suggestions do, as such, require further investigation with a more nuanced model such as the one developed in this and the previous chapter. Nevertheless along with Neighbour Joining it could play a useful role in suggesting phylogenetic tree structures and potential admixtures for deeper investigation using models such as the one developed here.

Chapter 6

Discussion

This thesis has focussed on examining and developing models of genetic drift and population history. There are some points that it would be useful to highlight in this closing chapter.

1. A model of drift has been developed that is sufficiently general that it can include admixture events. It builds on the foundation of Nicholson and others (Nicholson et al., 2002), placing an emphasis straight-forward demographic interpretation.

2. When modelling genetic drift over a great many generations, it is important to take proper account of the possibility of fixation if the intention is to obtain drift parameters that can be interpreted meaningfully.

These will be discussed in turn, some problems will be highlighted and suggestions for further development proposed which are aimed at mitigating or even eliminating these problems.

6.1 A Model of Genetic Drift that Accommodates Admixture Events

Increasingly general models of genetic drift and population history were developed over chapters 3 to 5. Models combining splitting events and isolated subpopulations, leading to tree relationships between the present-day subpopulations, were developed and these were further generalised to include admixture events.

A key practical issue with the admixture model was that of slow mixing, particularly of drift parameters adjacent to admixture events. This was dealt with by brute force, running the MCMC sampler for a much larger number of iterations. Many of the models described in chapter 5 took as long as 5 to 8 days on what was then a high-end Intel Core i7 processor to complete 100,000 iterations of the Gibbs sampler. In those situations, only 2,189 loci and 11 subpopulations were being modelled. If there were more loci to analyse and/or more subpopulations, the process would take at least proportionately longer. It might be argued that this renders the model impractical in such situations. However, many techniques that are commonly used today would have been impractical using past technology. For example, in 1981, the then common and affordable NEC D780C-1 processor running at 3.25MHz was capable of 0.5 MIPS (million instructions per second) (Gamia, 2013). In 2016, the Intel Core i7 6700K on which the analysis for chapter 5 was done is capable of over 160,000 MIPS at 4GHz (Scott, 2015). Moore's Law, which has held true or been surpassed over the past 40-50 years, predicts the number of transistors that can be crammed into the same area of an integrated circuit doubles every two years (Moore, 1975) and processing power has increased similarly as a consequence. If processing power continues to double every two years, a statistical model that takes a week to process in 2016 will take just 5.25 hours by 2026 and under 10 minutes by 2036. Something that seems barely feasible now can be reasonably expected to be imminently feasible. There is no reason to refrain from developing such models now to show they are conceptually sound in anticipa-

tion of future technology. Analysis time should also be put into the context of the time it takes to design and implement a population genetics study. Nevertheless, there are ways that could possibly be used now to speed the process up. Many of the parameters of the model have full conditionals that do not depend directly on each other. The drift parameters do not depend on each other and the allele frequency parameters for different loci do not depend directly on each other. The multi-core, multi-threaded nature of the processor was used during this project to run up to seven models simultaneously. However, if only one model was of interest, rewriting the Gibb's sampler to make use of multi-threading could allow many of these parameters to be sampled simultaneously.

The slow mixing was largely related to the weak non-identifiability of the drift parameters adjacent to an admixture event. This makes these parameters very correlated in their joint posterior distributions. Rewriting the Gibbs sampler to use block updating of these parameters, drawing them from joint full conditionals should improve mixing and potentially also speed things up by not requiring the sampler to be run for so many iterations (if sampling from those joint full conditionals is not a bottleneck). Weak priors were used for the drift parameters to reflect an a-priori position of being indifferent between all the possibilities. However, there may be situations where more information is available for the periods of drift leading up to admixture and could justify the use of much more informative priors. This ability to incorporate additional information is a key advantage of the flexibility of the Bayesian approach. For example, take the case of Iceland which was settled by a mixture of people of Norse and Celtic descent in the 9th and 10th century AD (Helgason et al., 2001). The Icelandic people wrote down a lot about their origins and early settlement of their island in documents such as the Landnámabók (Palsson and Edwards, 2007) as well as the Sagas of the Icelanders, the Íslendingasögur (Smiley, 2005). If study of these documents gave even rough information about the numbers, origins and time of the arrivals during the period of settlement, that would provide estimates or at least sensible ranges

for the parameters associated with the admixture event that could be reflected in much more informative priors for these parameters. Stronger priors for the drift parameters leading up to the admixture event would help mitigate the problem of weak non-identifiability.

A scenario that is not modelled is that of low-level recurrent migration between subpopulations. The model assumes that when each subpopulation splits at each bifurcation, the only way that they can come back into contact with the other subpopulation or any other subpopulation is through admixture. However, it is possible that enough low-level recurrent migration takes place between two subpopulations, to influence each other's allele frequencies, so that it cannot be ignored, but not enough that they can be reasonably treated as the same subpopulation with common allele frequencies. The effect of such low level migration would be to move the allele frequencies of the affected subpopulations closer to each other. If such a situation occured but were modelled by the present model, this would be expected to manifest in reduced genetic drifts if the subpopulations involved were close to each other in the phylogenetic tree or if they were not close to each other in the tree, in a pairwise posterior predictive F_{ST} that was low, indicating that the two subpopulations are more alike than the current model allows. Incorporating the possibility of low-level migrations may be another way that the model could be further refined. Models of isolation with migration, e.g., Hey (2009), can involve fitting a great many migration parameters which in that study were found to have some sensitivity to choice of prior so such a refinement might not be straightforward.

One of the differences between TreeMix and the model developed here from chapter 4 onwards is that TreeMix does suggest a phylogenetic tree while the model developed in this thesis does not. Instead other means such as outside knowledge or the Neighbour Joining algorithm are used to suggest a tree to analyse. Neither method is able to quantify uncertainty about the chosen tree or suggest how likely alternative trees are. A possible future development of the model could address

this issue. Huelsenbeck and Ronquist (2001) and Yang (2006) describe a method that could be adapted to do this. While these papers describe a substitution model as might be appropriate if mutation is being modelled, the ideas could be adapted to the drift model. If $\mathfrak{T}_{\mathfrak{l}}$ represents the $\mathfrak{l}$th possible tree topology out of $\mathfrak{P}$ possible trees, then letting D represent the data and for simplicity (and to focus on the impact of the tree toplologies), θ represent all the other parameters of the model, α, π and c, then the posterior probability of tree $\mathfrak{T}_{\mathfrak{l}}$ is

$$p(\mathfrak{T}_{\mathfrak{l}}|D) = \frac{p\left(D|\mathfrak{T}_{\mathfrak{l}}\right) f\left(\mathfrak{T}_{\mathfrak{l}}\right)}{\sum_{\mathfrak{n}=1}^{\mathfrak{P}} p\left(D|\mathfrak{T}_{\mathfrak{n}}\right) f\left(\mathfrak{T}_{\mathfrak{n}}\right)}, \tag{6.1}$$

where

$$p\left(D|\mathfrak{T}_{\mathfrak{l}}\right) = \int_{\theta} p\left(D|\mathfrak{T}_{\mathfrak{l}}, \theta\right) f\left(\theta\right) d\theta \tag{6.2}$$

The term $p\left(\mathfrak{T}_{\mathfrak{l}}\right)$ is a prior which can either be taken to have value $\frac{1}{\mathfrak{P}}$, a prior representing prior indifference between any of the possible trees, or set to other values that could reflect other outside (e.g., archaeological) information about how likely the tree topology is before considering the data. In this framework a full conditional for a given tree could be derived and tree topology could besampled as an additional Metropolis-Hastings within Gibbs step. After starting with an initial random tree, the steps of updating the parameters for the given tree are performed as usual followed by a proposal to change the tree topology to another similar but different one, for example, by proposing to remove a (non root or leaf) node and reattach one of the two subtrees attached to it to another edge on the tree, creating a new node.

This does, however, raise a problem. When a move to a new tree topology is proposed, the parameter values at that step will be appropriate to the existing one, making such a transition less likely to be accepted. Looking at it in likelihood space, there will be a lot of local maxima for combinations of trees and parameters, separated by chasms of low likelihood, with it being very unlikely for the Metropolis-Hastings step to successfully jump between the hypervolumes around

them. One way around this that is suggested by Huelsenbeck and Ronquist (2001) is to use Metropolis-coupled Markov chain Monte Carlo $(MC)^3$. This involves running a number of chains alongside the main chain. The difference is that in these additional chains, the chance of the chain moving between tree topologies is increased by different proportions. After all the chains have updated at the end of an iteration, there is a proposal to swap the states of two of the chains which are chosen at random. If accepted the two chains swap states and the MCMC processes continue at the next iteration. In this way, the main chain can be chosen as one of these two chains, making the proposal more likely to move near to one of the other likelihood peaks, offering the possibility for the main chain to leap over a likelihood chasm, changing to a state with another tree topology and set of parameter values in one move. One practical problem here is that the likelihood expressions that would need to be calculated to within proportionality for this step could be challenging.

While the idea of incorporating the tree topology in the model in this way appears theoretically possible, there are a number of practical drawbacks. Running the chains for models as described in chapters 4 and 5 could take as long as a week to produce 100,000 iterations for a particular tree or network even on a fast processor in C++. If, in addition, the chains were exploring the very large number of possible trees for 11 subpopulations, the time taken to produce a useful posterior distribution could stretch from weeks into many months, something that is not really practical at present but could become possible in the future with improved technology. Using $(MC)^3$ would be necessary and would also place even greater load on processor time. However, when it does become practically possible, it would enhance the models developed here because it would not only suggest the most likely tree topology but also describe other likely ones and how likely these are by examining the proportions that each tree topology was represented in the posterior distribution of the main chain. It may even be possible to extend this framework to suggest other changes to the tree or network topology, for example

proposing that a subpopulation becomes an admixture of the subpopulations at two other nodes, or to remove an admixture and attach it to only one of its two parent subpopulations. These operations would bring the added complications of changing the number of parameters and would increase the number of possible network topologies to be explored again, with another consequent increase in processor time but would allow uncertainty about the admixture events to be explored as well.

The issue of ascertainment that is covered by Nicholson et al. (2002) is not addressed in this thesis. The problem arises because the SNPs in historical datasets are not random loci in the genome but have been identified from variability in small samples. Often particular subpopulations will be over-represented in the ascertainment process. The SNPs will tend to have allele frequencies that are in the mid-range around 0.5 and fewer over 0.9 or under 0.1 than would be expected by random chance simply because they have more chance of being identified as SNPs. This is an issue with the HapMap dataset that has been used in this work. However, looking towards the future, with genome-wide sweeps taking in more and more loci and even full genome resequencing becoming increasingly common, the bias towards mid-range allele frequencies could be expected to be greatly reduced removing ascertainment effects as an issue.

On the topic of full genome resequencing, in the near future a greater amount of information will become available. The models developed in this thesis assume independence between loci and as they stand, cannot make full use of that information, needing to keep the loci used in analysis sufficiently spaced to avoid any serious linkage disequilibrium. The model could be developed to make more use of it. However, much of this additional information will be highly correlated due to linkage disequilibrium. It is unclear how far it would provide additional useful information. Developing and practically testing such a model would be necessary without any guarantee that it would improve greatly on the existing one. Even if it did, the additional processing time for the increased amount of data should be

balanced against any gains.

6.2 Importance of Fixation

The problems described in chapters 3 and 4 of using a beta distribution based model of genetic drift led to an unexpected finding as locus drifts towards fixation. A beta distribution model doesn't allow fixation but at least as importantly, in that model, it takes a greater and greater amount of drift to change the allele frequency by the same amount towards fixation, the nearer it starts to fixation. Very small changes in allele frequency of a rare allele can thus be taken as signals of disproportionately large amounts of drift, much larger than in the Wright-Fisher model that is the one that is being approximated. One option would have been to modify the beta distribution model to allow the possibility of fixation and modify its behaviour near allele frequencies of 0 and 1. However, an alternative model proposed by Nicholson et al. (2002) was readily available based on a normal distribution rectified at 0 and 1 which inherently took fixation into account. Models with this rectified normal model of drift fitted the data dramatically better as reflected in reduced WAIC values. This was a particularly interesting finding. Apart from the work of Nicholson et al. (2002), much of the previous work in the area of modelling genetic drift has avoided the issue of explicitly modelling fixation, often arguing that allele frequencies are insufficiently far from 0.5, populations are so large or time scales too small to warrant taking fixation into account. However, anthropological investigations inevitably involve large time scales. Newly discovered SNPs would be expected to have small allele frequencies. Fixation itself can be highly suggestive of ancestry; if a number of present-day subpopulations are observed to have reached fixation at a locus but others have not, it is possible that all these subpopulations reached fixation independently but it is also a likely explanation of their present day fixation that some or all of these subpopulations share a common ancestor in which fixation had already occurred. Treating fixation

did make the modelling process somewhat more complicated but repaid that effort in a model that better fitted the data and with drift parameters that are much more easy to interpret.

One of the arguments made for including fixation is that it is necessary in a model that allows long time scales to be modelled. Arguably the same could be said for mutation. The model developed in this thesis is open to the criticism that it assumes no mutation at any locus since the population at the root of the tree. When outlier residuals for some models were examined, for example during the analysis in chapter 3, it did happen very occasionally that these occured in situations where one subpopulation had a mid-range allele frequency while all the others were at fixation. This could always have happened by genetic drift, but another explanation is that a mutation event occurred in only one subpopulation some time near its foundation or during a population bottleneck and that mutation drifted to become common in that subpopulation. Although mutation events at a particular locus that result in a variant that does not soon die out are usually too uncommon to be worth modelling, over a sufficiently large number of generations and at a sufficiently large number of loci, the probability that they do happen could become appreciable. One possible future development of the model could be to take mutation into account. There would only be need to model a mutation event at a bifurcation where the allele frequency for one child node is at fixation and not at the other. A Metropolis-Hastings step could be added to decide whether to add (or remove) a mutation event that would have the effect of changing the allele frequency to that branch or not while leaving the parent node at the start of the bifurcation at fixation. However, this would be no easy task. In any event, the number of loci investigated as outliers where there was a suspicion that mutation might be the reason were very rare so the effort in making such a refinement was judged, for the purpose of this thesis, unlikely to repay the investment.

A drawback of using a model of drift based on the rectified normal distribution is that it is not exactly equivalent to the Wright-Fisher model, one of which is that

the expected value of an allele frequency is its last observed one no matter how many generations have elapsed since then. (Nevertheles, even the Wright-Fisher model is still just a model and does not fully reflect reality.) The mean of the rectified normal distribution used is not the same as the allele frequency before drift which is used as the location parameter in the distribution. The expected allele frequency after drift shifts slightly towards 0.5 compared to the allele frequency before drift using the rectified normal model of drift. In the human population context that this thesis is mostly concerned, a drift parameter larger than 0.4 seems unlikely. Then the largest expected shift of allele frequency towards 0.5 of the allele frequency that would result is only 0.036 for an allele frequency of 0.904 (or 0.096) so is unlikely to affect much. However, this issue does need to be borne in mind if it is intended to use the model on other species where larger genetic drift parameters might well be encountered.

6.3 Closing Remarks

Over the course of this thesis a model of population history has been developed that uses genetic data from present-day people grouped into subpopulations. It builds on earlier work by Nicholson and others and of Balding and Nichols and generalises it to, not only model the relationships and history of these subpopulations as phylogenetic trees starting from a common ancestral population, but also allows subpopulations to be formed from admixtures of earlier ones. The model could be used in a number of areas such as genome-wide association studies in medical genetics but most obviously lends itself to anthropological investigations. From that point of view, the parameters of the model, of genetic drift, admixture proportion and allele frequency have readily interpretable demographic and genetic meanings. Since draws from the joint posterior distribution of the parameters are produced by the model, point estimates of the parameters, uncertainty about those estimates and relationships between them can readily be assessed. It

is these properties that should be attractive to future investigators.

Bibliography

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium on Information Theory (eds. Petrov, B.N. and Csaki, F.), 267–281, Akadémiai Kiadó, Budapest.

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

Alexander, D., Shringapure, S., Novembre, J., Lange, K., 2017. Structure. Accessed on 19 September 2017.
URL www.genetics.ucla.edu/software/admixture/download.html

Balding, D., 2006. A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7, 781–791.

Balding, D., Bishop, M., Cannings, C., 2007. Handbook of Statistical Genetics. Wiley, Chichester.

Balding, D., Nichols, R., 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96, 3–12.

Baptist, E., 2001. Cuffy, fancy maids and one-eyed men: Rape, commodification, and the domestic slave trade in the United States. The American Historical Review 106, 1619–1650.

Barreiro, L., Laval, G., Quach, H., Patin, E., Quintana-Murci, L., 2008. Natural selection has driven population differentiation in modern humans. Nature Genetics 40, 340–345.

Bernstein, F., 1931. Die Geographische Verteillung der Blutgruppen und ihre Anthropologische Bedeutung. Comitato Italiano per lo Studio dei graphico dello Stato, 227–243.

BioNinja, 2017. Meiosis. Accessed on 3 August 2017.
URL http://www.vce.bioninja.com.au/aos-3-heredity/cell-reproduction/meiosis.html

Blattner, F., Plunkett, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N. W., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B., Shao, Y., 1997. The complete genome sequence of Escherichia coli K-12. Science 277, 1453–1462.

Borradale, D., Kimlin, M., 2012. Folate degradation due to ultraviolet radiation: possible implications for human health and nutrition. Nutrition Reviews 70, 414–422.

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., Mountain, J. L., 2015. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. American Journal of Human Genetics 96, 37–53.

Casella, G., Robert, C. P., Well, M. T., 2004. Generalized accept-reject sampling schemes. Institute of Mathematical Statistics Lecture Notes – Monograph Series 45, 342–347.

Cavalieri, B., 1635. Geometrica Indivisibilibus Continuorum Nova Quodam Ratione Promota.

Cavalli-Sforza, L., Menozzi, P., Piazza, A., 1994. The History and Geography of Human Genes, Abridged Paperback Edition. Princeton University Press, Princeton.

Chen, M.-H., Shao, Q.-M., 1999. Monte Carlo estimation of Bayesian credible and HPD intervals. Journal of Computational and Graphical Statistics 8, 69–92.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician 49, 327–335.

Chikhi, L., Bruford, M., Beaumont, M., 2001. Estimation of admixture proportions: A likelihood-based approach using Markov Chain Monte Carlo. Genetics 158, 1347–1362.

Choisy, M., Franck, P., Cornuet, J.-M., 2004. Estimating admixture proportions with microsatellites: Comparison of methods based on simulated data. Molecular Ecology 13, 955–968.

Collins, 2017. Collins free online dictionary. Accessed on 26 September 2017. URL https://www.collinsdictionary.com/dictionary/english/occams-razor

Conrad, D., Keebler, J., DePristo, M., Lindsay, S., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C., Torroja, C., Garimella, K., Zilversmit, M., Cartwright, R., Rouleau, G., Daly, M., Stone, E. A., Hurles, M., Awadalla, P., 2011. Variation in genome-wide mutation rates within and between human families. Nature Genetics 43, 712–714.

Crow, J., Kimura, M., 1970. An Introduction to Population Genetics Theory. Harper and Row, New York.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, A., Holmes, S., Destro-Bizol, G., Coia, V., Wallace, D., Oefner, P., Torroni, A., Cavalli-Sforza, L., Scozzari, R., Underhill, P., 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. American Journal of Human Genetics 70, 1197–1214.

Darwin, C. R., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life., 1st Edition. John Murray, London.

Ekirch, A. R., 1990. Bound for America: The Transportation of British Convicts to the Colonies 1718 to 1775. Oxford University Press, Oxford.

Elias, S. A., Short, S. K., Nelson, C. H., Birks, H. H., 1996. Life and times of the Bering land bridge. Nature 382, 60–63.

Ellis, L., Huang, W., Quinn, A., Ahuja, A., Alfrejd, B., Gomez, F., Hjelmen, C., Moore, K., Mackay, T., Johnston, S., Tarone, A., 2014. Intrapopulation genome size variation in D. melanogaster reflects life history variation and plasticity. PLOS Genetics 10, e1004522.

Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14, 2611–2620.

Excoffier, L., Estroup, A., Cornuet, J.-M., 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics 169, 1727–1738.

Excoffier, L., Smouse, P., Quattro, M., 1992. Analysis of molecular variance inferred from metric distance among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131, 479–491.

Falush, D., van Dorp, L., Lawson, D., July 2016. A tutorial on how (not) to over-interpret structure/admixture bar plots. Accessed on 19 September 2017. URL www.researchgate.net/publication/305985032_A_tutorial_on_how_not_to_over-interpret_STRUCTUREADMIXTURE_bar_plots

Felsenstein, J., 1981. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution 35, 1229–1242.

Fisher, R., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Wotton-under-Edge.

Foster, R., 1988. Modern Ireland 1600 to 1972. Penguin Group, London.

Frichot, E., Mathieu, F., Trouillon, T., Bourchard, G., Francois, O., April 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196, 973–983.

Friz, C., 1968. The biochemical composition of the free-living Amoebae Chaos chaos, Amoeba dubia and Amoeba proteus. Comparative Biochemistry and Physiology 26, 81–90.

Gamia, E., 2013. Instructions per second. Accessed on 31 August 2017.
URL http://gaming.wikia.com/wiki/Instructions_per_second#cite_ref-d780_13-5

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian Data Analysis, Third Edition. CRC Press, Abingdon.

Gelman, A., Roberts, G., Gilks, W., 1996. Efficient Metropolis jumping rules. Bayesian Statistics 5, 599–608 Oxford University Press, Oxford.

Gilks, W., Richardson, S., 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall, London.

Hamzelou, J., 2016. Exclusive: World's first baby born with new 3 parent technique. Accessed on 23 September 2017
URL www.newscientist.com/article/2107219-exclusive-worlds-first-baby-born-with-new-3-parent-technique/.

Hartl, D., Clark, A., 1997. Principles of Population Genetics. Sinauer, Massachusetts.

Hastings, W., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Helgason, A., Hickey, E., Goodacre, S., Bosnes, V., Stefánsson, K., Ward, R., Sykes, B., 2001. mtDNA and the islands of the North Atlantic: Estimating the proportions of Norse and Gaelic ancestry. American Journal of Human Genetics 68, 723–737.

Hey, J., 2009. Isolation with migration model for more than two populations. Molecular Biology and Evolution 27, 905–920.

Hou, L., Faraci, G., Chen, D., Kassem, L., Schulze, T., Shugart, Y., McMahon, F., 2013. Amish revisited: next generation sequencing studies of psychiatric disorders among the Plain people. Trends in Genetics 29, 412–418.

Huelsenbeck, J., Ronquist, F., 2001. MR BAYES: Bayesian Inference of Phylogenetic Trees. Bioinformatics 17, 754–755.

Human Genome Project, 2003. Human genome project information archive. Accessed on 3 August 2017.
URL http://web.ornl.gov/sci/techresources/Human_Genome/project/index.shtml

International HapMap Consortium, 2003. The International HapMap Project. Nature 426, 789–796.

Kayser, M., de Knijff, P., 2011. Improving human forensics through advances in genetics, genomics and molecular biology. Nature Reviews Genetics 12, 179–192.

Kumar, S., Bellis, C., Zlojutro, M., Melton, P. E., Blangero, J., Curran, J. E., 2011. Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. BMC Evolutionary Biology 11, 293.

Leibniz, G., 1684. Nova Methodus Pro Maximus Et Minimus.

Lipson, M., Loh, P.-R., Sankararaman, S., Patterson, N., Berger, B., Reich, D., 2015. Calibrating the human mutation rate via ancestral recombination density in diploid genomes. PLOS Genetics 11, e1004550.

Lisker, R., Ramirez, E., Gonzalez-Villalpando, C., Stern, M., 1995. Racial admixture in a Mestizo population from Mexico City. American Journal of Human Biology 7, 213–216.

Llorente, M. G., Jones, E. R., Eriksson, A., Siska, V., Arthur, K. W., Arthur, J. W., Curtis, M. C., Stock, J. T., M. Coltorti and, P. P., Stretton, S., Brock, F., Higham, T., Park, Y., Hofreiter, M., Bradley, D. G., J. Bhak and, R. P., Manica, A., 2015. Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science 350, 820–822.

Llorente, M. G., Jones, E. R., Eriksson, A., Siska, V., Arthur, K. W., Arthur, J. W., Curtis, M. C., Stock, J. T., M. Coltorti and, P. P., Stretton, S., Brock, F., Higham, T., Park, Y., Hofreiter, M., Bradley, D. G., J. Bhak and, R. P., Manica, A., 2016. Erratum for the report Ancient Ethiopian genome reveals extensive eurasian admixture in Eastern Africa. Science 351.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., Taha, A., Shaari, N., Raja, J., Ismail, P., Zainuddin, Z., Goodwin, W., Bulbeck, D., Bandelt, H., Oppenheimer, S., Torroni, A., Richards, M., 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308, 1034–1036.

Making SNPs Make Sense, 2017. University of utah. Accessed on 3 August 2017. URL http://learn.genetics.utah.edu/content/precision/snips/

Mandal, A., 2014. Types of junk DNA sequences. Accessed on 3 August 2017. URL https://www.news-medical.net/life-sciences/Types-of-Junk-DNA-Sequences.aspx

Marable, M., 1999. How Capitalism Underdeveloped Black America: Problems in Race, Political Economy, and Society (South End Press Classics Series). South End Press, New York.

Martínez-Cortés, G., Fernández-Rodríguez, J. S.-F. L. G., Rubi-Castellanos, R., Rodríguez-Loya, C., Velarde-Félix, J. S., Muñoz-Valle, J. F., Parra-Rojas, I., Rangel-Villalobos, H., 2012. Admixture and population structure in Mexican-Mestizos based on paternal lineages. Journal of Human Genetics 57, 568–574.

McConville, S., 1981. A History of English Prison Administration: Volume I 1750 to 1877. Routledge & Kegan Paul, Abingdon.

Mendel, G., 1866. Versuche über Plflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn Bd IV für das Jahr 1865, 3–47.

Meng, J., Zhang, J., Chen, Y., Huang, Y., 2011. Bayesian non-negative factor analysis for reconstructing transcription factor mediated regulatory networks. Proteome Science 9, S9.

Moore, G., 1975. Progress in digital integrated electronics. Accessed on 31 August 2017.
URL http://www.lithoguru.com/scientist/CHE323/Moore1975.pdf

Nicholson, G., Smith, A., Jónsson, F., Gústafsson, O., Stefánsson, K., Donnelly, P., 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society B 64, 695–715.

Oliver, R., Fage, J., 1970. A Short History of Africa. Penguin, London.

Palazzo, A., Gregory, R., 2014. The case for junk DNA. PLOS Genetics 10, e1004351.

Palsson, H., Edwards, P., 2007. The Book of Settlements: Landnamabok. University of Manitoba Press.

Patterson, M., Price, A., Reich, D., 2006. Population structure and eigenanalysis. PLOS Genetics 2, e190.

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

Penny, D., Hendy, M., Holland, B., 2007. Phylogenetics: Parsimony, Networks and Distance Methods. In Handbook of Statistical Genetics (3rd Edition) (eds Balding, D.J., Bishop, M., Cannings, C), Volume 1, 489–532.

Penny, W., Mattout, J., Trujillo-Barreto, N., 2006. Chapter 35: Bayesian Model Selection and Averaging.

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., Reich, D., 2014. Ancient west Eurasian ancestry in southern and eastern Africa. Proceedings of the National Academy of Sciences of the United States of America 111, 2632–2637.

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

Poznik, D., Henn, B., Yee, M.-C., Sliwerska, E., Euskirchen, G., Lin, A., Snyder, M., Quintana-Murci, L., Kidd, J., Underhill, P., Bustamante, C., 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341, 562–565.

Price, A., 2017. Eigensoft. Accessed on 19 September 2017.
URL www.hsph.harvard.edu/alkes-price/software/

Price, A., Patterson, N., Weinblatt, R., Shadick, N., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38, 904–909.

Pritchard, 2017. Structure. Accessed on 19 September 2017.
URL web.stanford.edu/group/pritchardlab/structure.html

Pritchard, J., Przeworski, M., 2001. Linkage disequilibrium in humans: models and data. American Journal of Human Genetics 69, 1–14.

Pritchard, J., Stevens, M., Donnelly, P., 2000. Inference of population structures using multilocus genotype data. Genetics 155, 945–959.

Pyle, A., Hudson, G., Wilson, I., Coxhead, J., Smertenko, T., Herbert, M., Santibanez-Koref, M., Chinnery, P., 2015. Extreme-depth re-sequencing of mitochondrial DNA finds no evidence of paternal transmission in humans. PLOS Genetics 11, e1005040.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/

Rasmussen, M., Anzick, S., Waters, M., Skoglund, P., DeGiorgio, M., et al., 2014. The genome of a late pleistocene human from a clovis burial site in western Montana. Nature 506, 225–229.

Rayment, W., 2017. The age of exploration. Accessed on 26 September 2017.
URL www.indepthinfo.com/history/age-of-exploration.htm

Republic of South Africa, June 1985. Immorality and prohibition of mixed marriages amendment act. Accessed on 28 February 2017.
URL http://www.gov.za/sites/www.gov.za/files/Act%2072%20of%201985.pdf

Roberts, D., Hiorns, R., 1962. The dynamics of racial intermixture. American Journal of Human Genetics 14, 261–277.

Roberts, D., Hiorns, R., 1965. Methods of analysis of the genetic composition of a hybrid population. Human Biology 37, 38–43.

Roberts, G., Rosenthal, J., 2009. Examples of adaptive MCMC. Journal of Computational and Graphical Statistics 18, 349–367.

Rosenthal, J., 2010. Optimal proposal distributions and adaptive MCMC. Accessed on 26 June 2014.
URL http://www.probability.ca/jeff/ftpdir/galinart.pdf

Rosenthal, J., 2012. Adaptive Metropolis and Gibbs samplers. Accessed on 18 May 2014.
URL http://www.birs.ca/workshops/2012/12w5105/files/Rosenthal.pdf

Saitou, N., Nei, M., 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425.

Saruni, S. L., 2016. Sons and daughters of Maa. Accessed on 28 February 2017.
URL http://www.sarunimara.com/experience/the-maasai-people.htm

Schneider, S., Roessli, D., Excoffier, L., 2000. A Software for Population Genetics Data Analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Scott, 2015. How far we've come: 40 years of processing power. Accessed on 31 August 2017
URL http://scottsoapbox.com/2015/08/15/how-far-weve-come-40-years-of-processing-power/.

Sharma, P., 2014. Identiying the chief trading emporiums in Indian Ocean maratime trade c.1000 - c.1500. Researchers World : Journal of Arts, Science and Commerce V, 131–142.

Smiley, J., 2005. The Sagas of the Icelanders. Penguin, London.

Smith, B., 2007. Boa: An R package for MCMC output convergence assessment and posterior inference. Journal of Statistical Software 21, 1–37.

Smithsonian, 2017. Modern human diversity - skin color. Accessed on 27 September 2017
URL http://humanorigins.si.edu/evidence/genetics/human-skin-color-variation/modern-human-diversity-skin-color.

Spiegelhalter, D., Best, N., Carlin, B., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society B 64, 583–639.

Studier, J., Keppler, K., 1988. A note on the neighbour-joining algorithm of Saitou and Nei. Molecular Biology and Evolution 5, 729–731.

Swofford, D., Olsen, G., Waddell, P., Hillis, D., 1996. Phyogenetic Inference. In Molecular Systematics, 2nd ed (eds. Hillis, D.M. and Moritz, D. and Mable, B.K. Sinauer, Massachusetts, 407-514.

Theobald, D., 2012. 29+ evidences for macroevolution: The scientific case for common descent. Accessed on 7 April 2016.
URL http://www.talkorigins.org/faqs/comdesc/phylo.html#phylointro

Thompson, 1973. The Icelandic admixture problem. Annals of Human Genetics 37, 69–80.

Urban, T., 2014. Your family: Past, present, and future. Accessed on 3 August 2017
URL https://waitbutwhy.com/2014/01/your-family-past-present-and-future.html.

US Supreme Court, 1967. Loving v Virginia. Accessed on 28 February 2017.
URL http://caselaw.findlaw.com/us-supreme-court/388/1.html

Vehtari, A., Gelman, A., 2014. WAIC and cross validation in Stan. Accessed on 9 June 2015.
URL http://www.stat.columbia.edu/~gelman/research/unpublished/waic_stan.pdf

Verdu, P., Rosenberg, N., 2011. A general mechanistic model of admixture histories of hybrid populations. Genetics 189, 1413–1426.

Wang, J., 2003. Maximum-likelihood estimation of admixture proportions frm genetic data. Genetics 164, 747–765.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research 11, 3571–3594.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.

Wright, S., 1951. The genetic structure of populations. Annals of Eugenics 15, 323–354.

Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press.

Appendix A

Proof of the Formula for the Variance After d Periods of Genetic Drift

The general formula for the variance for a particular locus i after d periods of drift between the ancestral population and a particular present day subpopulation is stated to be approximately

$$\mathrm{Var}\left(\frac{x_i}{n_i}\,\middle|\,\pi, c\right) = \frac{\pi_i\,(1-\pi_i)}{n_i}\big(1+(n_i-1)\,[1-(1-c_d)\,(1-c_{d-1})\cdots(1-c_1)]\big). \tag{A.1}$$

As usual, i labels the locus, π_i represents the proportion of the allele at that locus in the ancestral population, and $c = (c_1, ..., c_d)$ represent drift parameters for d consecutive periods of drift. Further, x_i repesents the number of counts of one variant allele out of a sample of n_i at locus i. One use of this formula is to standardise residuals. This appendix provides a proof by induction for this formula. This proof has three steps. First, the formula will be proved in the case $d = 1$. Then, it will be shown that if the formula is assumed to be true for $d = k$

periods of drift then it must also be true that

$$\mathrm{Var}\left(\alpha_{ik} \;\middle|\; \pi, c\right) = \pi_i \left(1 - \pi_i\right) \left[1 - \left(1 - c_k\right) \left(1 - c_{k-1}\right) \ldots \left(1 - c_1\right)\right], \tag{A.2}$$

where α_{ik} is the proportion of the allele at locus i after k periods of drift have taken place since the ancestral population. This result is used in the final step to show that if the formula is assumed true for $d = k$ periods of drift then it must also be true for $d = k + 1$ periods of drift, completing the proof.

The variance in the case where $d = 1$ is contained in Nicholson et al. (2002) and is given as:

$$\mathrm{Var}\left(\frac{x_i}{n_i} \;\middle|\; \pi, c\right) = \frac{\pi_i \left(1 - \pi_i\right) \left(1 - c_1\right)}{n_i} + \pi_i \left(1 - \pi_i\right) c_1. \tag{A.3}$$

This formula, which follows from the law of total variance (also known as the variance decomposition formula) assumes that $\mathrm{E}\left(\alpha_{i1} | c_1, \pi_i\right) = \pi_i$ and $\mathrm{Var}\left(\alpha_{i1} | c_1, \pi_i\right) = \pi_i \left(1 - \pi_i\right) c_1$ which is only approximately true under their rectified normal mode of drift (see chapter 4), can be rearranged to provide the basis for the induction, completing the first step of the proof.

$$\mathrm{Var}\left(\frac{x_i}{n_i} \;\middle|\; \pi, c\right) = \frac{\pi_i \left(1 - \pi_i\right)}{n_i} \big(1 + \left(n_i - 1\right) \left[1 - \left(1 - c_1\right)\right]\big). \tag{A.4}$$

For the inductive step, if it can be assumed that the formula is true for $d = k$ periods of drift then it must be shown that it must also be true for $d = k + 1$ periods of drift where $k \in \mathbb{N}$ (Figure A.1).

For simplicity, since this proof concentrates on a particular locus the subscript i will be suppressed in what follows. First notice that in the case where there are

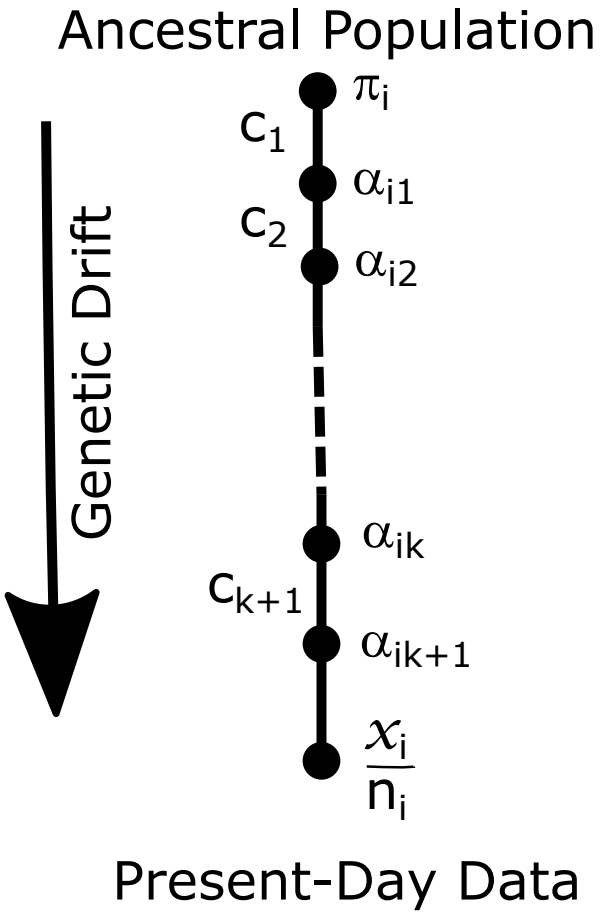

Figure A.1: Diagram of the Induction Step Showing $k+1$ Periods of Genetic Drift. *Diagram of the induction step showing $k+1$ periods of genetic drift between the ancestral population and the data at locus i. The direction of time is downward. The cs represent $k+1$ periods of genetic drift in series that have taken place since the time of the ancestral population whose proportion of the allele at locus i is represented by π_i. α_{ij} represents the proportion of the allele at every intermediate population, labelled byj, after each period of drift after the ancestral population. x_i represents the counts of the allele out of a sample of size n_i in the present-day data. The induction step is about showing that if the formula is true for k periods of drift then it is also true when the $(k+1)th$ period of drift is added.*

only $d = k$ periods of drift

$$\begin{aligned}
\mathrm{Var}\left(\frac{x}{n} \;\middle|\; \pi, c\right) &= \mathrm{E}_{\alpha_k}\left[\mathrm{Var}\left(\frac{x}{n} \;\middle|\; \alpha_k\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\frac{x}{n} \;\middle|\; \alpha_k\right)\right] \\
&= \frac{1}{n}\mathrm{E}_{\alpha_k}\left(\alpha_k\left(1-\alpha_k\right) \mid c_k, \alpha_{k-1}\right) + \mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right) \\
&= \frac{1}{n}\mathrm{E}_{\alpha_k}\left(\alpha_k | c_k, \alpha_{k-1}\right) - \frac{1}{n}\mathrm{E}_{\alpha_k}\left(\alpha_k^2 | c_k, \alpha_{k-1}\right) + \mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right) \\
&= \frac{1}{n}\mathrm{E}_{\alpha_k}\left(\alpha_k | c_k, \alpha_{k-1}\right) - \frac{1}{n}\left[\mathrm{E}_{\alpha_k}\left(\alpha_k | c_k, \alpha_{k-1}\right)^2 + \mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right)\right] \\
&\quad + \mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right) \qquad \text{(A.5)} \\
&= \frac{1}{n}\mathrm{E}_{\alpha_k}\left(\alpha_k | c_k, \alpha_{k-1}\right) - \frac{1}{n}\mathrm{E}_{\alpha_k}\left(\alpha_k | c_k, \alpha_{k-1}\right)^2 \qquad \text{(A.6)} \\
&\quad + \frac{n-1}{n}\mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right) \qquad \text{(A.7)}
\end{aligned}$$

However for any $m > 1$, where the subscripts on E and Var indicate what random variable is averaged over, $\mathrm{E}_{\alpha_m}\left(\alpha_m | c_m, \alpha_{m-1}\right) = \alpha_{m-1}$ and when $m = 1$, $\mathrm{E}_{\alpha_1}\left(\alpha_1 | c_1, \pi\right) = \pi$ by the assumption that the expected proportion of an allele is preserved under drift. This is a standard property of the Wright-Fisher model as was shown in Chapter 3. If it can be assumed, as Nicholson et al. did, that this is approximately true, then A.7 reduces to

$$\mathrm{Var}\left(\frac{x}{n} \;\middle|\; \pi, c\right) = \frac{1}{n}\pi\left(1-\pi\right) + \frac{n-1}{n}\mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right). \qquad \text{(A.8)}$$

Equating (A.8) and (A.1), it can be seen that the inductive assumption that the formula at (A.1) is true for $d = k$ then implies that

$$\begin{aligned}
&\frac{1}{n}\pi\left(1-\pi\right) + \frac{n-1}{n}\mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right) \\
&\qquad = \frac{\pi\left(1-\pi\right)}{n}\left(1 + \left(n-1\right)\left[1 - \left(1-c_k\right)\left(1-c_{k-1}\right)\ldots\left(1-c_1\right)\right]\right),
\end{aligned}$$

which, in turn implies that

$$\mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right) = \pi_i\left(1-\pi_i\right)\left[1 - \left(1-c_k\right)\left(1-c_{k-1}\right)\ldots\left(1-c_1\right)\right]. \qquad \text{(A.9)}$$

So this can be assumed to be true if the inductive assumption can be assumed to be true. This completes the second of the three steps of the proof.

To begin the third step, in the case of $d = k + 1$ periods of drift,

$$\begin{aligned}
\mathrm{Var}\left(\frac{x}{n} \,\middle|\, \pi, c\right) &= \mathrm{E}_{\alpha_{k+1}}\left(\mathrm{Var}_{\frac{x}{n}}\left(\frac{x}{n} \,\middle|\, \alpha_{k+1}\right)\right) + \mathrm{Var}_{\alpha_{k+1}}\left(\mathrm{E}_{\frac{x}{n}}\left(\frac{x}{n} \,\middle|\, \alpha_{k+1}\right)\right) \\
&= \frac{1}{n}\mathrm{E}_{\alpha_{k+1}}\left[\alpha_{k+1}\left(1-\alpha_{k+1}\right) \mid c_{k+1}, \alpha_k\right] + \mathrm{Var}_{\alpha_{k+1}}\left(\alpha_{k+1} \mid c_{k+1}, \alpha_k\right) \\
&= \frac{1}{n_i}\left[\mathrm{E}_{\alpha_{k+1}}\left(\alpha_{k+1}|c_{k+1}, \alpha_k\right) - \mathrm{E}_{\alpha_{k+1}}\left(\alpha_{k+1}|c_{k+1}, \alpha_k\right)^2\right] \\
&\quad + \frac{n_i - 1}{n_i}\mathrm{Var}_{\alpha_{k+1}}\left(\alpha_{k+1} \mid c_{k+1}, \alpha_k\right).
\end{aligned}$$

As before, if for any $m > 1$, $\mathrm{E}_{\alpha_m}\left(\alpha_m|c_m, \alpha_{m-1}\right) = \alpha_{m-1}$ and when $m = 1$, $\mathrm{E}_{\alpha_1}\left(\alpha_1|c_1, \pi\right) = \pi$ can be assumed to hold at least approximately, this can be reduced to

$$\begin{aligned}
\mathrm{Var}\left(\frac{x}{n} \,\middle|\, \pi, c\right) &= \frac{\pi\left(1-\pi\right)}{n} \\
&\quad + \frac{n-1}{n}\left[\mathrm{E}_{\alpha_k}\left(\mathrm{Var}_{\alpha_{k+1}}\left(\alpha_{k+1}|c_{k+1}, \alpha_k\right)\right) + \mathrm{Var}_{\alpha_k}\left(\mathrm{E}_{\alpha_{k+1}}\left(\alpha_{k+1}|c_{k+1}, \alpha_k\right)\right)\right] \\
&= \frac{\pi\left(1-\pi\right)}{n} \\
&\quad + \frac{n-1}{n}\left[\mathrm{E}_{\alpha_k}\left(\alpha_k\left(1-\alpha_k\right)c_{k+1} \mid c_k, \alpha_{k-1}\right) + \mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right)\right] \\
&= \frac{\pi\left(1-\pi\right)}{n} + \frac{n-1}{n}\left[c_{k+1}\mathrm{E}_{\alpha_k}\left(\alpha_k|c_k, \alpha_{k-1}\right) - c_{k+1}\mathrm{E}_{\alpha_k}\left(\alpha_k|c_k, \alpha_{k-1}\right)^2\right] \\
&\quad + \frac{n-1}{n}\left[\mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right) - c_{k+1}\mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right)\right]
\end{aligned}$$

Again, allowing the simplifications for any $m > 1$, $\mathrm{E}_{\alpha_m}\left(\alpha_m|c_m, \alpha_{m-1}\right) = \alpha_{m-1}$ and when $m = 1$, $\mathrm{E}_{\alpha_1}\left(\alpha_1|c_1, \pi\right) = \pi$ leads to:

$$\mathrm{Var}\left(\frac{x}{n} \,\middle|\, \pi, c\right) = \frac{\pi\left(1-\pi\right)}{n} + \frac{n-1}{n}\left[c_{k+1}\pi\left(1-\pi\right) + \left(1-c_{k+1}\right)\mathrm{Var}_{\alpha_k}\left(\alpha_k \mid c_k, \alpha_{k-1}\right)\right]. \tag{A.10}$$

Now substituting in (A.9) from the second step quickly gives

$$\mathrm{Var}\left(\frac{x}{n} \,\middle|\, \pi, c\right) = \frac{\pi\left(1-\pi\right)}{n}\left[1 + \left(n-1\right)\left\{1 - \left(1-c_{k+1}\right)\left(1-c_k\right)\cdots\left(1-c_1\right)\right\}\right], \tag{A.11}$$

which is the same as (A.1) where $d = k + 1$. The formula has been shown to be approximately true for $d = 1$ and that if it is approximately true for $d = k$ then it must also be approximately true for $d = k+1$, so by the principle of mathematical induction it must be approximately true for all $d \in \mathbb{N}$ completing the proof.

Comparing (A.1), the formula for d periods of drift, with (A.3), for one period of drift, a formula can be derived that gives a value of c for a single period of drift that is equivalent, at least in terms of variance, to that for d periods of drift in series:

$$c = \left[1 - (1 - c_d)\,(1 - c_{d-1}) \cdots (1 - c_1)\right]. \tag{A.12}$$

Noting that this implies that

$$1 - c = (1 - c_d)\,(1 - c_{d-1}) \cdots (1 - c_1)\,, \tag{A.13}$$

it can be seen that this ties in with the discussion on the interpretation of the drift parameter, c, in section 3.3.3.

Appendix B

Rectified Normal Distributions

Any random variable with a normal distribution can take values in the interval $(-\infty, \infty)$. This distribution shall be called the "conventional normal distribution". However, there are situations in which a normal distribution is a good approximation to the process being modelled but where only a subset of these values makes sense for that process. This has led to a number of modified versions of the normal distribution (figure B.1). One of the better known of these is the truncated normal distribution. Suppose for a given process only values in the interval $[a, b]$, where $a, b \in \mathbb{R}$:$a < b$, make sense; the truncated normal distribution uses only the part of the normal distribution that lies within that interval and renormalises it to produce a proper probability density function. This is equivalent to drawing from the conventional normal distribution, checking to see whether the draw lies in the interval $[a, b]$ and rejecting it and making a new draw if the draw lies outside the interval.

Another, lesser-known variant deals with the problem in a different way. This involves starting from the conventional normal distribution and taking all the probability below a and assigning it all to a probability mass at a. Similarly all the probability above b is assigned to a probability mass at b. This leads to a hybrid distribution which has a discrete part with probability masses at a and b and a

continuous part with probability density in the interval (a, b). Such a distribution is referred to by Meng et al. (2011) in the special case $a = 0$ and $b = \infty$. In that paper the distribution is referred to as a Rectified Normal Distribution. This takes its name from the result of putting a waveform through a half-rectifier circuit in electronics. In that case all the input signal that is below 0 leads to an output signal of 0 Volts but input signals above 0 remain unchanged. Nicholson et al. (2002) use this type of distribution with $a = 0$ and $b = 1$ in their paper. However, they refer to it as a truncated normal distribution. Such a nomenclature could be potentially confusing; it could be considered undesirable to have two probability distributions with quite different properties having the same name. For that reason, in this thesis this type of distribution will be referred to using Meng et al. (2011)'s name - the rectified normal distribution.

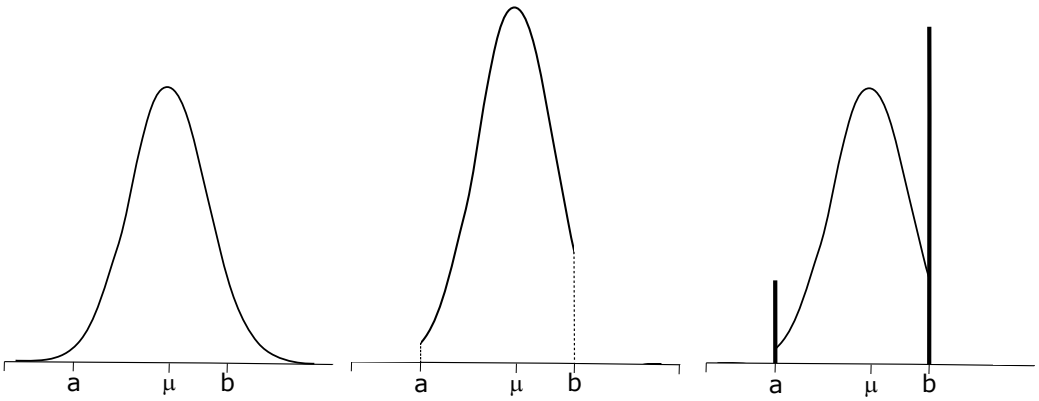

Figure B.1: Conventional Normal Distribution, Trucated Normal Distribution and Rectified Normal Distribution

A conventional Normal Distribution (left), a Truncated Normal Distribution with the same μ and σ^2 truncated at a and b (middle) and a Rectified Normal Distribution with the same μ and σ^2 rectified at a and b (right). The thick bars represent point masses at those values. These are probabilities rather than probability densities.

B.1 Notation for Rectified Normal Distributions

Meng et al. (2011) use the notation $N^R(\mu, \sigma^2)$ where μ is the mean and σ^2 the variance of the conventional normal distribution from which the rectified normal distribution is derived. It should be noted that the act of rectification means

that, in general, μ is no longer the mean and σ^2 no longer the variance of the rectified normal distribution. However Meng et al. (2011) were only referring to the distribution for the case where $a = 0$ and $b = \infty$. For this notation to be used more generally without ambiguity the range of acceptable values which the random variable can take must also be specified. The notation used in this thesis includes the interval by placing it after the R so that $N^{R[0,1]}(\mu, \sigma^2)$ would refer to the distribution used by Nicholson et al. (2002) and $N^{R[0,\infty)}(\mu, \sigma^2)$ would refer to that used by Meng et al. (2011); in deference to that paper, where the interval is omitted from the notation, the interval is assumed to be $[0, \infty)$ so that $N^R(\mu, \sigma^2) \equiv N^{R[0,\infty)}(\mu, \sigma^2)$. This last distribution is a case where rectification takes place only on the left of the distribution at 0. Such a distribution where the interval is of the form $[a, \infty)$ will be referred to as a left-rectified distribution. Similarly one where the interval is of the form $(-\infty, b]$ will be referred to as a right-rectified distribution (figure B.2).

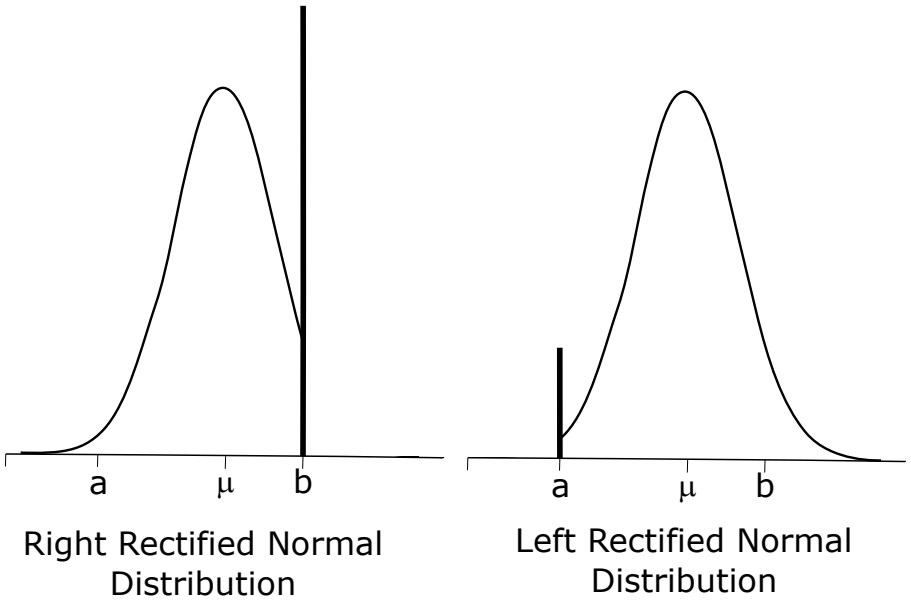

Figure B.2: A $N^{R(-\infty,b]}(\mu, \sigma^2)$ Right Rectified Normal Distribution and $N^{R[a,\infty)}(\mu, \sigma^2)$ Left Rectified Normal Distribution

B.2 First Two Moments of Rectified Normal Distribution

As has been mentioned, in general, μ is not the mean and σ^2 not the variance of $N^{R[a,b]}(\mu, \sigma^2)$. There is little literature on this distribution, which necessitated the moments to be found from first principles. These results are reproduced here.

B.2.1 Mean of a $N^{R[a,b]}(\mu, \sigma^2)$ Distribution

Splitting the distribution into three parts gives

$$\mathrm{E}(X) = \int_{-\infty}^{a} af(x)dx + \int_{a}^{b} xf(x)dx + \int_{b}^{\infty} bf(x)dx, \tag{B.1}$$

where $f(x)$ is the probability density function of the $N(\mu, \sigma^2)$ distribution. Hence

$$\mathrm{E}(X) = a\Phi\left(\frac{a-\mu}{\sigma}\right) + \int_{a}^{b} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx + b\left[1 - \Phi\left(\frac{b-\mu}{\sigma}\right)\right]. \tag{B.2}$$

Using the substitution $t = \frac{x-\mu}{\sqrt{2\sigma^2}}$ leads to

$$\int_{a}^{b} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx = \int_{\alpha}^{\beta} \frac{\sqrt{2\sigma^2}\, t + \mu}{\sqrt{2\pi\sigma^2}} \exp\left(-t^2\right) \sqrt{2\sigma^2} dt \tag{B.3}$$

where $\alpha = \frac{a-\mu}{\sqrt{2\sigma^2}}$ and $\beta = \frac{b-\mu}{\sqrt{2\sigma^2}}$. Performing the integral yields

$$\int_{a}^{b} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx = \sqrt{\frac{2}{\pi}}\sigma\left(\frac{1}{2}\left[\exp\left(-\alpha^2\right) - \exp\left(-\beta^2\right)\right]\right) + \frac{\mu}{2}\left[2\Phi\left(\sqrt{2}\beta\right) - 2\Phi\left(\sqrt{2}\alpha\right)\right]. \tag{B.4}$$

Substituting back a and b for α and β in (B.4) and substituting this back into (B.2) gives the the mean of the $N^{R[a,b]}(\mu, \sigma^2)$ distribution to be:

$$\begin{aligned} \mathrm{E}(X) = &(a-\mu)\,\Phi\left(\frac{a-\mu}{\sigma}\right) - (b-\mu)\,\Phi\left(\frac{b-\mu}{\sigma}\right) + b \\ &+ \frac{\sigma}{\sqrt{2\pi}}\left[\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) - \exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right)\right]. \end{aligned} \tag{B.5}$$

B.2.2 Variance of a $N^{R[a,b]}(\mu, \sigma^2)$ Distribution

Starting from

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2, \tag{B.6}$$

$\mathrm{E}(X)$ has already been found at (B.5). It remains to find $\mathrm{E}(X^2)$.

$$\begin{aligned} \mathrm{E}(X^2) = &\int_{-\infty}^{a} a^2 f(x)dx + \int_{a}^{b} x^2 f(x)dx + \int_{b}^{\infty} b^2 f(x)dx \\ =&a^2\Phi\left(\frac{a-\mu}{\sigma}\right) + \int_{a}^{b} \frac{x^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx \\ &+ b^2\left(1 - \Phi\left(\frac{b-\mu}{\sigma}\right)\right). \end{aligned} \tag{B.7}$$

Again using the substitution $t = \frac{x-\mu}{\sqrt{2\sigma^2}}$ on the second term leads to

$$\begin{aligned} &\int_{a}^{b} \frac{x^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx \\ =&\frac{1}{\sqrt{\pi}}\left[2\sigma^2 \int_{\alpha}^{\beta} t^2 \exp\left(-t^2\right) dt + 2\mu\sqrt{2\sigma^2} \int_{\alpha}^{\beta} t \exp\left(-t^2\right) dt + \mu^2 \int_{\alpha}^{\beta} \exp\left(-t^2\right) dt\right]. \end{aligned} \tag{B.8}$$

But

$$\int_{\alpha}^{\beta} t^2 \exp\left(-t^2\right) dt = \frac{1}{2}\alpha \exp\left(-\alpha^2\right) - \frac{1}{2}\beta \exp\left(-\beta^2\right) + \frac{\sqrt{\pi}}{4}\left[\operatorname{erf}(\beta) - \operatorname{erf}(\alpha)\right]. \tag{B.9}$$

Substituting this back into (B.8) and dealing with the other integrals gives

$$\begin{aligned}
&\int_a^b \frac{x^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx \\
=&\frac{1}{\sqrt{\pi}}\left[2\sigma^2\left(\frac{1}{2}\alpha\exp\left(-\alpha^2\right) - \frac{1}{2}\beta\exp\left(-\beta^2\right) + \frac{\sqrt{\pi}}{4}\left[\operatorname{erf}(\beta) - \operatorname{erf}(\alpha)\right]\right)\right] \\
+&\frac{1}{\sqrt{\pi}}\left[2\mu\sqrt{2\sigma^2}\left(\frac{1}{2}\left[\exp\left(-\alpha^2\right) - \exp\left(-\beta^2\right)\right]\right) + \mu^2\left(\frac{\sqrt{\pi}}{2}\left[\operatorname{erf}(\beta) - \operatorname{erf}(\alpha)\right]\right)\right].
\end{aligned} \tag{B.10}$$

Substituting back a and b for α and β in (B.10) and then in turn substituting into (B.7) leads to

$$\begin{aligned}
\mathrm{E}(X^2) = &\left(a^2 - \mu^2 - \sigma^2\right)\Phi\left(\frac{a-\mu}{\sigma}\right) + b^2 - \left(b^2 - \mu^2 - \sigma^2\right)\Phi\left(\frac{b-\mu}{\sigma}\right) \\
&+ \sqrt{\frac{2}{\pi}}\mu\sigma\left[\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) - \exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right)\right] \\
&+ \frac{\sigma^2}{\sqrt{\pi}}\left[\left(\frac{a-\mu}{\sqrt{2\sigma^2}}\right)\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) - \left(\frac{b-\mu}{\sqrt{2\sigma^2}}\right)\exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right)\right].
\end{aligned} \tag{B.11}$$

Finally, substituting equations (B.11) and (B.5) into (B.6) gives the variance of the $N^{R[a,b]}(\mu, \sigma^2)$ distribution:

$$\begin{aligned}
\operatorname{Var}(X) = &\left(a^2 - \mu^2 - \sigma^2\right)\Phi\left(\frac{a-\mu}{\sigma}\right) + b^2 - \left(b^2 - \mu^2 - \sigma^2\right)\Phi\left(\frac{b-\mu}{\sigma}\right) \\
&+ \sqrt{\frac{2}{\pi}}\mu\sigma\left[\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) - \exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right)\right] \\
&+ \frac{\sigma^2}{\sqrt{\pi}}\left[\left(\frac{a-\mu}{\sqrt{2\sigma^2}}\right)\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) - \left(\frac{b-\mu}{\sqrt{2\sigma^2}}\right)\exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right)\right] \\
&- \left[(a-\mu)\Phi\left(\frac{a-\mu}{\sigma}\right) - (b-\mu)\Phi\left(\frac{b-\mu}{\sigma}\right) + b + \frac{\sigma}{\sqrt{2\pi}}\left(\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) - \exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right)\right)\right]^2
\end{aligned} \tag{B.12}$$

B.2.3 Mean and Variance of a Right-Rectified $N^{R(-\infty,b]}(\mu, \sigma^2)$ Normal Distribution

The mean and variance of right-rectified normal distributions can be found by following an analogous procedure to that above.

The mean of a right-rectified $N^{R(-\infty,b]}(\mu, \sigma^2)$ normal distribution is

$$\mathrm{E}(X) = b - (b-\mu)\,\Phi\left(\frac{b-\mu}{\sigma}\right) - \frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right) \tag{B.13}$$

and its variance is

$$\begin{aligned}\mathrm{Var}(X) = b^2 &- \left(b^2 - \mu^2 - \sigma^2\right)\Phi\left(\frac{b-\mu}{\sigma}\right) - \sqrt{\frac{2}{\pi}}\mu\sigma\exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right) \\ &- \frac{\sigma^2}{\sqrt{\pi}}\left(\frac{b-\mu}{\sqrt{2\sigma^2}}\right)\exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right) \\ &- \left[b - (b-\mu)\,\Phi\left(\frac{b-\mu}{\sigma}\right) - \frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(b-\mu)^2}{2\sigma^2}\right)\right]^2 .\end{aligned} \tag{B.14}$$

B.2.4 Mean and Variance of a Left-Rectified $N^{R[a,\infty)}(\mu, \sigma^2)$ Normal Distribution

Finally, the mean of a left-rectified $N^{R[a,\infty)}(\mu, \sigma^2)$ normal distribution is

$$\mathrm{E}(X) = (a-\mu)\,\Phi\left(\frac{a-\mu}{\sigma}\right) + \frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) + \mu \tag{B.15}$$

and its variance is

$$\begin{aligned}\mathrm{Var}(X) = \mu^2 + \sigma^2 + \left(a^2 - \mu^2 - \sigma^2\right)\Phi\left(\frac{a-\mu}{\sigma}\right) + \sqrt{\frac{2}{\pi}}\mu\sigma\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) \\ + \frac{\sigma^2}{\sqrt{\pi}}\left(\frac{a-\mu}{\sqrt{2\sigma^2}}\right)\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) \\ - \left[(a-\mu)\,\Phi\left(\frac{a-\mu}{\sigma}\right) + \frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) + \mu\right]^2. \qquad \text{(B.16)}\end{aligned}$$

Appendix C

Results from Applying the Phylogenetic Tree Model of Chapter 4 to the HapMap Data

Tables C.1-C.22 contain the results of applying the phylogenetic tree model with Nicholson–Donnely drift to the HapMap dataset for each of the 22 autosomes.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0004	0.0026	0.0012
1	0.0030	0.0055	0.0042
2	0.0002	0.0010	0.0005
3	0.0010	0.0027	0.0018
4	0.0161	0.0238	0.0198
5	0.0081	0.0122	0.0101
6	0.0028	0.0054	0.0041
7	0.0079	0.0135	0.0105
8	0.0103	0.0148	0.0125
9	0.0017	0.0042	0.0029
10	0.0098	0.0128	0.0112
11	0.0022	0.0057	0.0038
12	0.0329	0.0401	0.0364
13	0.0169	0.0207	0.0187
14	0.1166	0.1379	0.1271
15	0.0015	0.0068	0.0040
16	0.0064	0.0107	0.0085
17	0.0164	0.0237	0.0199
18	0.0246	0.0423	0.0331
19	0.1081	0.1381	0.1227

Table C.1: Estimated Drift Parameters for Chromosome 1
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson-Donnelly model applied to Chromosome 1.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0015	0.0007
1	0.0048	0.0076	0.0062
2	0.0002	0.0011	0.0006
3	0.0005	0.0021	0.0012
4	0.0187	0.0266	0.0225
5	0.0080	0.0119	0.0099
6	0.0032	0.0059	0.0046
7	0.0107	0.0165	0.0136
8	0.0099	0.0146	0.0122
9	0.0014	0.0038	0.0026
10	0.0098	0.0127	0.0112
11	0.0042	0.0073	0.0057
12	0.0350	0.0428	0.0387
13	0.0127	0.0161	0.0144
14	0.1298	0.1526	0.1406
15	0.0003	0.0023	0.0010
16	0.0151	0.0209	0.0180
17	0.0132	0.0198	0.0164
18	0.0170	0.0306	0.0237
19	0.1101	0.1362	0.1230

Table C.2: Estimated Drift Parameters for Chromosome 2
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 2.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0004	0.0029	0.0012
1	0.0032	0.0057	0.0044
2	0.0002	0.0011	0.0006
3	0.0010	0.0029	0.0020
4	0.0152	0.0225	0.0188
5	0.0065	0.0104	0.0083
6	0.0009	0.0036	0.0023
7	0.0111	0.0171	0.0141
8	0.0092	0.0141	0.0116
9	0.0030	0.0058	0.0043
10	0.0105	0.0138	0.0121
11	0.0025	0.0057	0.0040
12	0.0323	0.0406	0.0363
13	0.0166	0.0207	0.0186
14	0.1015	0.1212	0.1110
15	0.0003	0.0036	0.0014
16	0.0100	0.0155	0.0127
17	0.0174	0.0249	0.0209
18	0.0201	0.0374	0.0277
19	0.1120	0.1444	0.1281

Table C.3: Estimated Drift Parameters for Chromosome 3

The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 3.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0017	0.0007
1	0.0023	0.0048	0.0035
2	0.0001	0.0010	0.0004
3	0.0011	0.0030	0.0021
4	0.0195	0.0282	0.0237
5	0.0092	0.0131	0.0112
6	0.0041	0.0067	0.0053
7	0.0172	0.0243	0.0205
8	0.0062	0.0111	0.0084
9	0.0014	0.0039	0.0027
10	0.0084	0.0114	0.0099
11	0.0006	0.0034	0.0018
12	0.0330	0.0419	0.0373
13	0.0100	0.0135	0.0117
14	0.1057	0.1272	0.1160
15	0.0005	0.0064	0.0026
16	0.0152	0.0218	0.0185
17	0.0121	0.0200	0.0158
18	0.0225	0.0420	0.0316
19	0.1186	0.1546	0.1364

Table C.4: Estimated Drift Parameters for Chromosome 4
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 4.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0017	0.0008
1	0.0026	0.0054	0.0039
2	0.0002	0.0009	0.0004
3	0.0008	0.0027	0.0017
4	0.0138	0.0210	0.0173
5	0.0080	0.0127	0.0101
6	0.0030	0.0057	0.0044
7	0.0153	0.0225	0.0187
8	0.0084	0.0133	0.0107
9	0.0017	0.0045	0.0030
10	0.0086	0.0116	0.0100
11	0.0005	0.0048	0.0029
12	0.0298	0.0373	0.0334
13	0.0118	0.0155	0.0136
14	0.1192	0.1441	0.1313
15	0.0004	0.0047	0.0020
16	0.0115	0.0173	0.0144
17	0.0118	0.0188	0.0151
18	0.0206	0.0389	0.0297
19	0.1037	0.1363	0.1193

Table C.5: Estimated Drift Parameters for Chromosome 5
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 5.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0017	0.0007
1	0.0027	0.0055	0.0041
2	0.0001	0.0009	0.0004
3	0.0017	0.0037	0.0027
4	0.0195	0.0278	0.0235
5	0.0080	0.0125	0.0101
6	0.0030	0.0058	0.0043
7	0.0133	0.0205	0.0169
8	0.0108	0.0161	0.0134
9	0.0010	0.0038	0.0023
10	0.0093	0.0127	0.0109
11	0.0032	0.0070	0.0051
12	0.0351	0.0442	0.0395
13	0.0142	0.0183	0.0162
14	0.1096	0.1332	0.1209
15	0.0003	0.0038	0.0015
16	0.0067	0.0123	0.0094
17	0.0108	0.0179	0.0142
18	0.0384	0.0605	0.0490
19	0.0780	0.1084	0.0925

Table C.6: Estimated Drift Parameters for Chromosome 6
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 6.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0001	0.0012	0.0006
1	0.0019	0.0047	0.0033
2	0.0002	0.0014	0.0006
3	0.0004	0.0019	0.0011
4	0.0181	0.0276	0.0225
5	0.0073	0.0122	0.0097
6	0.0039	0.0069	0.0054
7	0.0078	0.0144	0.0112
8	0.0038	0.0084	0.0061
9	0.0011	0.0038	0.0024
10	0.0094	0.0129	0.0111
11	0.0027	0.0067	0.0046
12	0.0274	0.0352	0.0313
13	0.0103	0.0140	0.0121
14	0.1118	0.1367	0.1239
15	0.0019	0.0087	0.0048
16	0.0171	0.0235	0.0202
17	0.0109	0.0189	0.0147
18	0.0279	0.0493	0.0379
19	0.0894	0.1234	0.1058

Table C.7: Estimated Drift Parameters for Chromosome 7

The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 7.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0016	0.0007
1	0.0042	0.0075	0.0058
2	0.0001	0.0012	0.0006
3	0.0006	0.0025	0.0015
4	0.0207	0.0304	0.0254
5	0.0051	0.0098	0.0072
6	0.0018	0.0046	0.0031
7	0.0130	0.0214	0.0172
8	0.0068	0.0121	0.0093
9	0.0008	0.0038	0.0022
10	0.0105	0.0141	0.0122
11	0.0026	0.0073	0.0049
12	0.0330	0.0426	0.0375
13	0.0135	0.0179	0.0156
14	0.0996	0.1231	0.1108
15	0.0041	0.0121	0.0080
16	0.0110	0.0169	0.0140
17	0.0068	0.0145	0.0104
18	0.0253	0.0489	0.0365
19	0.1035	0.1428	0.1218

Table C.8: Estimated Drift Parameters for Chromosome 8
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 8.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0023	0.0009
1	0.0022	0.0057	0.0039
2	0.0001	0.0013	0.0006
3	0.0010	0.0034	0.0022
4	0.0110	0.0197	0.0152
5	0.0071	0.0131	0.0099
6	0.0019	0.0053	0.0036
7	0.0124	0.0216	0.0169
8	0.0100	0.0168	0.0133
9	0.0021	0.0057	0.0039
10	0.0106	0.0152	0.0128
11	0.0023	0.0075	0.0049
12	0.0299	0.0399	0.0347
13	0.0146	0.0200	0.0172
14	0.1088	0.1373	0.1226
15	0.0010	0.0118	0.0057
16	0.0050	0.0115	0.0082
17	0.0150	0.0256	0.0202
18	0.0259	0.0510	0.0377
19	0.0914	0.1312	0.1105

Table C.9: Estimated Drift Parameters for Chromosome 9
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 9.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0014	0.0006
1	0.0028	0.0059	0.0042
2	0.0001	0.0010	0.0005
3	0.0005	0.0021	0.0012
4	0.0149	0.0241	0.0193
5	0.0089	0.0143	0.0115
6	0.0033	0.0067	0.0049
7	0.0068	0.0140	0.0103
8	0.0081	0.0144	0.0111
9	0.0005	0.0034	0.0020
10	0.0081	0.0117	0.0098
11	0.0011	0.0053	0.0029
12	0.0311	0.0406	0.0357
13	0.0118	0.0161	0.0139
14	0.1000	0.1250	0.1120
15	0.0066	0.0155	0.0109
16	0.0106	0.0173	0.0139
17	0.0159	0.0259	0.0206
18	0.0369	0.0644	0.0498
19	0.0695	0.1023	0.0856

Table C.10: Estimated Drift Parameters for Chromosome 10
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 10.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0004	0.0031	0.0013
1	0.0031	0.0065	0.0046
2	0.0001	0.0013	0.0006
3	0.0006	0.0027	0.0016
4	0.0201	0.0297	0.0247
5	0.0065	0.0117	0.0090
6	0.0027	0.0061	0.0045
7	0.0150	0.0237	0.0191
8	0.0090	0.0149	0.0118
9	0.0007	0.0045	0.0028
10	0.0096	0.0136	0.0115
11	0.0058	0.0105	0.0080
12	0.0266	0.0356	0.0309
13	0.0124	0.0170	0.0147
14	0.0969	0.1215	0.1088
15	0.0006	0.0059	0.0025
16	0.0078	0.0146	0.0112
17	0.0108	0.0191	0.0148
18	0.0134	0.03389672	0.0237
19	0.1056	0.1434871	0.1233

Table C.11: Estimated Drift Parameters for Chromosome 11
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 11.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0016	0.0007
1	0.0029	0.0062	0.0044
2	0.0001	0.0009	0.0004
3	0.0004	0.0023	0.0012
4	0.0123	0.0216	0.0170
5	0.0074	0.0127	0.0100
6	0.0018	0.0052	0.0034
7	0.0159	0.0242	0.0199
8	0.0091	0.0148	0.0119
9	0.0025	0.0062	0.0044
10	0.0104	0.0144	0.0123
11	0.0020	0.0063	0.0039
12	0.0305	0.0404	0.0352
13	0.0156	0.0205	0.0180
14	0.1268	0.1583	0.1418
15	0.0002	0.0045	0.0014
16	0.0086	0.0149	0.0117
17	0.0147	0.0237	0.0190
18	0.0168	0.0372	0.0258
19	0.1058	0.1460	0.1252

Table C.12: Estimated Drift Parameters for Chromosome 12
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 12.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0033	0.0013
1	0.0017	0.0052	0.0034
2	0.0001	0.0014	0.0006
3	0.0019	0.0046	0.0032
4	0.0174	0.0291	0.0229
5	0.0055	0.0116	0.0084
6	0.0024	0.0059	0.0040
7	0.0069	0.0159	0.0113
8	0.0034	0.0099	0.0064
9	0.0022	0.0058	0.0038
10	0.0085	0.0126	0.0105
11	0.0009	0.0071	0.0043
12	0.0322	0.0436	0.0375
13	0.0074	0.0119	0.0096
14	0.1037	0.1349	0.1185
15	0.0005	0.0098	0.0040
16	0.0140	0.0231	0.0184
17	0.0137	0.0251	0.0191
18	0.0306	0.0618	0.0448
19	0.0805	0.1247	0.1015

Table C.13: Estimated Drift Parameters for Chromosome 13
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 13.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0025	0.0009
1	0.0053	0.0092	0.0071
2	0.0002	0.0015	0.0007
3	0.0004	0.0031	0.0016
4	0.0142	0.0253	0.0195
5	0.0084	0.0151	0.0117
6	0.0020	0.0055	0.0036
7	0.0165	0.0277	0.0218
8	0.0081	0.0159	0.0118
9	0.0002	0.0027	0.0011
10	0.0084	0.0129	0.0105
11	0.0021	0.0075	0.0047
12	0.0354	0.0488	0.0417
13	0.0118	0.0173	0.0145
14	0.1011	0.1323	0.1159
15	0.0004	0.0065	0.0022
16	0.0091	0.0172	0.0131
17	0.0134	0.0243	0.0185
18	0.0191	0.0487	0.0328
19	0.0848	0.1322	0.1076

Table C.14: Estimated Drift Parameters for Chromosome 14
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 14.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0001	0.0015	0.0006
1	0.0053	0.0092	0.0072
2	0.0001	0.0011	0.0004
3	0.0002	0.0023	0.0009
4	0.0172	0.0300	0.0233
5	0.0063	0.0131	0.0096
6	0.0027	0.0074	0.0050
7	0.0162	0.0292	0.0223
8	0.0040	0.0114	0.0074
9	0.0002	0.0025	0.0010
10	0.0103	0.0157	0.0129
11	0.0026	0.0080	0.0051
12	0.0349	0.0492	0.0417
13	0.0105	0.0162	0.0133
14	0.1121	0.1488	0.1292
15	0.0046	0.0166	0.0104
16	0.0149	0.0248	0.0197
17	0.0051	0.0165	0.0107
18	0.0168	0.0469	0.0304
19	0.1077	0.1651	0.1349

Table C.15: Estimated Drift Parameters for Chromosome 15
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 15.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0025	0.0009
1	0.0032	0.0070	0.0050
2	0.0001	0.0014	0.0006
3	0.0007	0.0036	0.0021
4	0.0151	0.0290	0.0216
5	0.0083	0.0156	0.0118
6	0.0031	0.0080	0.0054
7	0.0134	0.0256	0.0191
8	0.0067	0.0146	0.0105
9	0.0002	0.0029	0.0012
10	0.0095	0.0152	0.0122
11	0.0008	0.0060	0.0030
12	0.0355	0.0508	0.0428
13	0.0151	0.0221	0.0183
14	0.1026	0.1399	0.1200
15	0.0013	0.0096	0.0045
16	0.0120	0.0221	0.0169
17	0.0176	0.0310	0.0239
18	0.0074	0.0282	0.0170
19	0.1278	0.1806	0.1534

Table C.16: Estimated Drift Parameters for Chromosome 16
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 16.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0036	0.0013
1	0.0024	0.0071	0.0047
2	0.0001	0.0011	0.0005
3	0.0004	0.0025	0.0013
4	0.0070	0.0183	0.0126
5	0.0043	0.0108	0.0074
6	0.0027	0.0068	0.0046
7	0.0071	0.0174	0.0119
8	0.0115	0.0203	0.0156
9	0.0026	0.00737	0.0048
10	0.0072	0.0120	0.0095
11	0.0034	0.0097	0.0064
12	0.0303	0.0438	0.0369
13	0.0171	0.0242	0.0205
14	0.1037	0.1384	0.1202
15	0.0055	0.0188	0.0118
16	0.0014	0.0108	0.0063
17	0.0157	0.0301	0.0224
18	0.0149	0.0423	0.0277
19	0.1186	0.1743	0.1452

Table C.17: Estimated Drift Parameters for Chromosome 17
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 17.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0001	0.0023	0.0009
1	0.0013	0.0050	0.0030
2	0.0001	0.0011	0.0005
3	0.0005	0.0033	0.0017
4	0.0110	0.0225	0.0162
5	0.0073	0.0141	0.0105
6	0.0037	0.0080	0.0057
7	0.0072	0.0167	0.0119
8	0.0098	0.0183	0.0137
9	0.0029	0.0072	0.0050
10	0.0070	0.0116	0.0092
11	0.0009	0.0063	0.0034
12	0.0274	0.0388	0.0328
13	0.0105	0.0162	0.0132
14	0.0965	0.1290	0.1118
15	0.0009	0.0085	0.0039
16	0.0066	0.0153	0.0109
17	0.0123	0.0238	0.0179
18	0.0113	0.0347	0.0215
19	0.1026	0.1509	0.1255

Table C.18: Estimated Drift Parameters for Chromosome 18
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 18.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0003	0.0038	0.0016
1	0.0016	0.0056	0.0034
2	0.0002	0.0019	0.0008
3	0.0002	0.0024	0.0009
4	0.0175	0.0337	0.0251
5	0.0071	0.0145	0.0106
6	0.0007	0.0051	0.0028
7	0.0102	0.0229	0.0162
8	0.0058	0.0143	0.0098
9	0.0004	0.0048	0.0023
10	0.0065	0.0117	0.0090
11	0.0009	0.0069	0.0037
12	0.0261	0.0408	0.0329
13	0.0106	0.0173	0.0136
14	0.0875	0.1243	0.1047
15	0.0004	0.0064	0.0024
16	0.0065	0.0162	0.0113
17	0.0120	0.0257	0.0184
18	0.0161	0.0463	0.0288
19	0.0771	0.1254	0.1003

Table C.19: Estimated Drift Parameters for Chromosome 19
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 19.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0025	0.0010
1	0.0028	0.0076	0.0051
2	0.0002	0.0024	0.0010
3	0.0016	0.0052	0.0032
4	0.0192	0.0344	0.0263
5	0.0031	0.0101	0.0064
6	0.0040	0.0089	0.0063
7	0.0124	0.0245	0.0182
8	0.0063	0.0150	0.0103
9	0.0010	0.0053	0.0028
10	0.0095	0.0154	0.0122
11	0.0028	0.0095	0.0059
12	0.0274	0.0410	0.0338
13	0.0089	0.0150	0.0119
14	0.1069	0.1490	0.1261
15	0.0003	0.0088	0.0031
16	0.0135	0.0238	0.0185
17	0.0131	0.0278	0.0199
18	0.0167	0.0557	0.0342
19	0.0951	0.1621	0.1269

Table C.20: Estimated Drift Parameters for Chromosome 20
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 20.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0002	0.0059	0.0022
1	0.0017	0.0081	0.0047
2	0.0002	0.0025	0.0010
3	0.0003	0.0036	0.0017
4	0.0126	0.0298	0.0205
5	0.0086	0.0198	0.0140
6	0.0013	0.0080	0.0045
7	0.0196	0.0386	0.0284
8	0.0108	0.0224	0.0165
9	0.0006	0.0061	0.0030
10	0.0093	0.0175	0.0131
11	0.0010	0.0100	0.0047
12	0.0232	0.0411	0.0313
13	0.0126	0.0227	0.0173
14	0.0916	0.1414	0.1146
15	0.0002	0.0097	0.0032
16	0.0005	0.0091	0.0035
17	0.0054	0.0189	0.0118
18	0.0224	0.0743	0.0452
19	0.0632	0.1342	0.0963

Table C.21: Estimated Drift Parameters for Chromosome 21
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 21.

Genetic Drift Along Edge	Central 95% Credible Interval Bounds for Genetic Drift, c		Point Estimate
	low	high	mean
0	0.0000	0.0019	0.0004
1	0.0027	0.0061	0.0044
2	0.0001	0.0009	0.0003
3	0.0003	0.0023	0.0012
4	0.0130	0.0218	0.0171
5	0.0078	0.0130	0.0103
6	0.0019	0.0051	0.0034
7	0.0159	0.0244	0.0201
8	0.0094	0.0152	0.0122
9	0.0027	0.0062	0.0044
10	0.0103	0.0144	0.0123
11	0.0013	0.0057	0.0035
12	0.0307	0.0405	0.0354
13	0.0157	0.0206	0.0180
14	0.1264	0.1576	0.1413
15	0.0002	0.0044	0.0017
16	0.0084	0.0148	0.0115
17	0.0142	0.0233	0.0186
18	0.0149	0.0363	0.0248
19	0.1064	0.1478	0.1263

Table C.22: Estimated Drift Parameters for Chromosome 22
The estimated value and 95% central credible intervals for c for each of the 20 periods of drift for the 11 subpopulations in the HapMap dataset estimated by the bifurcating Nicholson–Donnelly model applied to Chromosome 22.

Appendix D

Results from Applying the Admixture Models of Chapter 5 to the HapMap Data

Tables D.1-D.31 contain the main parameter estimates and post predictive checking tables for the models discussed in chapter 5.

Table D.1: Parameter Estimates for the Model in Figure 5.6 at 100,000 iterations.

	Bounds on 95% Credible Interval		
Parameter	lower	upper	Median
c_0	0.0002	0.0016	0.0008
c_1	0.0047	0.0073	0.0060
c_2	0.0002	0.0011	0.0006
c_3	0.0005	0.0019	0.0012
c_4	0.0182	0.0257	0.0220
c_5	0.0081	0.0121	0.0100
c_6	0.0002	0.0012	0.0006
c_7	0.0113	0.0168	0.0140
c_8	0.0007	0.0035	0.0020
c_9	0.0015	0.0041	0.0028
c_{10}	0.0023	0.0047	0.0035
w_{11}	0.1891	0.2107	0.1998
c_{12}	0.0003	0.0125	0.0039
c_{13}	0.0002	0.0026	0.0013
c_{14}	0.0038	0.0072	0.0055
c_{15}	0.1298	0.1527	0.1410
c_{16}	0.0137	0.0202	0.0168
c_{17}	0.0021	0.0143	0.0082
c_{18}	0.0223	0.0367	0.0296
c_{19}	0.0002	0.0016	0.0008
c_{20}	0.0916	0.1168	0.1039
c_{21}	0.0110	0.0144	0.0127
c_{22}	0.0364	0.0434	0.0398
c_{23}	0.0356	0.0529	0.0441

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8151	0.8409	0.8151	0.5838	0.7458	0.5477	0.8723	0.2477	0.4448	0.7472
CEU	0.8151	X	0.7919	**0.9953**	0.0692	0.9388	**0.9959**	0.6920	0.4374	0.5982	**0.9996**
CHB	0.8409	0.7919	X	0.1220	0.6817	0.2191	0.8640	**0.0001**	0.3084	0.9038	0.9636
CHD	0.8151	**0.9953**	0.1220	X	0.9060	0.6524	0.7535	**0.0004**	0.2888	**0.9957**	0.9419
GIH	0.5838	0.0692	0.6817	0.9060	X	0.8672	0.7943	**0.9949**	0.0560	**0.0070**	0.9481
JPT	0.7458	0.9388	0.2191	0.6524	0.8672	X	0.8161	**0.0001**	0.2789	0.9629	0.9655
LWK	0.5477	**0.9959**	0.8640	0.7535	0.7943	0.8161	X	0.9671	0.0827	0.7928	0.2475
MEX	0.8723	0.6920	**0.0001**	**0.0004**	**0.9949**	**0.0001**	0.9671	X	0.5990	0.9602	**0.9969**
MKK	0.2477	0.4374	0.3084	0.2888	0.0560	0.2789	0.0827	0.5990	X	**0.0005**	0.8265
TSI	0.4448	0.5982	0.9038	**0.9957**	**0.0070**	0.9629	0.7928	0.9602	**0.0005**	X	0.9248
YRI	0.7472	**0.9996**	0.9636	0.9419	0.9481	0.9655	0.2475	**0.9969**	0.8265	0.9248	X

Table D.2: Predictive p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.6.

Table D.3: Parameter Estimates for the Model in Figure 5.7 at 102,000 iterations.

	Bounds of 95% Credible Interval		
Parameter	lower	upper	Median
c_0	0.0002	0.0016	0.0008
c_1	0.0051	0.0077	0.0064
c_2	0.0001	0.0011	0.0005
c_3	0.0004	0.0019	0.0012
c_4	0.0083	0.0146	0.0113
c_5	0.0082	0.0123	0.0103
c_6	0.0002	0.0013	0.0006
c_7	0.0001	0.0011	0.0005
c_8	0.0006	0.0032	0.0018
c_9	0.0012	0.0035	0.0023
c_{10}	0.0023	0.0047	0.0035
c_{11}	0.0035	0.0069	0.0052
c_{12}	0.0418	0.0671	0.0541
w_{13}	0.3374	0.4136	0.3784
c_{14}	0.1442	0.2743	0.2019
c_{15}	0.0002	0.0014	0.0006
c_{16}	0.0801	0.1117	0.0953
c_{17}	0.0224	0.0317	0.0269
c_{18}	0.0052	0.0160	0.0101
w_{19}	0.1921	0.2148	0.2033
c_{20}	0.0003	0.0121	0.0041
c_{21}	0.0002	0.0026	0.0013
c_{22}	0.0048	0.0177	0.0113
c_{23}	0.0012	0.0106	0.0056
c_{24}	0.0876	0.1131	0.0998
c_{25}	0.0111	0.0144	0.0127
c_{26}	0.0367	0.0436	0.0402
c_{27}	0.0362	0.0541	0.0448

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8260	0.7877	0.7629	0.7218	0.6614	0.5491	0.3519	0.2696	0.5820	0.7415
CEU	0.8620	X	0.6456	**0.9853**	0.2836	0.8531	**0.9974**	0.0368	0.5637	0.6068	**0.9997**
CHB	0.7877	0.6456	X	0.1575	0.3699	0.2716	0.8419	0.3375	0.2154	0.8239	0.9585
CHD	0.7629	**0.9853**	0.1575	X	0.7040	0.7026	0.7258	0.6261	0.2052	**0.9887**	0.9359
GIH	0.7218	0.2836	0.3699	0.7040	X	0.6051	0.8510	0.8950	0.1226	0.1140	0.9676
JPT	0.6614	0.8531	0.2716	0.7026	0.6051	X	0.7737	0.3819	0.1730	0.9123	0.9559
LWK	0.5491	**0.9974**	0.8419	0.7258	0.8510	0.7737	X	0.5758	0.0736	0.9041	0.2404
MEX	0.3519	0.0368	0.3375	0.6261	0.8950	0.3819	0.5758	X	0.0399	0.3635	0.8452
MKK	0.2696	0.5637	0.2154	0.2052	0.1226	0.1730	0.0736	0.0399	X	**0.0056**	0.8084
TSI	0.5820	0.6068	0.8239	**0.9887**	0.1140	0.9123	0.9041	0.3635	0.0056	X	0.9737
YRI	0.7415	**0.9997**	0.9585	0.9359	0.9676	0.9559	0.2404	0.8452	0.8084	0.9737	X

Table D.4: Predictive p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.7

Table D.5: Parameter Estimates for the Model in Figure 5.8 at 87,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0001	0.0015	0.0007
c_1	0.0046	0.0073	0.0059
c_2	0.0000	0.0009	0.0004
c_3	0.0005	0.0020	0.0012
c_4	0.0081	0.0123	0.0101
c_5	0.0000	0.0010	0.0004
c_6	0.0001	0.0010	0.0004
c_7	0.0007	0.0035	0.0020
c_8	0.0015	0.0041	0.0028
c_9	0.0023	0.0049	0.0036
w_{10}	0.3920	0.4843	0.4349
w_{11}	0.1907	0.2144	0.2023
c_{12}	0.0457	0.1166	0.0780
c_{13}	0.0000	0.0013	0.0004
c_{14}	0.0001	0.0122	0.0043
c_{15}	0.0000	0.0027	0.0011
c_{16}	0.0034	0.0071	0.0053
c_{17}	0.0635	0.0968	0.0794
c_{18}	0.0739	0.1082	0.0904
c_{19}	0.0046	0.0191	0.0118
c_{20}	0.0046	0.0221	0.0125
c_{21}	0.0085	0.0378	0.0241
c_{22}	0.0109	0.0145	0.0127
c_{23}	0.0362	0.0433	0.0398
c_{24}	0.0366	0.0561	0.0458
c_{25}	0.0703	0.0970	0.0833

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

Table D.6: Parameter Estimates for the Model in Figure 5.9 at 102,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0016	0.0008
c_1	0.0049	0.0075	0.0062
c_2	0.0002	0.0011	0.0005
c_3	0.0005	0.0019	0.0012
c_4	0.0153	0.0225	0.0189
c_5	0.0082	0.0121	0.0101
c_6	0.0002	0.0012	0.0006
c_7	0.0001	0.0017	0.0007
c_8	0.0004	0.0033	0.0018
c_9	0.0014	0.0038	0.0026
c_{10}	0.0023	0.0047	0.0035
c_{11}	0.0038	0.0072	0.0055
c_{12}	0.1353	0.1594	0.1471
c_{13}	0.0122	0.0190	0.0156
w_{14}	0.5610	0.6812	0.6237
c_{15}	0.0002	0.0024	0.0009
c_{16}	0.1225	0.2917	0.1938
c_{17}	0.0002	0.0041	0.0016
c_{18}	0.0022	0.0146	0.0086
w_{19}	0.1894	0.2117	0.2008
c_{20}	0.0003	0.0116	0.0037
c_{21}	0.0002	0.0026	0.0013
c_{22}	0.0165	0.0317	0.0241
c_{23}	0.0002	0.0048	0.0015
c_{24}	0.0896	0.1154	0.1021
c_{25}	0.0110	0.0143	0.0127
c_{26}	0.0367	0.0437	0.0401
c_{27}	0.0349	0.0531	0.0438

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8204	0.8418	0.8205	0.6524	0.7501	0.5454	0.6208	0.2634	0.4980	0.7361
CEU	0.8204	X	0.7829	**0.9942**	0.1558	0.9347	0.9967	0.1755	0.4827	0.5956	**0.9996**
CHB	0.8418	0.7829	X	0.1328	0.6516	0.2228	0.8566	**0**	0.3057	0.9195	0.9600
CHD	0.8205	**0.9942**	0.1328	X	0.8934	0.6489	0.7466	**0.0002**	0.2865	**0.9970**	0.9382
GIH	0.6524	0.1558	0.6516	0.8934	X	0.8476	0.8196	0.9033	0.0849	0.0368	0.9552
JPT	0.7501	0.9347	0.2228	0.6489	0.8476	X	0.8091	**0**	0.2694	0.9715	0.9629
LWK	0.5454	**0.9967**	0.8566	0.7466	0.8196	0.8091	X	0.8399	0.0786	0.8337	0.2389
MEX	0.6208	0.1755	**0**	**0.0002**	0.9033	**0**	0.8399	X	0.2126	0.6460	0.9656
MKK	0.2634	0.4827	0.3057	0.2865	0.0849	0.2694	0.0786	0.2126	X	**0.0014**	0.8194
TSI	0.4980	0.5956	0.9195	**0.9970**	0.0368	0.9715	0.8337	0.6460	**0.0014**	X	0.9459
YRI	0.7361	**0.9996**	0.9600	0.9382	0.9552	0.9629	0.2389	0.9656	0.8194	0.9459	X

Table D.7: Predictive p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.9

Table D.8: Parameter Estimates for the Model in Figure 5.10 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0016	0.0008
c_1	0.0049	0.0074	0.0061
c_2	0.0002	0.0011	0.0005
c_3	0.0005	0.0019	0.0012
c_4	0.0198	0.0267	0.0232
c_5	0.0080	0.0121	0.0101
c_6	0.0002	0.0012	0.0006
c_7	0.0002	0.0012	0.0005
c_8	0.0005	0.0035	0.0020
c_9	0.0014	0.0038	0.0026
c_{10}	0.0023	0.0048	0.0035
w_{11}	0.3621	0.4501	0.4019
c_{12}	0.0960	0.1982	0.1462
c_{13}	0.0001	0.0015	0.0007
w_{14}	0.1898	0.2116	0.2007
c_{15}	0.0002	0.0127	0.0039
c_{16}	0.0003	0.0029	0.0014
c_{17}	0.0036	0.0071	0.0053
c_{18}	0.0030	0.0146	0.0083
c_{19}	0.0553	0.0822	0.0682
c_{20}	0.0715	0.1004	0.0856
c_{21}	0.0017	0.0160	0.0093
c_{22}	0.0152	0.0270	0.0213
c_{23}	0.0001	0.0008	0.0004
c_{24}	0.0110	0.0143	0.0127
c_{25}	0.0365	0.0435	0.0399
c_{26}	0.0878	0.1132	0.1003
c_{27}	0.0360	0.0538	0.0446

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8165	0.8124	0.7925	0.8820	0.7085	0.5332	0.2530	0.2357	0.4836	0.7243
CEU	0.8165	X	0.7231	**0.9912**	0.4472	0.9029	**0.9966**	0.1472	0.4585	0.6118	**0.9996**
CHB	0.8124	0.7231	X	0.1427	0.0492	0.2561	0.8594	0.3294	0.2880	0.8674	0.9627
CHD	0.7925	**0.9912**	0.1427	X	0.2242	0.6918	0.7507	0.6212	0.2711	**0.9933**	0.9410
GIH	0.8820	0.4472	0.0492	0.2242	X	0.1759	0.9632	0.8006	0.4084	0.2007	**0.9949**
JPT	0.7085	0.9029	0.2561	0.6918	0.1759	X	0.8051	0.3928	0.2508	0.9433	0.9632
LWK	0.5332	**0.9966**	0.8594	0.7507	0.9632	0.8051	X	0.4123	0.0809	0.8308	0.2457
MEX	0.2530	0.1472	0.3294	0.6212	0.8006	0.3928	0.4123	X	**0.0129**	0.6130	0.7198
MKK	0.2357	0.4585	0.2880	0.2711	0.4084	0.2508	0.0809	**0.0129**	X	**0.0011**	0.8182
TSI	0.4836	0.6118	0.8674	**0.9933**	0.2007	0.9433	0.8308	0.6130	**0.0011**	X	0.9430
YRI	0.7243	**0.9996**	0.9627	0.9410	**0.9949**	0.9632	0.2457	0.7198	0.8182	0.9430	X

Table D.9: Predictive p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.10

Table D.10: Parameter Estimates Table of the Model in Figure 5.11 after 72,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0017	0.0008
c_1	0.0052	0.0078	0.0065
c_2	0.0002	0.0010	0.0005
c_3	0.0004	0.0019	0.0012
c_4	0.0080	0.0141	0.0110
c_5	0.0082	0.0122	0.0102
c_6	0.0001	0.0011	0.0005
c_7	0.0001	0.0011	0.0005
c_8	0.0060	0.0132	0.0097
c_9	0.0002	0.0020	0.0009
c_{10}	0.0022	0.0046	0.0033
w_{11}	0.1939	0.2161	0.2042
c_{12}	0.0002	0.0115	0.0037
c_{13}	0.0002	0.0028	0.0013
w_{14}	0.1905	0.2311	0.2109
c_{15}	0.0333	0.1083	0.0693
c_{16}	0.0017	0.0129	0.0077
w_{17}	0.3443	0.4077	0.3772
c_{18}	0.1497	0.2668	0.2038
c_{19}	0.0001	0.0014	0.0006
c_{20}	0.0036	0.0069	0.0052
c_{21}	0.0401	0.0651	0.0524
c_{22}	0.0819	0.1133	0.0973
c_{23}	0.0216	0.0310	0.0262
c_{24}	0.0005	0.0027	0.0015
c_{25}	0.0040	0.0150	0.0096
c_{26}	0.0046	0.0163	0.0106
c_{27}	0.0030	0.0128	0.0074
c_{28}	0.1066	0.1412	0.1235
c_{29}	0.0110	0.0144	0.0127
c_{30}	0.0037	0.0117	0.0078
c_{31}	0.0717	0.1012	0.0859

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8253	0.7928	0.7719	0.7022	0.6718	0.3403	0.2920	0.6895	0.5788	0.7145
CEU	0.8253	X	0.6658	**0.9858**	0.2938	0.8663	0.9452	0.0549	0.9688	0.4534	**0.9992**
CHB	0.7928	0.6658	X	0.1619	0.3664	0.2765	0.6612	0.3334	0.4554	0.8387	0.9490
CHD	0.7719	**0.9858**	0.1619	X	0.6976	0.7038	0.5047	0.6213	0.4350	**0.9905**	0.9222
GIH	0.7022	0.2938	0.3664	0.6976	X	0.6043	0.5261	0.8896	0.4068	0.1217	0.9422
JPT	0.6718	0.8663	0.2765	0.7038	0.6043	X	0.5786	0.3820	0.3810	0.9255	0.9463
LWK	0.3403	0.9452	0.6612	0.5047	0.5261	0.5786	X	0.2123	0.2056	0.5702	0.2888
MEX	0.2920	0.0549	0.3334	0.6213	0.8896	0.3820	0.2123	X	0.1315	0.4280	0.7395
MKK	0.6895	0.9688	0.4554	0.4350	0.4068	0.3810	0.2056	0.1315	X	0.2662	0.9173
TSI	0.5788	0.4534	0.8387	**0.9905**	0.1217	0.9255	0.5702	0.4280	0.2662	X	0.9585
YRI	0.7145	**0.9992**	0.9490	0.9222	0.9422	0.9463	0.2888	0.7395	0.9173	0.9585	X

Table D.11: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.11

Table D.12: Parameter Estimates Table of the Model in Figure 5.12 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0013	0.0006
c_1	0.0053	0.0079	0.0066
c_2	0.0002	0.0011	0.0006
c_3	0.0004	0.0019	0.0012
c_4	0.0081	0.0144	0.0111
c_5	0.0082	0.0123	0.0102
c_6	0.0033	0.0058	0.0046
c_7	0.0002	0.0012	0.0005
c_8	0.0096	0.0142	0.0119
c_9	0.0010	0.0034	0.0022
c_{10}	0.0098	0.0126	0.0112
w_{11}	0.3365	0.3994	0.3694
c_{12}	0.1593	0.2813	0.2148
c_{13}	0.0002	0.0013	0.0006
c_{14}	0.0035	0.0069	0.0052
c_{15}	0.0396	0.0640	0.0518
c_{16}	0.0829	0.1142	0.0981
c_{17}	0.0216	0.0305	0.0261
c_{18}	0.0172	0.0249	0.0210
c_{19}	0.0026	0.0124	0.0073
c_{20}	0.1062	0.1341	0.1200
c_{21}	0.0128	0.0161	0.0144
c_{22}	0.0152	0.0211	0.0182
c_{23}	0.0158	0.0308	0.0227

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	**0**	**0.0036**	**0.0032**	**0**	**0.0020**	0.8601	**0**	0.9595	**0**	**0.0001**
CEU	**0**	X	0.6321	**0.9844**	0.2652	0.8506	**0.9987**	0.0466	0.8241	0.5840	**1**
CHB	**0.0036**	0.6321	X	0.1450	0.3789	0.2623	0.8377	0.3188	0.3304	0.8398	**0.9979**
CHD	**0.0032**	**0.9844**	0.1450	X	0.7137	0.7083	0.7232	0.6141	0.3190	**0.9913**	**0.9955**
GIH	**0**	0.2652	0.3789	0.7137	X	0.6206	0.8985	0.8884	0.2963	0.1295	**0.9995**
JPT	**0.0020**	0.8506	0.2623	0.7083	0.6206	X	0.7695	0.3725	0.2820	0.9253	**0.9970**
LWK	0.8601	**0.9987**	0.8377	0.7232	0.8985	0.7695	X	0.6226	**0**	0.9445	0.5701
MEX	**0**	0.0466	0.3188	0.6141	0.8884	0.3725	0.6226	X	0.0951	0.4274	0.9284
MKK	0.9595	0.8241	0.3304	0.3190	0.2963	0.2820	**0**	0.0951	X	0.0416	0.9415
TSI	**0**	0.5840	0.8398	**0.9913**	0.1295	0.9253	0.9445	0.4274	0.0416	X	**0.9998**
YRI	**0.0001**	**1**	**0.9979**	**0.9955**	**0.9995**	**0.9970**	0.5701	0.9284	0.9415	**0.9998**	X

Table D.13: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.12

Table D.14: Parameter Estimates Table of the Model in Figure 5.13 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0012	0.0006
c_1	0.0051	0.0078	0.0064
c_2	0.0002	0.0011	0.0006
c_3	0.0005	0.0019	0.0012
c_4	0.0239	0.0307	0.0272
c_5	0.0078	0.0118	0.0098
c_6	0.0034	0.0058	0.0046
c_7	0.0003	0.0039	0.0019
c_8	0.0100	0.0148	0.0123
c_9	0.0011	0.0036	0.0023
c_{10}	0.0097	0.0126	0.0112
c_{11}	0.0041	0.0075	0.0058
c_{12}	0.1340	0.1564	0.1449
c_{13}	0.0068	0.0129	0.0098
c_{14}	0.0260	0.0329	0.0293
c_{15}	0.0154	0.0222	0.0187
c_{16}	0.1073	0.1346	0.1205
c_{17}	0.0127	0.0161	0.0143
c_{18}	0.0149	0.0207	0.0177
c_{19}	0.0172	0.0316	0.0241

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	**0**	**0.0072**	**0.0068**	**0**	**0.0062**	0.8683	**0.0036**	0.9619	**0**	**0.0001**
CEU	**0**	X	0.7587	**0.9939**	0.8772	0.9366	**0.9971**	**0.0088**	0.6779	0.6241	**1**
CHB	**0.0072**	0.7587	X	0.1287	**0.0104**	0.2092	0.8592	0.4500	0.4040	0.9344	**0.9982**
CHD	**0.0068**	**0.9939**	0.1287	X	0.0752	0.6431	0.7548	0.7419	0.3875	**0.9981**	**0.9969**
GIH	**0**	0.8772	**0.0104**	0.0752	X	0.0686	0.7846	0.9663	0.1425	0.7708	**0.9971**
JPT	**0.0062**	0.9366	0.2092	0.6431	0.0686	X	0.8273	0.5424	0.4002	**0.9812**	**0.9985**
LWK	0.8683	**0.9971**	0.8592	0.7548	0.7846	0.8273	X	**0.9962**	**0**	0.8510	0.5671
MEX	**0.0063**	**0.0088**	0.4500	0.7419	0.9663	0.5424	**0.9962**	X	0.8748	0.2198	**1**
MKK	0.9619	0.6779	0.4040	0.3875	0.1425	0.4002	**0**	0.8748	X	**0.0067**	0.9488
TSI	**0**	0.6241	0.9344	**0.9981**	0.7708	**0.9812**	0.8510	0.2198	**0.0067**	X	**0.9979**
YRI	**0.0001**	**1**	**0.9982**	**0.9969**	**0.9971**	**0.9985**	0.5671	**1**	0.9488	**0.9979**	X

Table D.15: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.13

Table D.16: Parameter Estimates Table of the Model in Figure 5.14 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0013	0.0006
c_1	0.0051	0.0077	0.0064
c_2	0.0002	0.0011	0.0006
c_3	0.0005	0.0019	0.0012
c_4	0.0077	0.0178	0.0127
c_5	0.0081	0.0120	0.0099
c_6	0.0033	0.0058	0.0045
c_7	0.0001	0.0016	0.0007
c_8	0.0100	0.0147	0.0123
c_9	0.0011	0.0036	0.0024
c_{10}	0.0098	0.0127	0.0112
w_{11}	0.2123	0.2673	0.2394
c_{12}	0.1662	0.4374	0.2940
c_{13}	0.0085	0.0249	0.0168
c_{14}	0.0002	0.0052	0.0014
c_{15}	0.0004	0.0062	0.0041
c_{16}	0.1408	0.1642	0.1523
c_{17}	0.0069	0.0139	0.0104
c_{18}	0.0086	0.0161	0.0123
c_{19}	0.0284	0.0396	0.0338
c_{20}	0.1073	0.1352	0.1207
c_{21}	0.0127	0.0161	0.0143
c_{22}	0.0148	0.0206	0.0176
c_{23}	0.0172	0.0322	0.0245

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	**0**	**0.0080**	**0.0069**	**0**	**0.0056**	0.8616	**0.0035**	0.9597	**0**	**0.0001**
CEU	**0**	X	0.6719	**0.9882**	0.8050	0.8919	**0.9980**	0.0251	0.7143	0.6179	**1**
CHB	**0.0080**	0.6719	X	0.1223	0.5382	0.2064	0.8729	0.0595	0.4121	0.8438	**0.9986**
CHD	**0.0069**	**0.9882**	0.1223	X	0.8226	0.6303	0.7653	0.2145	0.3874	**0.9909**	**0.9968**
GIH	**0**	0.8050	0.5382	0.8226	X	0.7156	0.6351	0.3363	0.0531	0.6723	**0.9902**
JPT	**0.0056**	0.8919	0.2064	0.6303	0.7156	X	0.8248	0.0969	0.3790	0.9387	**0.9981**
LWK	0.8616	**0.9980**	0.8729	0.7653	0.6351	0.8248	X	**0.9961**	**0**	0.8762	0.5688
MEX	**0.0035**	0.0251	0.0595	0.2145	0.3363	0.0969	**0.9961**	X	0.8726	0.3350	**1**
MKK	0.9597	0.7143	0.4121	0.3874	0.0531	0.3790	**0**	0.8726	X	**0.0079**	0.9471
TSI	**0**	0.6179	0.8438	**0.9909**	0.6723	0.9387	0.8762	0.3350	**0.0079**	X	**0.9985**
YRI	**0.0001**	**1**	**0.9986**	**0.9968**	**0.9902**	**0.9981**	0.5688	**1**	0.9471	**0.9985**	X

Table D.17: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.14

Table D.18: Parameter Estimates Table of the Model in Figure 5.15 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0013	0.0006
c_1	0.0030	0.0067	0.0050
c_2	0.0002	0.0011	0.0006
c_3	0.0004	0.0019	0.0011
c_4	0.0079	0.0172	0.0123
c_5	0.0081	0.0122	0.0101
c_6	0.0034	0.0057	0.0045
c_7	0.0002	0.0016	0.0006
c_8	0.0098	0.0145	0.0121
c_9	0.0010	0.0034	0.0022
c_{10}	0.0098	0.0127	0.0112
w_{11}	0.2048	0.2656	0.2355
c_{12}	0.1048	0.3389	0.2121
c_{13}	0.0090	0.0237	0.0161
w_{14}	0.5093	0.5865	0.5474
c_{15}	0.0268	0.0624	0.0439
c_{16}	0.0002	0.0035	0.0011
c_{17}	0.0002	0.0055	0.0015
c_{18}	0.0003	0.0058	0.0036
c_{19}	0.0825	0.1079	0.0951
c_{20}	0.0551	0.0785	0.0666
c_{21}	0.0003	0.0038	0.0017
c_{22}	0.0117	0.0221	0.0169
c_{23}	0.0435	0.0610	0.0520
c_{24}	0.0796	0.1044	0.0917
c_{25}	0.0127	0.0161	0.0143
c_{26}	0.0149	0.0208	0.0179
c_{27}	0.0167	0.0307	0.0234

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	**0**	**0.0097**	**0.0087**	**0**	**0.0063**	0.8699	**0**	0.9605	**0**	**0.0001**
CEU	**0**	X	0.4331	0.9557	0.7228	0.7246	**0.9980**	0.4379	0.7411	0.5096	**1**
CHB	**0.0097**	0.4331	X	0.1369	0.4792	0.2245	0.8827	0.4157	0.4516	0.6189	**0.9985**
CHD	**0.0087**	0.9557	0.1369	X	0.8000	0.6899	0.7910	0.7138	0.4432	0.9584	**0.9975**
GIH	**0**	0.7228	0.4792	0.8000	X	0.6548	0.7753	0.5184	0.1442	0.6309	**0.9968**
JPT	**0.0063**	0.7246	0.2245	0.6899	0.6548	X	0.8314	0.4839	0.4025	0.7938	**0.9983**
LWK	0.8699	**0.9980**	0.8827	0.7910	0.7753	0.8314	X	0.8382	**0**	0.9249	0.5695
MEX	**0**	0.4379	0.4157	0.7138	0.5184	0.4839	0.8382	X	0.3105	0.8352	**0.9974**
MKK	0.9605	0.7411	0.4516	0.4432	0.1442	0.4025	**0**	0.3105	X	**0.0211**	0.9487
TSI	**0**	0.5096	0.6189	0.9584	0.6309	0.7938	0.9249	0.8352	**0.0211**	X	**0.9995**
YRI	**0.0001**	**1**	**0.9985**	**0.9975**	**0.9968**	**0.9983**	0.5695	**0.9974**	0.9487	**0.9995**	X

Table D.19: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.15

Table D.20: Parameter Estimates Table of the Model in Figure 5.16 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0016	0.0008
c_1	0.0034	0.0068	0.0052
c_2	0.0002	0.0011	0.0006
c_3	0.0005	0.0019	0.0011
c_4	0.0083	0.0180	0.0129
c_5	0.0082	0.0122	0.0102
c_6	0.0002	0.0012	0.0006
c_7	0.0002	0.0016	0.0007
c_8	0.0008	0.0037	0.0021
c_9	0.0013	0.0037	0.0025
c_{10}	0.0023	0.0049	0.0036
w_{11}	0.2110	0.2725	0.2418
c_{12}	0.1011	0.3383	0.2144
c_{13}	0.0091	0.0244	0.0165
w_{14}	0.5030	0.5843	0.5449
c_{15}	0.0269	0.0674	0.0450
c_{16}	0.0002	0.0033	0.0011
w_{17}	0.1868	0.2094	0.1976
c_{18}	0.0004	0.0153	0.0047
c_{19}	0.0003	0.0029	0.0014
c_{20}	0.0002	0.0049	0.0016
c_{21}	0.0004	0.0059	0.0035
c_{22}	0.0803	0.1075	0.0939
c_{23}	0.0556	0.0808	0.0681
c_{24}	0.0001	0.0030	0.0012
c_{25}	0.0002	0.0128	0.0029
c_{26}	0.0007	0.0177	0.0117
c_{27}	0.0425	0.0601	0.0511
c_{28}	0.0663	0.0889	0.0772
c_{29}	0.0109	0.0143	0.0126
c_{30}	0.0362	0.0432	0.0397
c_{31}	0.0348	0.0516	0.0429

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8416	0.6761	0.6479	0.5423	0.5404	0.5527	0.4123	0.2454	0.5992	0.7371
CEU	0.8416	X	0.4654	0.9609	0.7407	0.7474	**0.9968**	0.4654	0.4567	0.4692	**0.9997**
CHB	0.6761	0.4654	X	0.1370	0.4908	0.2345	0.8921	0.4202	0.3579	0.6219	**0.9782**
CHD	0.6479	0.9609	0.1370	X	0.8011	0.6815	0.8043	0.7136	0.3447	0.9567	0.9654
GIH	0.5423	0.7407	0.4908	0.8011	X	0.6556	0.6918	0.4879	0.0396	0.5865	0.9044
JPT	0.5404	0.7474	0.2345	0.6815	0.6556	X	0.8435	0.4876	0.3048	0.7944	**0.9775**
LWK	0.5527	**0.9968**	0.8921	0.8043	0.6918	0.8435	X	0.7522	0.0813	0.8893	0.2352
MEX	0.4123	0.4654	0.4202	0.7136	0.4879	0.4876	0.7522	X	0.1402	0.8486	0.9389
MKK	0.2454	0.4567	0.3579	0.3447	0.0396	0.3048	0.0813	0.1402	X	**0.0025**	0.8281
TSI	0.5992	0.4692	0.6219	0.9567	0.5865	0.7944	0.8893	0.8486	**0.0025**	X	0.9671
YRI	0.7371	**0.9997**	**0.9782**	0.9654	0.9044	**0.9775**	0.2352	0.9389	0.8281	0.9671	X

Table D.21: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.16

Table D.22: Parameter Estimates Table of the Model in Figure 5.17 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0017	0.0008
c_1	0.0032	0.0069	0.0052
c_2	0.0002	0.0011	0.0006
c_3	0.0004	0.0019	0.0011
c_4	0.0082	0.0178	0.0131
c_5	0.0082	0.0123	0.0102
c_6	0.0002	0.0012	0.0006
c_7	0.0002	0.0017	0.0007
c_8	0.0059	0.0131	0.0096
c_9	0.0002	0.0018	0.0008
c_{10}	0.0023	0.0047	0.0034
w_{11}	0.2100	0.2759	0.2433
c_{12}	0.0996	0.3323	0.2138
c_{13}	0.0090	0.0247	0.0166
w_{14}	0.4998	0.5774	0.5373
c_{15}	0.0280	0.0671	0.0467
c_{16}	0.0002	0.0038	0.0013
w_{17}	0.1890	0.2111	0.1996
c_{18}	0.0002	0.0133	0.0037
c_{19}	0.0003	0.0029	0.0014
w_{20}	0.1937	0.2315	0.2130
c_{21}	0.0344	0.1051	0.0680
c_{22}	0.0025	0.0136	0.0080
c_{23}	0.0002	0.0055	0.0023
c_{24}	0.0002	0.0056	0.0028
c_{25}	0.0791	0.1054	0.0922
c_{26}	0.0504	0.0747	0.0625
c_{27}	0.0002	0.0032	0.0013
c_{28}	0.0006	0.0031	0.0019
c_{29}	0.0003	0.0116	0.0030
c_{30}	0.0005	0.0177	0.0109
c_{31}	0.0487	0.0690	0.0585
c_{32}	0.0838	0.1143	0.0986
c_{33}	0.0109	0.0142	0.0125
c_{34}	0.0038	0.0114	0.0074
c_{35}	0.0684	0.0961	0.0818

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8284	0.7022	0.6768	0.5180	0.5675	0.3182	0.4098	0.6904	0.5688	0.7058
CEU	0.8284	X	0.4905	0.9667	0.7661	0.7637	0.9236	0.4995	0.9637	0.3227	**0.9990**
CHB	0.7022	0.4905	X	0.1417	0.4869	0.2242	0.7740	0.4261	0.5238	0.6479	0.9716
CHD	0.6768	0.9667	0.1417	X	0.8010	0.6746	0.6422	0.7191	0.5196	0.9637	0.9544
GIH	0.5180	0.7661	0.4869	0.8010	X	0.6590	0.2846	0.4970	0.2553	0.5903	0.8363
JPT	0.5675	0.7637	0.2242	0.6746	0.6590	X	0.7058	0.4895	0.4623	0.8121	0.9701
LWK	0.3182	0.9236	0.7740	0.6422	0.2846	0.7058	X	0.4448	0.1842	0.4733	0.2873
MEX	0.4098	0.4995	0.4261	0.7191	0.4970	0.4895	0.4448	X	0.3779	0.8417	0.9105
MKK	0.6904	0.9637	0.5238	0.5196	0.2553	0.4623	0.1842	0.3779	X	0.2314	0.9162
TSI	0.5688	0.3227	0.6479	0.9637	0.5903	0.8121	0.4733	0.8417	0.2314	X	0.9399
YRI	0.7058	**0.9990**	0.9716	0.9544	0.8363	0.9701	0.2873	0.9105	0.9162	0.9399	X

Table D.23: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.17

Table D.24: Parameter Estimates Table of the Model in Figure 5.18 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0016	0.0009
c_1	0.0050	0.0076	0.0063
c_2	0.0002	0.0011	0.0006
c_3	0.0005	0.0019	0.0012
c_4	0.0086	0.0180	0.0133
c_5	0.0081	0.0120	0.0100
c_6	0.0002	0.0012	0.0006
c_7	0.0002	0.0015	0.0007
c_8	0.0062	0.0132	0.0098
c_9	0.0002	0.0017	0.0008
c_{10}	0.0022	0.0046	0.0034
w_{11}	0.2216	0.2763	0.2489
c_{12}	0.1686	0.4149	0.2844
c_{13}	0.0100	0.0256	0.0173
w_{14}	0.1896	0.2108	0.2004
c_{15}	0.0003	0.0135	0.0044
c_{16}	0.0003	0.0027	0.0013
w_{17}	0.1942	0.2329	0.2139
c_{18}	0.0345	0.1073	0.0689
c_{19}	0.0025	0.0130	0.0078
c_{20}	0.0003	0.0059	0.0026
c_{21}	0.0002	0.0058	0.0028
$w22$	0.1405	0.1639	0.1520
c_{23}	0.0033	0.0107	0.0071
c_{24}	0.0006	0.0031	0.0019
c_{25}	0.0004	0.0114	0.0066
c_{26}	0.0002	0.0101	0.0031
c_{27}	0.0318	0.0448	0.0386
c_{28}	0.1105	0.1460	0.1278
c_{29}	0.0110	0.0142	0.0126
c_{30}	0.0034	0.0111	0.0072
c_{31}	0.0708	0.1000	0.0850

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8016	0.7799	0.7570	0.4673	0.6806	0.3305	0.8777	0.6981	0.4667	0.7083
CEU	0.8016	X	0.6998	**0.9895**	0.8254	0.9082	0.9023	0.0289	0.9558	0.4748	**0.9988**
CHB	0.7799	0.6998	X	0.1251	0.5284	0.2020	0.7267	0.0535	0.5193	0.8543	0.9622
CHD	0.7570	**0.9895**	0.1251	X	0.8229	0.6345	0.5823	0.2022	0.5023	**0.9924**	0.9400
GIH	0.4673	0.8254	0.5284	0.8229	X	0.7216	0.1675	0.2297	0.1566	0.6569	0.7297
JPT	0.6806	0.9802	0.2020	0.6345	0.7216	X	0.6717	0.0906	0.4789	0.9479	0.9651
LWK	0.3305	0.9023	0.7267	0.5823	0.1675	0.6717	X	0.9248	0.1846	0.3623	0.2833
MEX	0.8777	0.0289	0.0535	0.2022	0.2297	0.0906	0.9248	X	0.8545	0.3170	**0.9984**
MKK	0.6981	0.9558	0.5193	0.5023	0.1566	0.4789	0.1846	0.8545	X	0.1525	0.9159
TSI	0.4667	0.4748	0.8543	**0.9924**	0.6569	0.9749	0.3623	0.3170	0.1525	X	0.8984
YRI	0.7083	**0.9988**	0.9622	0.9400	0.7297	0.9651	0.2833	**0.9984**	0.9159	0.8984	X

Table D.25: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.18

Table D.26: Parameter Estimates Table of the Model in Figure 5.19 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0017	0.0008
c_1	0.0031	0.0067	0.0051
c_2	0.0002	0.0011	0.0006
c_3	0.0004	0.0019	0.0012
c_4	0.0110	0.0172	0.0141
c_5	0.0081	0.0122	0.0101
c_6	0.0002	0.0013	0.0006
c_7	0.0002	0.0014	0.0006
c_8	0.0061	0.0133	0.0097
c_9	0.0002	0.0019	0.0009
c_{10}	0.0023	0.0047	0.0035
w_{11}	0.4828	0.5507	0.5165
c_{12}	0.0481	0.0896	0.0669
c_{13}	0.0002	0.0031	0.0011
w_{14}	0.1872	0.2065	0.1968
c_{15}	0.0003	0.0166	0.0055
c_{16}	0.0003	0.0027	0.0013
w_{17}	0.1947	0.2330	0.2144
c_{18}	0.0342	0.1045	0.0674
c_{19}	0.0030	0.0133	0.0081
c_{20}	0.0037	0.0071	0.0053
c_{21}	0.0695	0.0915	0.0804
c_{22}	0.0402	0.0597	0.0497
c_{23}	0.0002	0.0030	0.0012
c_{24}	0.0007	0.0031	0.0019
c_{25}	0.0002	0.0031	0.0011
c_{26}	0.0373	0.0476	0.0424
c_{27}	0.0264	0.0404	0.0334
c_{28}	0.0941	0.1262	0.1100
c_{29}	0.0108	0.0141	0.0124
c_{30}	0.0030	0.0112	0.0072
c_{31}	0.0675	0.0951	0.0809

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8117	0.7078	0.6823	0.7337	0.5832	0.3257	0.2519	0.7102	0.4740	0.7030
CEU	0.8117	X	0.6603	**0.9877**	0.7538	0.8751	0.8826	0.5075	0.9506	0.3844	**0.9984**
CHB	0.7078	0.6603	X	0.1400	0.0305	0.2410	0.7790	0.4269	0.5531	0.8275	0.9705
CHD	0.6823	**0.9877**	0.1400	X	0.1670	0.6904	0.6459	0.7213	0.5386	**0.9902**	0.9540
GIH	0.7337	0.7538	0.0305	0.1670	X	0.1221	0.6559	0.6956	0.6190	0.5525	**0.9755**
JPT	0.5832	0.8751	0.2410	0.6904	0.1221	X	0.7190	0.4887	0.4922	0.9267	0.9716
LWK	0.3257	0.8826	0.7790	0.6459	0.6559	0.7190	X	0.2487	0.1699	0.3020	0.2910
MEX	0.2519	0.5075	0.4269	0.7213	0.6956	0.4887	0.2487	X	0.2042	0.8382	0.7985
MKK	0.7102	0.9506	0.5531	0.5386	0.6190	0.4922	0.1699	0.2042	X	0.1330	0.9129
TSI	0.4740	0.3844	0.8275	**0.9902**	0.5525	0.9267	0.3020	0.8382	0.1330	X	0.8760
YRI	0.7030	**0.9984**	0.9705	0.9540	**0.9755**	0.9716	0.2910	0.7985	0.9129	0.8760	X

Table D.27: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.19

Table D.28: Parameter Estimates Table of the Model in Figure 5.20 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0003	0.0023	0.0011
c_1	0.0030	0.0069	0.0052
c_2	0.0002	0.0011	0.0006
c_3	0.0005	0.0019	0.0012
c_4	0.0081	0.0174	0.0125
c_5	0.0082	0.0122	0.0101
c_6	0.0033	0.0057	0.0045
c_7	0.0002	0.0016	0.0006
c_8	0.0099	0.0179	0.0140
c_9	0.0002	0.0021	0.0010
c_{10}	0.0098	0.0127	0.0112
w_{11}	0.2071	0.2691	0.2373
c_{12}	0.0905	0.3406	0.2130
c_{13}	0.0092	0.0238	0.0161
w_{14}	0.5051	0.5797	0.5413
c_{15}	0.0277	0.0661	0.0459
c_{16}	0.0002	0.0033	0.0010
w_{17}	0.1165	0.1583	0.1385
c_{18}	0.0088	0.1385	0.0650
c_{19}	0.0053	0.0160	0.0105
c_{20}	0.0002	0.0058	0.0028
c_{21}	0.0002	0.0055	0.0023
c_{22}	0.0809	0.1066	0.0935
c_{23}	0.0534	0.0775	0.0652
c_{24}	0.0003	0.0037	0.0014
c_{25}	0.0004	0.0027	0.0015
c_{26}	0.0105	0.0215	0.0160
c_{27}	0.0453	0.0637	0.0543
c_{28}	0.0948	0.1249	0.1094
c_{29}	0.0114	0.0148	0.0131
c_{30}	0.0002	0.0044	0.0019
c_{31}	0.0249	0.0427	0.0336

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	**0**	**0.0016**	**0.0013**	**0**	**0.0011**	0.9538	**0**	0.9376	**0**	**0.0012**
CEU	**0**	X	0.4793	0.9613	0.7535	0.7546	**0.9982**	0.4747	0.9657	0.3753	**1**
CHB	**0.0016**	0.4793	X	0.1267	0.4860	0.2381	0.8958	0.4250	0.5508	0.6508	**0.9988**
CHD	**0.0013**	0.9613	0.1267	X	0.7916	0.6716	0.8006	0.7093	0.5275	0.9628	**0.9977**
GIH	**0**	0.7535	0.4860	0.7916	X	0.6664	0.7506	0.5281	0.3162	0.6259	**0.9962**
JPT	**0.0011**	0.7546	0.2381	0.6716	0.6664	X	0.8431	0.4873	0.4906	0.8120	**0.9986**
LWK	0.9538	**0.9982**	0.8958	0.8006	0.7506	0.8431	X	0.8124	**0**	0.9238	0.5664
MEX	**0**	0.4747	0.4250	0.7093	0.5281	0.4873	0.8124	X	0.4361	0.8387	**0.9966**
MKK	0.9376	0.9657	0.5508	0.5275	0.3162	0.4906	**0**	0.4361	X	0.2232	0.9433
TSI	**0**	0.3753	0.6508	0.9628	0.6259	0.8120	0.9238	0.8387	0.2232	X	**0.9995**
YRI	**0.0012**	**1**	**0.9988**	**0.9977**	**0.9962**	**0.9986**	0.5664	**0.9966**	0.9433	**0.9995**	X

Table D.29: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.20

Table D.30: Parameter Estimates Table of the Model in Figure 5.21 after 100,000 iterations

	95% HPD Interval Bounds		
Parameter	lower	upper	Median
c_0	0.0002	0.0016	0.0008
c_1	0.0029	0.0064	0.0048
c_2	0.0002	0.0011	0.0006
c_3	0.0004	0.0019	0.0012
c_4	0.0129	0.0193	0.0160
c_5	0.0081	0.0121	0.0101
c_6	0.0002	0.0012	0.0006
c_7	0.0001	0.0013	0.0006
c_8	0.0008	0.0041	0.0024
c_9	0.0015	0.0039	0.0027
c_{10}	0.0024	0.0048	0.0036
w_{11}	0.4922	0.5644	0.5290
c_{12}	0.0427	0.0831	0.0615
c_{13}	0.0002	0.0033	0.0010
w_{14}	0.1846	0.2043	0.1947
c_{15}	0.0002	0.0131	0.0046
c_{16}	0.0003	0.0027	0.0013
c_{17}	0.0037	0.0071	0.0054
c_{18}	0.0724	0.0948	0.0835
c_{19}	0.0433	0.0631	0.0530
c_{20}	0.0002	0.0031	0.0014
c_{21}	0.0002	0.0038	0.0013
c_{22}	0.0347	0.0451	0.0399
c_{23}	0.0225	0.0351	0.0288
c_{24}	0.0757	0.0995	0.0873
c_{25}	0.0108	0.0141	0.0125
c_{26}	0.0357	0.0429	0.0392
c_{27}	0.0357	0.0530	0.0440

The table shows the posterior mean and 95% HPD intervals for the drift, c, and admixture, w, parameters.

p-value	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
ASW	X	0.8260	0.6897	0.6594	0.7605	0.5641	0.5821	0.2642	0.2320	0.4916	0.7404
CEU	0.8260	X	0.6626	**0.9870**	0.7843	0.8763	**0.9955**	0.5123	0.3351	0.5153	**0.9995**
CHB	0.6897	0.6626	X	0.1433	0.1842	0.2429	0.8999	0.4292	0.3859	0.8285	**0.9774**
CHD	0.6594	**0.9870**	0.1433	X	0.1210	0.6825	0.8096	0.7156	0.3679	**0.9901**	0.9637
GIH	0.7605	0.7843	0.1842	0.1210	X	0.0918	0.9326	0.6698	0.2533	0.6025	**0.9883**
JPT	0.5641	0.8763	0.2429	0.6825	0.0918	X	0.8570	0.4973	0.3421	0.9259	**0.9783**
LWK	0.5821	**0.9955**	0.8999	0.8096	0.9326	0.8570	X	0.6003	0.0916	0.7959	0.2448
MEX	0.2642	0.5123	0.4292	0.7156	0.6698	0.4973	0.6003	X	0.0431	0.8381	0.8599
MKK	0.2320	0.3351	0.3859	0.3679	0.2533	0.3421	0.0916	0.0431	X	**0.0001**	0.8413
TSI	0.4916	0.5153	0.8285	**0.9901**	0.6025	0.9259	0.7959	0.8381	**0.0001**	X	0.9223
YRI	0.7404	**0.9995**	**0.9774**	0.9637	**0.9883**	**0.9783**	0.2448	0.8599	0.8413	0.9223	X

Table D.31: p-values for Pairwise F_{ST} for Each Pair of Subpopulations Produced from Post Predictive Checking of the Model in Figure 5.21