Department of Statistics

MSc Statistics
February 2010

Bayesian analysis of models of population divergence for SNP variation data

by

Stuart MacIntyre

A thesis submitted to the University of Glasgow for the degree of Master of Science

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

February 2010

Acknowledgements

I would like to thank Dr. Vincent Macaulay for his guidance throughout my research. Without his kind advice and encouragement along the way the completion of this thesis would not have been possible. I would also like to thank my friends and family for their enduring support before, during and hopefully after the completion of this thesis.

Abstract

Probabilistic models to describe genetic differentiation between populations typically fail to include the effect of complex ancestry. A Bayesian hierarchical model proposed by Nicholson et al. (2002) (ND) provides a framework for assessing differentiation using population-wise parameters for single-nucleotide polymorphism (SNP) data under certain assumptions regarding the evolution of allele frequencies over time. Although the ND model offers a coherent method to estimate population divergence, a rather simplistic assumption must be made about the historical evolution of populations. Since shared ancestry between populations results in correlations in allele frequencies, it is the potential capture of such correlations that motivates the development of the new model reported here.

This thesis presents a review of the ND model using simulated and newly available SNP data, highlighting situations where the ND model does and does not fit the data well. The model was fitted using Markov-chain Monte-Carlo (MCMC) methods, and the fit assessed using residual diagnostics. Nicholson et al. (2002) reported instability in parameter estimates when a population was removed from the data set and the model re-fitted. Analysis of simulated data ensured that this is not an inherent property of the ND model and therefore can be used to highlight discrepancies with the model. Analyses on real data show that the ND model works well for groups of Europeans with low levels of genetic differentiation between populations, but a lack of fit is found when groups of populations dispersed across continents are considered. Data are also simulated under an alternative ancestral configuration and it is shown that lack of fit, manifest in residuals and estimator instability, is present when analysed using the ND model. An extension to the ND model is developed and fitted, supposing that discrepancies in the modelling assumptions of the ND model are due to the effect of alternative ancestral relationships. The ND and the new model are compared, as regards their fit to various data sets, and it is found that in some cases the new model does provide a better fit and in other cases the distinction is unclear. The new model is also used to infer the most likely ancestral relationships between populations sampled from the Human Genome Diversity Panel.

Keywords: Bayesian model, population differentiation, residuals, ancestry

Contents

Chapter 1
Introduction

Understanding the structure of human populations is crucial to many areas of scientific research such as the mapping of genes associated with common diseases, forensics and the environmental sciences**.** For example, when conducting genetic association studies, a failure to acknowledge differences in population structure between cases and controls can lead to spurious results, in particular an inflation of type I error (Marchini et al., 2004). If we are prepared to make some assumptions about the evolutionary processes responsible for patterns of variation observed in DNA samples from a collection of populations, inferences can be made about the history and relationships of such populations.

Over the last two decades major advancements have been made in the experimental manipulation of DNA fragments, giving scientists access to huge volumes of genetic data. Such data are the result of various complex processes and attempts to understand the patterns of variation have led to the development of statistical models which rely on existing population genetics theory. Single nucleotide polymorphisms (SNPs, pronounced "snip") have become the marker of choice for genetic studies in recent years, a genetic marker being a piece of DNA, variable between individuals, whose position on the genome is known and whose inheritance can be traced. A SNP is simply a single position in the DNA at which there is known to be variation between individuals within a species (Nicholson et al., 2002). Modelling the complex mechanisms which generated the observed data using traditional likelihood methods has in the past been problematic as maximisation of the likelihood function over a large number of parameters is a computationally difficult task. The recent surge in popularity of Bayesian approaches to statistical inference in population genetics is largely due to the potential for parameter-rich models with inter-dependency to be handled with relative ease (Beaumont and Rannala, 2004). Nicholson et al. (2002) proposed a Bayesian hierarchical model for SNP data in a pure drift setting using population-specific parameters to describe population differentiation and isolation and a simple structure of evolutionary history. This thesis will develop new methods to account for uncertainty in the ancestry of sampled populations while adhering to the probabilistic structure of the model suggested by Nicholson et al. (2002).

1.1 Context

It is often of importance to scientists in many differing fields to have an idea of human population structure and also some notion of the extent of differentiation between populations. Any interpretation of the observed pattern of genetic diversity found in a sample can be potentially misleading without the formal assessment of hidden population structure (Excoffier, 2007). Intellectual interest in divergence between populations is common in areas such as anthropology, in the case of humans, where quantitative measures are used to aid understanding and further knowledge of the processes responsible for the observed variation. Demographic history (i.e. historic population sizes and migration patterns) is also important in elucidating patterns of genetic variation. For a group of populations it may be of interest to quantify the genetic distance between populations but also to infer the historical evolutionary path such populations have taken. For example, knowledge of the relationships between sampled populations and the most recent common ancestral population (MRCAP) is of obvious relevance to scientists interested in the history of such populations. Where humans are concerned, language or phenotypic differences may be used to classify populations and distinguish between them. Thus qualitative estimates of differentiation can be obtained. However over the last 30 years, advances in biotechnology, leading to the availability of DNA sequence data, have permitted the development of methods to quantify genetic diversity and differentiation. The assessment of populations at the DNA level leads to a much broader perspective than would be gained through simple qualitative methods. Human populations have been studied extensively since the advent of genetic sequencing techniques and are the focus of this study.

Humans are diploid organisms; that is, their genome consists of pairs of chromosomes, of which there are 23. In every pair of chromosomes one is maternally inherited and the other paternally, and so offspring contain a sample of genetic material from their parents. Single chromosomes have a double helix structure (see Figure 1-1) where each strand contains a DNA (deoxyribonucleic acid) sequence complementary to the sequence on the corresponding strand; a phenomenon known as base-pairing, critical in replication processes. A DNA sequence consists of four basic molecules called nucleotides: adenine (A), guanine (G), thymine (T), cytosine (C), and it is the precise linear order of these chemicals along the chromosomes that comprise an individual’s genetic constituent. At any given position, or ‘locus’, on a chromosome pair in an individual there are two ‘bases’, one on each of the pair of chromosomes. If the two bases are the same the individual is said to be homozygous at that locus, and if they are different the individual is heterozygous at the given locus. Bearing in mind the complementary relationship between strands of DNA within the same chromosome, it is useful while referring to Figure 1-1 to focus attention on a particular strand, let’s say the red strand in both of the chromosomes. In the first chromosome of the pair at the highlighted locus a ‘C’ is found whereas in the second a ‘T’ is found. The individual from whom this sample was taken would then be heterozygous at the highlighted locus. The variants found at a locus are known as alleles and it is the frequency of such alleles in populations that are used to describe differences between and within those populations (Lewin, 2004). This concept can be extended to genes where a locus is no longer a single base position but the location of a gene, which has specified functions. However, for reasons that become clear in the following description of SNPs, we will consider a locus to be the location of a single base. The two bases present at a SNP in an individual are called the genotype at that SNP and the process of determining the genotype is called genotyping.

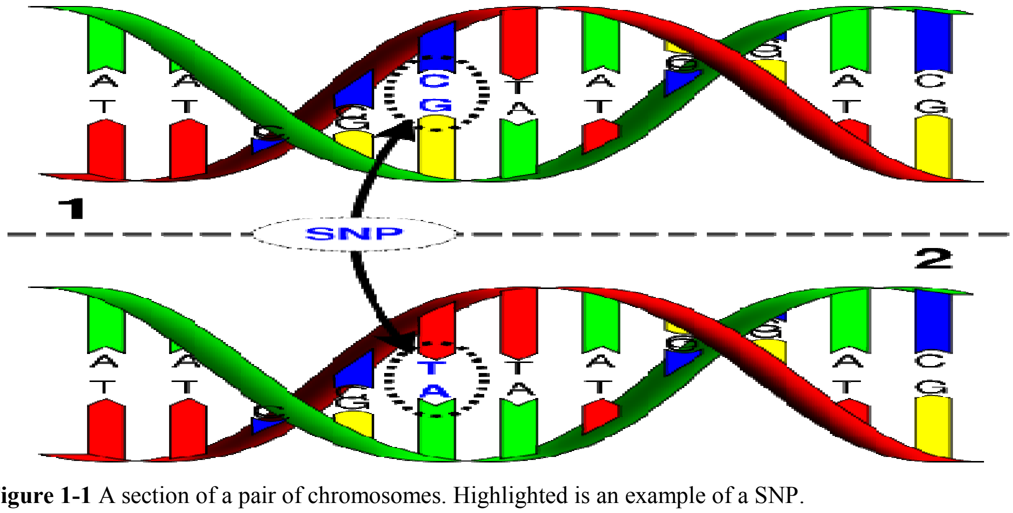


Figure 1-1 A section of a pair of chromosomes. Highlighted is an example of a SNP.

It has recently become economically feasible to genotype individuals at a large number of SNP loci, which are then used to study genetic variation in populations. Each SNP’s existence stems from an error in the DNA copying process at some time in the past, known as a mutation. It is generally the case, and will be assumed throughout, that SNPs only exhibit two variants: in the jargon they are bi-allelic. Such locations are found during small-scale identification or ascertainment studies after which it would be typical to genotype a sample of individuals at a number of SNP loci. Previously unseen levels of data are becoming available through projects such as the International HapMap Project (http://www.hapmap.org/) and the Human Genome Diversity Project (HGDP) (http://www.stanford.edu/group/morrinst/hgdp.html), both international collaborations between scientists with the aim of providing publicly-available resources to aid the understanding of human genetic diversity.

A particular collaboration between the Human Genome Diversity Panel and CEPH (Centre d'Etude du Polymorphism Humain, translated as Human Polymorphism Study Center) in Paris, has resulted in a collection of DNA samples from 1050 individuals in 51 world populations being banked and subsequently the availability of genotype data at 650,000 SNP loci for the 1050 individuals. Not only the remarkable volume of data available but also the geographic area covered by the sampled populations makes these data well suited for studies of human diversity.

Current technology provides the means to genotype an individual at a huge number of loci simultaneously. Many different SNP genotyping techniques are currently in use but in general they can be categorised into hybridisation-based and enzyme-based methods. Both rely on the base-pairing property of DNA alluded to above where adenine (A) pairs with thymine (T) and guanine (G) pairs with cytosine (C). Hybridisation methods use two short pieces of synthetic DNA for each SNP called primers, designed to complement the target sequence. A heating process breaks the weak hydrogen bonds between the two strands of the double-helix and the primers are then exposed to the denatured DNA. If the sample DNA contains the allele of interest then the complementary relationship between the two will form a hybrid segment. Hybridisation can then be assessed using various visualisation techniques. Enzyme-based methods cover a wide variety of different techniques, commonly using either DNA ligase, DNA nucleases or DNA polymerase to catalyse specific reactions designed to yield detectable mutations at specific sites on a DNA sequence (Ye et al., 2001).

Table 1. Format of SNP data collected on *P* populations at *L* SNPs.

	SNP Locus (Reference Nucleotide)				
Population	***j*=1 (A)**	***j*=2 (G)**	***j*=3 (C)**	**…**	***j*=*L* (C)**
***i*=1**	$n_{11} = 110$ $x_{11} = 75$	$n_{12} = 100$ $x_{12} = 43$	$n_{13} = 120$ $x_{13} = 100$		$n_{1L} = 100$ $x_{1L} = 90$
⋮	⋮	⋮	⋮	⋱	⋮
i*=*P	$n_{P1} = 114$ $x_{P1} = 80$	$n_{P2} = 110$ $x_{P2} = 0$	$n_{P3} = 100$ $x_{P3} = 100$	…	$n_{PL} = 110$ $x_{PL} = 75$

Table 1 illustrates some SNP data from *P* populations at *L* SNPs where n_{ij} is twice the number of individuals typed in population *i* at SNP *j*, as we are dealing with diploid individuals; and x_{ij} is the number of copies of the randomly chosen reference nucleotide at SNP *j* in population *i,* shown bracketed in Table 1. Therefore x_{ij} / n_{ij} is the sample allele frequency of the reference nucleotide in population *i* at SNP *j*. The raw data consists of the genotype of each individual at each SNP locus; so table 1 is a compact version of the tallied genotype data. It is also worth reiterating that the reference nucleotide is chosen at random from two possible candidates as SNPs are assumed to be bi-allelic.

The information contained across many independent SNP loci can lead to accurate inferences about demographic characteristics and the historical and contemporary relationships between populations. As is so often the case in a statistical analysis, it is advantageous for observations, in this case SNPs, to be independent, as the mathematical manipulation becomes more difficult if this property cannot be assumed. For the independence assumption to hold it must be the case that the transmission of genetic information from parent to child at a particular SNP locus has no bearing on the probability of inheritance of the information at another SNP locus. If alleles at different SNP loci tend to be co-inherited from the same parent then independence would be violated. During meiosis, the production of sex cells called gametes, the closer loci are positioned in relation to one another on the chromosome the more likely they are to be co-inherited. On the other hand, if loci are sufficiently distant from one another, recombination, the shuffling of genes during meiosis, allows independence to be assumed.

The majority of positions on the genome are identical across individuals and so SNPs are a more cost effective way of studying variation. The alternative would be to sequence stretches of DNA, most of which are not variable. However, the efficiency of SNPs comes with a penalty. Generally speaking, the more polymorphic a locus is (i.e the closer the allele frequencies are to 0.5) the more likely it is to be discovered in the ascertainment process, which brings with it the possibility of biased estimates. Researchers have included in their modelling procedure the effect of SNP ascertainment, with differing conclusions. Nicholson et al. (2002) found that estimates were not sensitive to the inclusion of an ascertainment effect whereas Balding & Nichols (1995) and M. Sharif (2007) found that it was important to model ascertainment in their analysis in certain circumstances. The difficulty in modelling such an effect is the variation in ascertainment procedures carried out and it is unlikely that any particular method to account for ascertainment is appropriate for all ascertainment schemes (Nielson, 2004). To formulate a meaningful probability expression one must have available the details of the procedure which can be difficult to obtain and not always reliably stated. Nevertheless it seems appropriate to model this effect whenever possible.

Genetic differentiation simply means that allele frequencies among populations are different (Hartl and Clark, 2007) and implies some population structure. This can be due to differences in the frequencies of founder individuals of the populations, chance fluctuations caused by the sampling involved in reproduction, known as random genetic drift, or selection favouring different alleles within sub-populations, perhaps corresponding to variable environmental conditions. Traditional measures of differentiation are based around the fixation index or F_{ST}, proposed by Sewall Wright in the 1920's, which gives a *single* quantitative measure of the proportion of the overall genetic variability ascribable to a certain level of population sub-division. Formal definitions of F_{ST} have evolved and multiplied since its inception. However definitions tend to rely on arguments relating to heterozygosity, one being the reduction in heterozygosity expected with random mating at any one level of a population hierarchy relative to another, more inclusive level of the hierarchy (Hartl and Clark, 2007). Heterozygosity is a measure of the genetic variability of a population and is the frequency of heterozygotes averaged over the tested loci (Falconer, 1989). The definition of F_{ST} makes intuitive sense as levels of heterozygosity decrease in the presence of population sub-division, relative to a randomly mating population. A common mathematical description is

$$F_{ST} = \frac{H_T - H_S}{H_T}. \qquad [1]$$

In words Equation [1] is the difference in total heterozygosity and that of a given level of sub-structuring $(H_T - H_S)$ relative to the total (H_T).

F_{ST} is limited as an estimator of divergence as it is an average over all populations. Thus estimates of population-wise divergence, characteristic of the model proposed below, offer more insight into diversity. An analogy can be made with the ANOVA procedure where one may be investigating the effectiveness of a group of treatments. The first stage would involve an analysis of the overall treatment effect; however if this was found to be statistically significant the natural continuation would be to carry out some comparisons to find where the differences occur. In our setting the F_{ST} value could act, loosely speaking, as an indicator of overall sub-division from which we can proceed to investigate more precisely, using population-wise parameters, the patterns of differentiation.

Probabilistic gene frequency models have been developed using population genetics theory, inherently statistical in nature, in an attempt to quantify differentiation between populations (Gillespie, 2004). Nicholson et al. (2002) proposed a Bayesian hierarchical model for SNP data to describe differentiation using population-specific parameters, closely analogous to $F_{ST.}$ Such parameters appear in the variance structure of the imposed normal distributions characterising allele frequencies at a given SNP in a given population. Population genetics interpretations and justifications are given and will be discussed in later sections. Since frequencies are necessarily on [0, 1], the normal distribution used to model them has to be truncated at 0 and 1, by placing point masses there. This implies a mixed distribution so that in (0, 1) the distribution is continuous (and so densities are evaluated) whereas at the boundaries the distribution is discrete and hence mass is evaluated. The advantage of using such a distribution is that it mirrors the feature of allele frequencies in a population called fixation. This occurs when an allele is lost and without mutation cannot return to the gene pool, an inevitable event in a pure-drift setting (see below) over a large number of generations. The use of normal distributions also permits the assessment of model fit, a rarity in population genetics, using various residual diagnostic plots.

Underlying the Nicholson et al. (ND) model is an assumption about demographic history; the sampled populations diverged simultaneously from an ancestral population some time in the past and have continued to evolve independently. This implies that gene flow is not occurring between populations or its effect is small and so negligible, a setting sometimes called pure drift. An alternative approach, proposed by Balding and Nichols (BN) (1995) and widely used in forensic DNA profiling, assumes that equilibrium has been reached through the contributions of both migration (gene flow) and random drift such that the levels of variation between populations are constant, and uses beta distributions to characterise allele frequencies. The BN approach accounts for the effect of gene-flow though assumes that differentiation has been and continues to be constant through time. The question is then to consider the context and hence the more useful model. In our case the BN model does not give an idea of the history of the sampled populations and since one of the aims of this thesis is to develop a new model for representing alternative evolutionary histories of sampled populations, the ND model is more attractive.

To make any practical sense a model must be related to or derived from a physical process. The formal assumption in the ND model which stipulates an ancestral population splitting into descendant populations can be related to a historical event where members of a population migrated to another location but were then unable to return, perhaps due to a geographical barrier. These populations would then have evolved through time independently as the exchange of genes between populations was impossible. As the earth's landscape has changed dramatically over time this type of event is plausible.

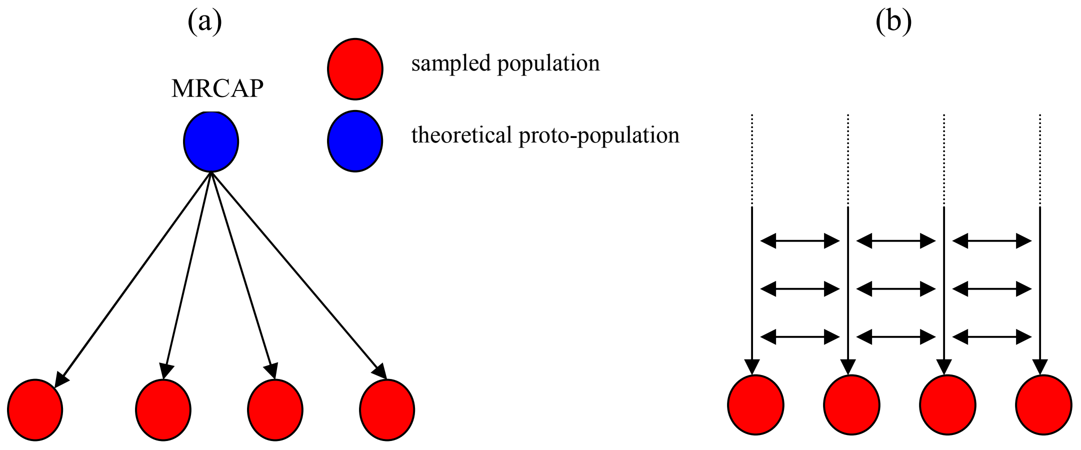


Figure 1-2 (a) Evolutionary pattern inferred by ND model for 4 populations. Notice the single ancestral population splitting simultaneously into four and evolving over time until the present. (b) BN model. Notice that distance between populations is constant through time. The single arrows represent the direction of time and the dashes reflect that the process has been occurring indefinitely over time. The double-headed arrows represent gene-flow.

If populations split relatively cleanly, with little subsequent gene flow between sub-populations, then a tree is an accurate representation of the relationships between populations and their historical path, as in Figures 1-2(a) and 1-3. Figures 1-2(a) and 1.2(b) illustrate the ND and BN models respectively. It is important to note that in Figure 1-2, both depictions are of *populations* evolving over time. This is distinct from another situation that can be represented using this tree format: to represent the relationships between a group of individuals. We will call the representation in Figure 1-3 (a) a topology. A topology defines the history of a set of populations (or individuals) without specific labelling. Figure 1-3(b) illustrates a labelled history, which is a topology with a specific labelling. Within a topology, any proto-population is the most recent common ancestral population (MRCAP) of any populations below it on the tree. The MRCAP of all sampled populations is called the root of the tree and represents the single population from which all the sampled contemporary populations are descended.

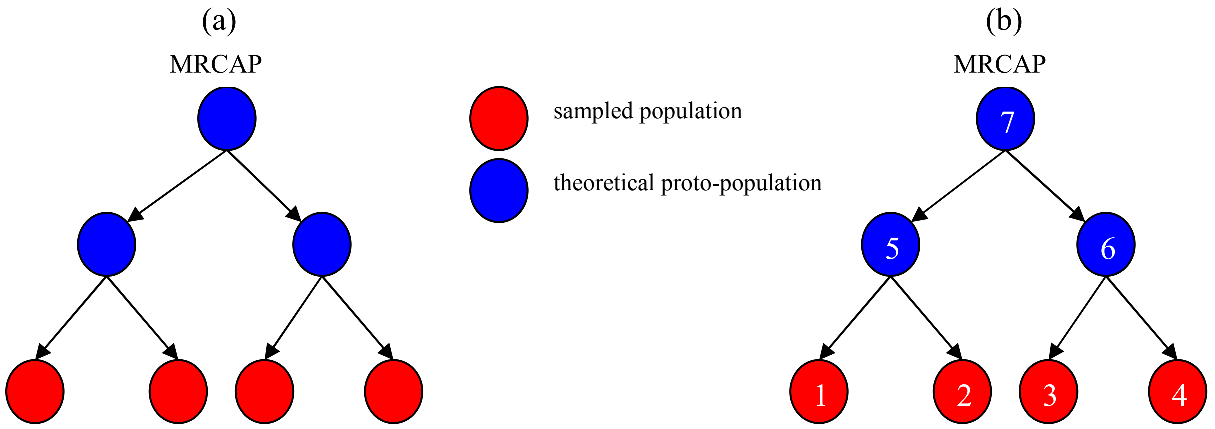


Figure 1-3 (a) A topology with distinction between sampled and theoretical populations. (b) A labelled history; populations 1-4 represent sampled populations, 5-7 represent theoretical populations. The arrows represent the direction of time.

Tree representations are useful as they display the hierarchical structure of models such as the ND model. Figure 1-3(b) illustrates the historical relationships of the sampled populations but we can also state that population 1 is more closely related to population 2 than populations 3 or 4 are to 1 or 2 and a similar relationship is evident between populations 3 and 4. The tree structures in Figure 1-3 are bifurcating which means that any proto-population splits into two populations. This is important as we will only be considering bifurcating trees when specifying more complex models. It may well be the case that a bifurcating tree is not the correct representation of the sampled populations, but since a bifurcating tree can give a good approximation to other topologies, for example a trifurcating tree, using very small intermediate branches, and the possible number of trees can become unmanageable if not handled sensibly, it is for our purposes a practical necessity.

Many of the developments in population genetics in recent years have been the consequence of increasing computing power, allowing researchers to use models with large numbers of parameters, in an attempt to reflect the complex processes influencing the data. Taking a Bayesian statistical approach and utilising Markov Chain Monte Carlo (MCMC) simulation methods offers the potential for a large number of parameters, with inter-dependency, to be handled in a practical manner such, that meaningful conclusions can be drawn (Beaumont and Rannala, 2004).

Within a Bayesian framework parameters are considered random quantities and inference is based on the marginal probability distribution of the parameters of interest conditional on the observed data, called the marginal posterior distribution. Basic summaries of these distributions such as means and variances are used to make inferential statements and draw conclusions. A probability model is specified according to some notion of the underlying process and prior distributions are used to quantify what is known about the parameters a priori*,* i.e. before the new data is taken into account. The assignment of prior distributions is the primary concern for the critics of Bayesian methodology as it is a fundamental requirement, whether or not cogent prior knowledge is available. However under such circumstances one can test the sensitivity of results to the prior thus gaining insight into the influence of this distribution.

The essence of Bayesian analysis is Bayes' rule (Bayes, 1763),

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \qquad [2]$$

where θ is the set of model parameters and y is the data. Therefore the posterior distribution of θ, $p(\theta|y)$, i.e. conditional on y, is proportional to the product of the prior distribution of θ, $p(\theta)$, and the distribution of y given θ, $p(y|\theta)$, commonly called the likelihood. The intermediate step in calculating $p(\theta|y)$ in equation [2] is the term $p(\theta, y)$ which is the joint distribution of θ and y. This represents the full probability model which is developed using relevant knowledge and theory from the field of study. The term $p(y)$ does not depend on θ

and with fixed data can be considered a constant, yielding the un-normalised posterior distribution

$$p(\theta|y) \propto p(y|\theta)p(\theta). \qquad [3]$$

Various techniques can then be employed to attain $p(\theta|y)$. MCMC methods have been used extensively over the last 10 years in studies of genetic variation, particularly for Bayesian hierarchical models, to take samples from $p(\theta|y)$.

MCMC methods refer to the use of Monte Carlo integration using Markov chains. Iterative in nature, the objective of MCMC is to sample from the posterior distribution of quantities of interest by repeated sampling over the parameter space. This is achieved by defining a Markov chain which has as its stationary distribution the required posterior density and running the algorithm for a sufficient length of time. At each stage in the process values of θ are drawn from approximate distributions and then corrected so that those draws are a better approximation of the posterior density (Gelman et al., 2004a). Inference is then based on simple summaries of the posterior distribution such as the mean and variance, after removing the initial period before convergence of the chain known as burn-in. Many algorithms have been proposed to carry out this task but most are similar to or special cases of the Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970), the details of which will be discussed in the next chapter.

1.2 Aims

I propose to fulfil the aims set out in this section.

- Develop an MCMC algorithm to fit the ND model of SNP allele frequencies to simulated and real data sets.

- Assess the fit of the ND model for both simulated data and newly available real data using residual and population-removal diagnostic techniques, highlighting situations where the model does and does not fit the data.

- Develop an extension to the ND model which allows flexibility in the evolutionary histories of contemporary populations, implemented again using MCMC methods.

- Assess the fit of the new model for real and simulated data sets and investigate whether there is information in the data to infer the most appropriate labelled history for a set of populations, using residual diagnostic techniques.

Chapter 2
Methods

The first step in any Bayesian statistical analysis is the formulation of a probability model which effectively represents the processes responsible for variation observed in the data. In our context this refers to a model to describe variation in allele frequencies at many independent SNP loci for a set of populations. The ND model proposed by Nicholson et al. (2002) is defined below, specifically the probabilistic structure of the model with statistical and population-genetics justifications.

2.1 ND Model for SNP Allele Frequencies

This model was proposed to describe SNP allele frequencies for structured populations while simultaneously estimating population-wise parameters aiming to capture historical differences between populations. In this setting a population is simply a breeding unit, meaning that only within the population can an individual find a mate to produce offspring. Mating is also considered to be random with respect to genotype, a standard assumption in population genetics. That is, mates are not chosen directly or indirectly for their genotype.

Suppose we have a sample of SNP data collected from P populations at L SNPs. Then let n_{ij} be the number of chromosomes typed in the ith population at the jth SNP which corresponds to twice the number of individuals typed. As mentioned in section 1.1 an arbitrarily selected nucleotide is chosen for every SNP and the number of copies of the chosen allele in population i at SNP j is x_{ij}, $0 \leq x_{ij} \leq n_{ij}$. The unobserved frequency of the chosen allele in the ith population at the jth SNP is denoted by α_{ij}, $0 \leq \alpha_{ij} \leq 1$. For ease of representation the omission of subscripts will denote the entire collection of quantities; so, for example, x will represent the set of all x_{ij} , $i = 1, 2, \ldots, P, j = 1, 2, \ldots, L$.

At the lowest level of the hierarchical model, we have binomial data: conditional on n and α,

$$x_{ij} \sim \text{Binomial}(n_{ij}, \alpha_{ij}), \qquad i = 1, 2, \ldots, P; j = 1, 2, \ldots, L, \qquad [4]$$
$$\text{independently} \ \forall \ i, j.$$

As we have taken a sample from the whole population α is unknown and so assigned a probability distribution. It is worth noting that the maximum likelihood estimate of α is x / n from the properties of binomial random variables. The population allele frequency α_{ij} is then modelled as

$$\alpha_{ij} \sim \text{Normal}(\pi_j, c_i \pi_j (1 - \pi_j)), \qquad i = 1, 2, \ldots, P; j = 1, 2, \ldots, L, \qquad [5]$$
$$\text{independently} \ \forall \ i, j.$$

The distributional expression in [5] is the basis of the ND model and will be justified in detail in what follows. The introduction of the unobserved quantities π and c can be explained by reference to the Wright-Fisher model of evolution in an idealized population. At present it is sufficient to define π_j as ($0 < \pi_j < 1$) the allele frequency at SNP j in the population ancestral to all sampled populations. Note that π_j has no population index as the model assumes a single ancestral population split into P populations at some time in the past. For simplicity it is also assumed that there was variation in the ancestral population at every SNP. If this was not assumed then conceptually without mutation or migration no variation would be present at that SNP. Since mutation and migration are not assumed to be present or their effects negligible then it is necessary to stipulate that $\pi_j \neq 0$ or 1. This is probably a reasonable assumption since most SNPs are variable in most populations and SNP mutation rates are known to be low (International HapMap Consortium, 2005), so a mutation arising in many populations independently has a low probability.

The population-wise parameters c_i are those which we aim to estimate and describe the amount of genetic drift population i has been subjected to since splitting from its ancestral population. In a statistical sense, c governs the amount, in terms of variance, the contemporary population allele frequencies tend to be different from typical values (Nicholson et al., 2002). Its relation to the variance of the allele frequencies leads to the stipulation that c is strictly non-negative.

To complete the hierarchy, we place independent prior distributions on π and c:

$\pi_1, \ldots, \pi_L$ are independent and identically distributed with density f; [6]

$c_1, \ldots, c_P$ are independent and identically distributed with density g. [7]

A discussion of particular prior distributions will follow in section 2.1.2.4 but for now the general statements in [6] and [7] will suffice.

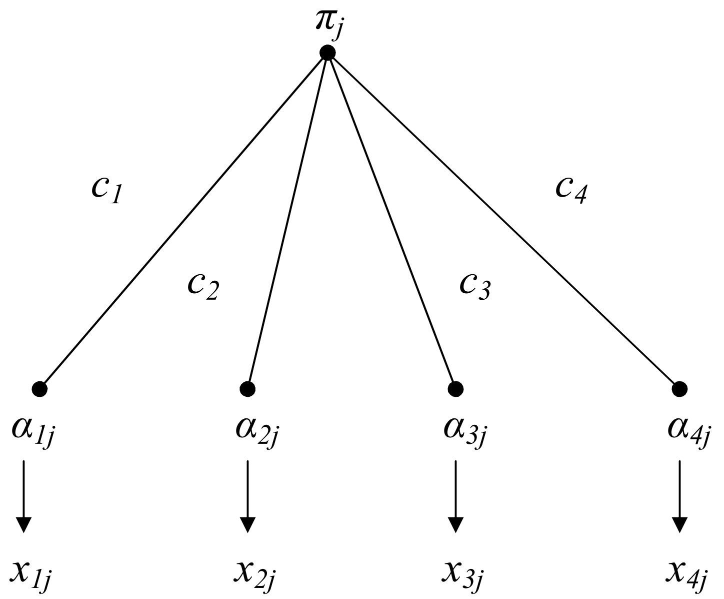


Figure 2-1 The phylogenetic structure of the ND model for a single SNP j, for $P = 4$.

Figure 2-1 is a useful representation of the ND model as the phylogenetic structure is apparent; for a If we take a prospective approach to describe the model, then, at some time in the past, a single population split simultaneously into four populations. The plausibility of such an event was discussed in section 1.1. If the populations evolve to the present day in the manner to be discussed in this section, in subsequent isolation and also if SNPs are not under selective pressure, then the ND model is an accurate representation. So not only is it desirable to assure that the SNPs to be analysed under the ND model are independent, but also that they are chosen from a region which is thought to be of no functional value to the individual, to avoid, as much as possible, the effects of selection. In fact most of the genome likely evolves without selection (Kimura, 1983). The assumption of populations evolving independently of one another, or, in population genetic terms, without migration, is of greater concern to the legitimacy of the modelling assumptions. If correlations between populations are present then it can be difficult to unravel the underlying forces affecting parameter values.

For example, a large estimate of *c* could be the due to the long period of isolation since splitting from the ancestral population but it could also be due to gene flow between other sampled populations which in turn exaggerates the isolation of the population in question. However, it is the possibility and potential capture of such correlations between populations which motivates the extension of the ND model proposed in the latter part of this thesis.

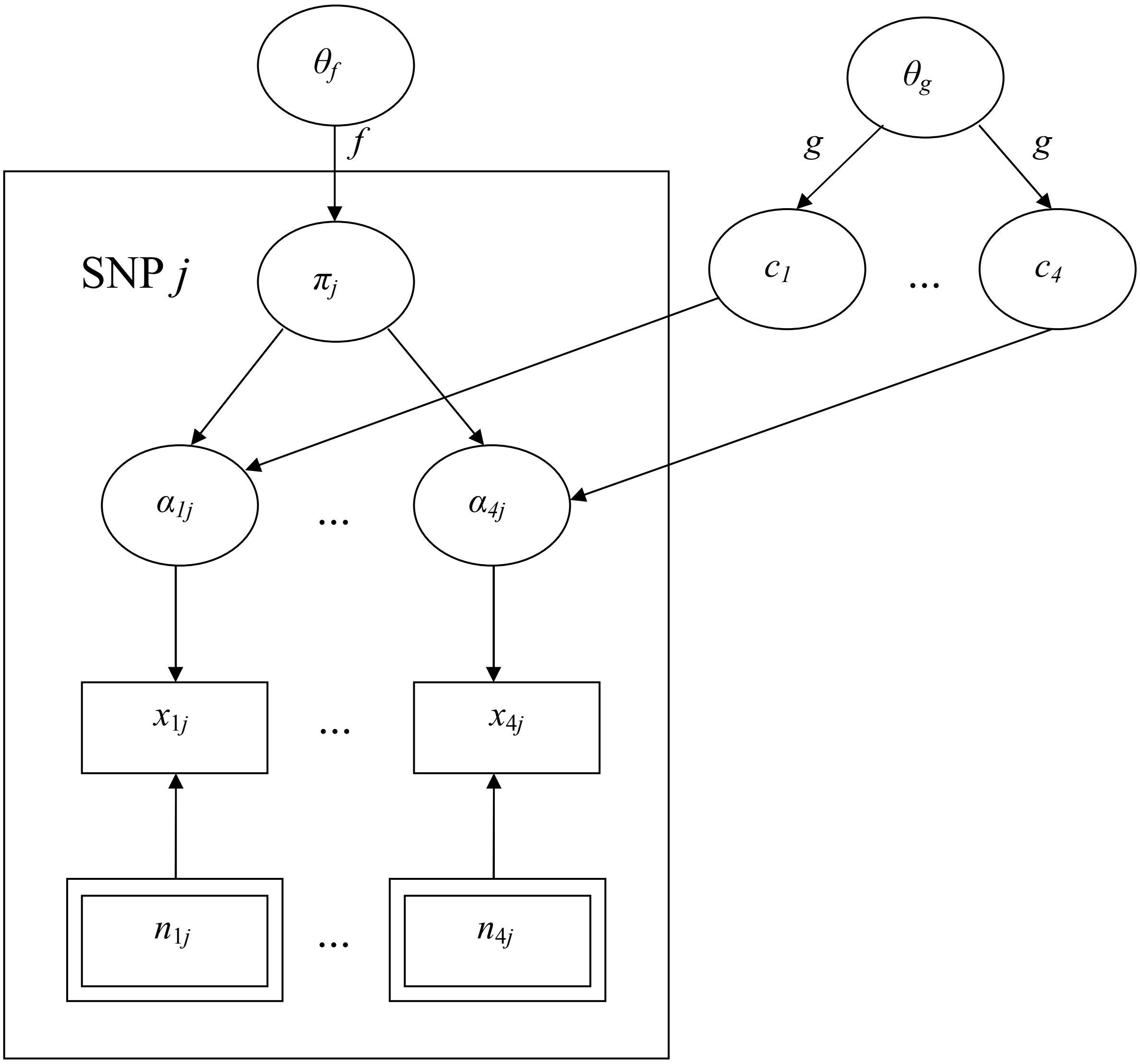


Figure 2-2 A directed acyclic graph (DAG) of the ND model for four populations.

The DAG in Figure 2-2 illustrates the probabilistic relationships between the parameters and the random variables defined in the ND model. In this format circles represent parameters; rectangles represent random variables (i.e. the data) and double rectangles represent quantities assumed fixed by design (i.e. sample sizes). The direction of an arrow specifies the condition that a quantity below the arrow is conditionally independent given the quantity at the top of the arrow. This equates to saying that the random variables are conditionally independent given the parameters that characterise their distribution. In this particular DAG, θ_f is the set of parameters characterising the distribution f (prior on the π's) and θ_s is the set of parameters characterising the distribution g (prior on the c's). Repetitive structures, of SNPs within populations, are shown as stacked "sheets".

Another feature of the ND model is the mixed distribution in expression [5]. This means that continuous distributions will be specified for quantities within the range (0, 1), with atoms at 0 and 1, whose size is the total mass of the relevant distribution on (-∞, 0] and [1, ∞), respectively. Figure 2-3 illustrates a standard normal mixed distribution. The distribution is continuous on (-1, 1) and has point masses at -1 and 1 equal to the "missing" tails of the normal distribution tales discrete on (-∞, -1] and [1, ∞). The mass at both atoms -1 and 1 is 0.1592, calculated from the standard normal cumulative distribution function.

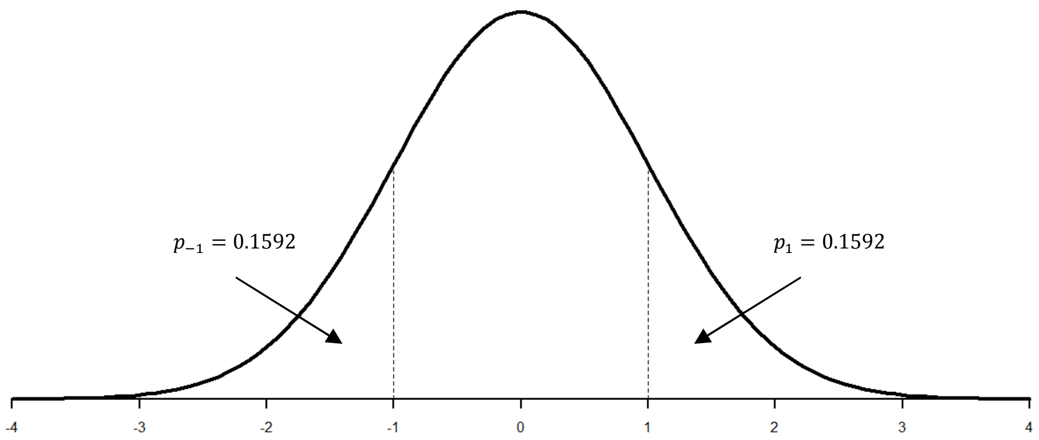


Figure 2-3 A mixed standard normal distribution, i.e. $\mu = 0$, $\boldsymbol{\sigma^2 = 1}$. Within the range (-1, 1) the distribution is continuous; point masses of 0.1592 are found at atoms -1 and 1, calculated from the tails of the normal distribution.

The use of a mixed distribution for allele frequencies reflects the need to handle probabilistically the situation where α_{ij} = 0 or 1, called fixation, where an allele is lost and therefore only a single allele remains, given that we are referring to bi-allelic SNPs.

The model defined above does not include an ascertainment effect as mentioned in section 1.1. The decision to leave out this aspect of the sampling procedure reflects the aims set out in section 1.2. The primary concern is to develop a new model to account for uncertainty in the topology, not necessarily to acquire unbiased estimates using the new model at this stage. However with the correct information regarding the ascertainment procedure, particularly the size of the panel used in the SNP discovery process and the population the individuals were sampled from, it would be possible to include the effect into the model and it is suggested at least to explore this possibility in any future analysis. The difficulty lies in obtaining the relevant information and providing an adequate characterization of the process. The ascertainment protocol used for the HapMap data is extremely complex and it would be impossible to model such a process mathematically (Clark et al., 2005). In any further analyses using these data, where an ascertainment correction is sought, a simplified version of the SNP discovery process should be modelled.

2.1.1 **The Wright - Fisher Model**

The binomial distribution in expression [4] reflects the sampling process involved when using SNP data. However, binomial sampling is also a feature of the evolution of allele frequencies over time in an idealized population. Consider a population comprising $2N$ chromosomes with two alleles at a given locus. Then assume that the population size is constant through time, that generations are non-overlapping, meaning that parents do not survive into the offspring generation and mating between individuals is random with respect to genotype. Deviations from the assumption of random mating could be due to differential reproductive fitness of particular genotypes, inbreeding or age-structured populations where the fertility of an individual is a function of age; however populations under random mating are often assumed in population genetics models due to their mathematical tractability. If mutation and recombination cannot introduce new alleles and new genotypes respectively and there exists no differential reproductive fitness between the two alleles, then the evolution of allele frequencies can be described by a Markov chain and this model is known as the Wright–Fisher model of evolution under genetic drift (Fisher, 1930; Wright, 1931). The total number of the reference allele in a given generation specifies the state of the chain and it follows that the allele frequency is easily calculated given the state. In our example, the state S_t at time t can take the values $0, \dots, 2N$. Note that time in this context is discrete as we are dealing with non-overlapping generations. There are probabilities associated with every possible transition from any state at time t to any state at time $t + 1$ known as the transition probabilities and these are calculated using the binomial formula. If $S_t = j$ where $j = 1, \ldots,$ $2N$ and $k = 0, \ldots, 2N,$ then

$$prob\left(S_{t+1} = k \middle| S_t = j\right) = \frac{(2N)!}{k!\left(2N-k\right)!}\left(j/2N\right)^k\left(1-j/2N\right)^{2N-k}.$$

Under these idealized conditions the changes in allele frequency are purely stochastic since the genes in the offspring generation are a random sample from the parent generation. This stochastic change in allele frequency is known as random genetic drift and is the evolutionary force we focus on in this thesis. The *c* parameters in expression [5] reflect the amount of drift a population has undergone since splitting from the MRCAP. Of the parameters estimable within the ND model, the *c*'s are the most informative in our context; the allele frequencies at particular SNPs do not provide any practical information and are in effect treated as nuisance parameters.

The inherent binomial property of evolution under the Wright-Fisher model leads us to the parameterization of the variance in expression [5]. From the properties of binomial random variables, if $z \sim \text{Bi}(\text{n},\theta)$, then the natural estimator of θ is $\hat{\theta} = z/n,$ and $\text{Var}\left(\hat{\theta}\right) = \frac{\theta(1-\theta)}{n},$ which has the same form as the variance component in [5] with π in place of θ and $c = 1/n$. Since the effect of genetic drift is inversely proportional to population size, such that, in a small population drift is more pronounced, compared to a large population where it has a minor effect, the *c*'s are consistent with this property.

In this simplified situation binomial sampling is present at every generational step. However even with such simplifications and the known relationship from generation to generation, to derive the mathematical properties of the Wright-Fisher model through many generations is an extremely demanding task as the random fluctuations between every generation must be captured. However a result exists, although not proved here due to the highly involved mathematics of diffusion equations, which approximates the allele frequency after many generations using a normal distribution (Kimura, 1983). The result is derived assuming the population size is large, and in practice works well as long as some of the populations are not very small. The *c* parameters are also proportional to time, since the amount of genetic drift increases over time, so large values of *c* suggest a longer time since splitting from the MRCAP and vice versa. This result also mirrors the approximation of the binomial distribution by the normal distribution, a well-known statistical property.

2.1.2 **Markov-Chain Monte-Carlo Methods**

Markov-Chain Monte-Carlo (MCMC) methods are used to sample from distributions of interest when direct sampling is not possible. These situations occur frequently in a Bayesian setting as models tend to involve large numbers of parameters with non-standard marginal posterior distributions. The Metropolis-Hastings algorithm is an iterative procedure closely related to a random walk used to sample from posterior distributions. The algorithm relies on the Markov property, which states that the future state of the chain only depends on the present state. Hence at every step in the algorithm a draw is made from a proposal distribution depending only on the current state and evaluated using an acceptance/rejection criterion. The proposal distribution and the acceptance/rejection rule are carefully constructed such that the stationary distribution of the Markov chain is the posterior distribution of interest, the idea being that if the algorithm is run for a sufficient length of time the draws will be from the target distribution. Inference is then based on summarising the marginal posterior distributions of interest using moments and appropriate plots (Gelman et al., 2004a).

The output from an MCMC chain is often thinned (only every *n* draws are kept, where *n* is typically around 10) to reduce the correlation between samples, this correlation often being assessed via autocorrelation plots (e.g. the acf() function in R). In this case it was not done as a matter of course, since (i) independent samples are not *essential* to compute summaries of the posterior distribution (they just lead to somewhat more stable estimates) and, in any case, (ii) in most examples there were not any notable cases of strong correlation effects, after appropriate tuning of the proposal distribution.

2.1.2.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm was employed to sample from posterior distributions of interest. Other strategies could have been employed, such as the Gibbs sampler; however, the Metropolis–Hastings algorithm is simpler to implement since it does not require the calculation of and sampling from full conditional distributions.

The general algorithm proceeds as follows:

1. Choose a vector of starting values θ_0, for all parameters in the model.

2. For $t = 1, 2, 3, \dots,$

   (a) Sample a proposal θ^* from a proposal distribution at time *t*, $J_t(\theta^* | \theta^{t-1})$.

   (b) Calculate the ratio of the densities,

$$r = \frac{p(\theta^* | y) / J_t(\theta^* | \theta^{t-1})}{p(\theta^{t-1} | y) / J_t(\theta^{t-1} | \theta^*)},$$

   where $p(.|y)$ is the posterior density.

   (c) Set

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \quad \text{“acceptance”} \\ \theta^{t-1} & \text{otherwise} \quad \text{“rejection.”} \end{cases}$$

The factor $J_t(\theta^{t-1} | \theta^*) / J_t(\theta^* | \theta^{t-1})$ is used in the calculation of r to account for a non-symmetric proposal distribution. If $J_t(.|.)$ is symmetric i.e. $J_t(\theta^{t-1} | \theta^*) = J_t(\theta^* | \theta^{t-1})$, the factor reduces to unity. Note also that if the proposed value is rejected, such that $\theta^t = \theta^{t-1}$, this counts as an iteration in the chain.

2.1.2.2 Implementation

In this section the Metropolis-Hastings algorithm implemented in our analysis is described in detail referring back to the general form in the previous section. The updating strategy employed updates “batches” of parameters sequentially, using the appropriate formula for calculating *r* (see (b) below), since each of the three groups of parameter, π, α and c, has a unique formula.

1. The starting values $\underline{\theta}^0 = (\pi, \alpha, c)^0$ are chosen using approximate values of the parameters taken from simple estimators, for example, F_{ST} for the *c* parameters. It is crucial that samples from the chain eventually become independent of the starting values. That is to say, regardless of the starting values the chain should converge to the target distribution, on a reasonable time scale. Therefore multiple simulations using starting values dispersed throughout the parameter space are considered. If it appears that the chain has not converged on the target distribution then more iterations may be needed.

2. (a) Gelman et al. (2004) suggest the following criteria for choosing the proposal distribution $J_t(\theta^t | \theta^{t-1})$, at a given *t*.

   - For any θ, it is easy to sample from $J(\theta^* | \theta)$.

   Since updates are performed in groups, each of the parameter sets has a unique proposal distribution. Using normal proposal distributions for π, α and *c* ensures that the above property is satisfied and it is also clear that the assumption of a symmetric proposal distribution is upheld. See the section 2.1.2.3 for a further discussion of particular proposal distributions.

   - It is easy to compute the ratio *r*

   See below (b).

   - Each jump goes a reasonable distance in the parameter space (otherwise the random walk moves too slowly).

   - The jumps are not rejected too frequently (otherwise the random walk wastes too much time standing still).

   The last two conditions can be grouped together as they both refer to the ‘mixing’ of the chain. To achieve satisfactory mixing, a balance between jumping a sufficient amount and not jumping too far must be made, for example, by adjusting the standard deviation of the proposal distribution.

(b) Some useful simplifications can be made when calculating the ratio of densities r depending on which type of parameter is being updated that greatly increase the efficiency of the algorithm.

Note that

$$p(\pi,\alpha,c|x) \propto \left[\prod_{ij} p(\alpha_{ij}|\pi_j,c_i)p(x_{ij}|n_{ij},\alpha_{ij})\right]\left[\prod_{j} p(\pi_j|\mu_\pi,\sigma_\pi^2)\right]\left[\prod_{i} p(c_i|\mu_c,\sigma_c^2)\right],$$

where $\mu_\pi,\sigma_\pi,\mu_c,\sigma_c$ are hyper-parameters indexed by the prior distribution they parameterise. This is (up to a constant that depends only on the data) the full joint posterior distribution of the ND model expressed in terms of known conditional distributions.

When updating c_i ($i = 1, 2, \ldots, P$),

$$r = \frac{\left\{\prod_{j=1}^{L} p(\alpha_{ij}|\pi_j,c_i^*)p(x_{ij}|n_{ij},\alpha_{ij})p(\pi_j|\mu_\pi,\sigma_\pi^2)\right\}p(c_i^*|\mu_c,\sigma_c^2)}{\left\{\prod_{j=1}^{L} p(\alpha_{ij}|\pi_j,c_i)p(x_{ij}|n_{ij},\alpha_{ij})p(\pi_j|\mu_\pi,\sigma_\pi^2)\right\}p(c_i|\mu_c,\sigma_c^2)} = \frac{\left\{\prod_{j=1}^{L} p(\alpha_{ij}|\pi_j,c_i^*)\right\}p(c_i^*|\mu_c,\sigma_c^2)}{\left\{\prod_{j=1}^{L} p(\alpha_{ij}|\pi_j,c_i)\right\}p(c_i|\mu_c,\sigma_c^2)},$$

where c_i^* is the proposed value and c_i is the current value.

When updating π_j ($j = 1, 2, \ldots, L$),

$$r = \frac{\left\{\prod_{i=1}^{P} p(\alpha_{ij}|\pi_j^*,c_i)p(x_{ij}|n_{ij},\alpha_{ij})p(c_i|\mu_c,\sigma_c^2)\right\}p(\pi_j^*|\mu_\pi,\sigma_\pi^2)}{\left\{\prod_{i=1}^{P} p(\alpha_{ij}|\pi_j,c_i)p(x_{ij}|n_{ij},\alpha_{ij})p(c_i|\mu_c,\sigma_c^2)\right\}p(\pi_j|\mu_\pi,\sigma_\pi^2)} = \frac{\left\{\prod_{i=1}^{P} p(\alpha_{ij}|\pi_j^*,c_i)\right\}p(\pi_j^*|\mu_\pi,\sigma_\pi^2)}{\left\{\prod_{i=1}^{P} p(\alpha_{ij}|\pi_j,c_i)\right\}p(\pi_j|\mu_\pi,\sigma_\pi^2)},$$

where π_j^* is the proposed value and π_j is the current value.

When updating α_{ij} ($i = 1, 2, \ldots, P; j = 1, 2, \ldots, L$),

$$r = \frac{p(\alpha_{ij}^*|\pi_j,c_i)p(x_{ij}|n_{ij},\alpha_{ij}^*)p(c_i|\mu_c,\sigma_c^2)p(\pi_j|\mu_\pi,\sigma_\pi^2)}{p(\alpha_{ij}|\pi_j,c_i)p(x_{ij}|n_{ij},\alpha_{ij})p(c_i|\mu_c,\sigma_c^2)p(\pi_j|\mu_\pi,\sigma_\pi^2)} = \frac{p(\alpha_{ij}^*|\pi_j,c_i)p(x_{ij}|n_{ij},\alpha_{ij}^*)}{p(\alpha_{ij}|\pi_j,c_i)p(x_{ij}|n_{ij},\alpha_{ij})},$$

where α_{ij}^* is the proposed value and α_{ij} is the current value.

(c) The acceptance/rejection condition is considered using the ratio *r*. The simple case is where $r \geq 1$ then the proposed value is accepted, since it is at least as probable as the current value. If $r < 1$ then the proposed value is accepted if $r > s$, where *s* is a draw from a Un(0, 1) distribution. That is, the proposal is accepted with probability *r*. It is also worth mentioning that when calculating *r*, sums and differences of log probabilities are used and then exponentiated at the end (to avoid over or under-flow during computation).

2.1.2.3 Proposal Distributions

The use of a mixed distribution to describe contemporary allele frequencies poses complications when drawing values from the proposal distribution. When drawing from a normal proposal distribution for *α,* it does not suffice to simply reject values outwith the range as the distribution has mass at the boundaries 0 and 1. To overcome this problem the following re-parameterisation of the ND model was used. First we define a function $t(x)$ such that,

$$t(x) = \begin{cases} x & \text{for } 0 \leq x \leq 1, \\ 0 & \text{for } x < 0, \\ 1 & \text{for } x > 1. \end{cases}$$

Then introduce the quantity $\beta_{ij} \in \Re$ $(i = 1, \ldots, P, j = 1, \ldots, L)$ such that $\alpha_{ij} = t(\beta_{ij})$ and

$$\beta_{ij} \sim \text{Normal}\left(\pi_j, c_i \pi_j (1 - \pi_j)\right). \qquad [8]$$

Expression [1] can now be written

$$x_{ij} \sim \text{Binomial}\left(n_{ij}, t(\beta_{ij})\right). \qquad [9]$$

Therefore the contemporary allele frequencies are expressed in terms of *β* whose parameter space spans the real line. The function $t(x)$ is required in expression [9] since *β* can potentially be < 0 or > 1 in which case the expression becomes invalid. If we were interested in inferring *α* then the function $t(x)$ would be used to transform back to the *α* scale and hence valid allele frequencies.

With no restrictions on the parameter space of β_{ij}, normal distributions can be used for all of the proposal distributions, effectively implementing random walks through each of the parameter spaces (Gelman et al., 2004a). However there are some other issues which must be dealt with to ensure adherence to the modelling assumptions. Firstly, proposed values of π outwith the range (0, 1) are immediately rejected under the ND model. Also, since c is strictly non-negative, a transformation onto the log-scale was used to enable the use of a normal proposal distribution which is on the preferred real line scale.

Formally, the proposal distributions $J_t(.|.)$ at iteration t for π, β and ln$c$ are,

$$J_t(\pi^t|\pi^{t-1}) \sim \text{Normal}(\pi^{\text{t-1}}, \sigma_1^2), \qquad [10]$$

$$J_t(\beta^t|\beta^{t-1}) \sim \text{Normal}(\beta^{\text{t-1}}, \sigma_2^2), \qquad [11]$$

$$J_t\left(\ln(c^t)|\ln(c^{t-1})\right) \sim \text{Normal}\left(\ln(c^{t-1}), \sigma_3^2\right) \qquad [12]$$

Characterising the variance in the proposal distributions are the parameters $\sigma_i^2\ (i = 1, 2, 3)$, which affect the efficiency of the algorithm and are adjusted accordingly (see section 2.3).

2.1.2.4 Prior Distributions

Prior distributions represent knowledge about parameters before considering the data. There are two useful ways of conceptualising a prior distribution: the first being the probabilistic characterisation of the investigator's knowledge about the parameter(s), maybe drawing on results from previous studies; the second supposes that the current parameter value is a draw from a population of possible values, which the prior distribution reflects (Gelman et al., 2004b). Both standpoints are equally valid in their respective contexts, however, regardless of interpretation, a prior distribution must have within its range the possible values of the parameter it describes and also it must quantify the uncertainty in the knowledge about the parameter. The distinction between a prior distribution whose form highlights more probable parameter values and one in which all values are equally probable is an important one; the latter being a non-informative prior and the former informative. Both are used in our analysis and the particular distributions used will be illustrated in this section. In choosing a sensible prior distribution an informed judgement has been made, drawing on relevant information when available, in the context of the particular problem, while also testing the influence of particular assignments on estimates.

Two prior distributions *f* and *g*, characterising π and *c* respectively, are specified in the ND model (see [6] and [7]). Let's consider the prior distribution of π, with density *f*. Recall that π_j is the allele frequency of the SNP *j* in the ancestral population and it is assumed that variation was present at SNP *j* in the ancestral population therefore $\pi_j \neq 0$ or 1. So the prior on π must be on the range (0, 1). Both the Beta and the Un(0, 1) distribution have this property (in fact, the Uniform is a special case of the Beta) and so both are considered in our analyses. It is a consequence of the way in which SNPs are discovered that loci with more variation tend to be found. Since an allele frequency value of 0.5 represents the maximum amount of polymorphism, a Beta(2, 2) distribution is a useful way of reflecting this property in the prior distribution of π (see Figure 2.4).

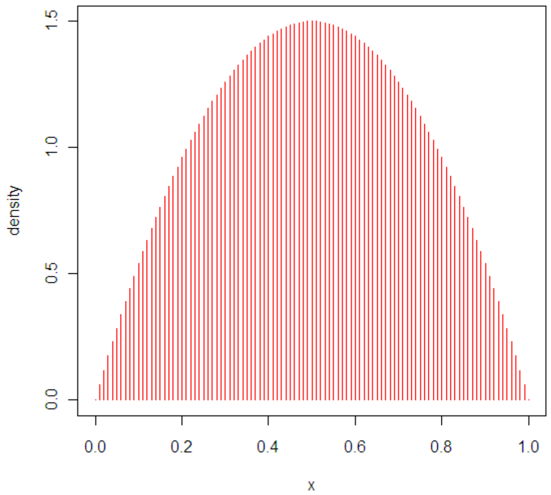


Figure 2-4 Probability density function of Beta(2, 2) distribution.

We can incorporate, if appropriate, prior information about the drift parameters into the prior distribution using estimates such as F_{ST}. Our degree of uncertainty in such an estimate is quantified in the variance of the prior distribution. Throughout the analyses this has been set at values corresponding to large variation, representing our uncertainty as well as ensuring that undue influence is not placed on the posterior distribution by the prior. A log-normal prior distribution was used for *c* meaning that the natural logarithm of *c* is normally distributed. As the proposal distribution of *c* is on the log-scale, it was decided that a prior distribution on log-scale should be sought. The normal distribution is a standard prior for parameters on the real-line and was therefore a natural choice.

2.1.3 Assessment of Model Fit

The use of normal distributions in the ND model allows for an assessment of fit, using standard residual analysis, giving a useful way of highlighting possible discrepancies in the modelling assumptions.

Given that

$$\alpha_{ij} \sim \mathrm{Normal}\big(\pi_j, c_i\pi_j(1-\pi_j)\big),$$

then

$$\left\{ \frac{x_{ij}/n_{ij} - \hat{\pi}_j}{\left\{(\hat{c}_i + (1-\hat{c}_i)/n_{ij})\,\hat{\pi}_j(1-\hat{\pi}_j)\right\}^{1/2}}; i = 1,\ldots,P, j = 1,\ldots,L \right\} \qquad [13]$$

are taken as the set of standardised residuals (where $\hat{\pi}_j$ and $\hat{c}_i$ denote the posterior mean of π_j and c_i, respectively) (Nicholson et al., 2002). The formula in [13] has been modified slightly in relation to the conventional form of standardised residuals by the inclusion of the estimated value of α_{ij}, x_{ij}/n_{ij}. If α_{ij} were known then [13] would be

$$\left\{ \frac{\alpha_{ij} - \hat{\pi}_j}{\left\{\hat{c}_i\hat{\pi}_j(1-\hat{\pi}_j)\right\}^{1/2}} \right\}.$$

A derivation of the formula in [13] is given the Appendix A but on inspection it makes intuitive sense as when the sample size n_{ij} tends to infinity and therefore the estimate of α_{ij} becomes more precise, the term $(1-\hat{c}_i)/n_{ij}$ tends to zero leaving the standard formula above.

If the normal assumptions are reasonable the standardised residuals ought to resemble a sample from a standard normal distribution. A *Q-Q* plot is used to assess this feature of the residuals by plotting them against theoretical quantiles of the standard normal distribution and the variance structure can be analysed by plotting the residuals against the fitted π's. The robustness of the estimates can also be checked by removing a population from the data set and re-fitting the model to see whether the estimates are stable. Nicholson et al. (2002) observed that estimates were highly unstable in certain situations, particularly when populations were analysed whose evolutionary history was not represented well by the ND model, given current understanding. For example, a data set including African, Melanesian, European and Chinese populations showed high instability under the leave-one-out diagnostic. It is highly unlikely that the ND model represents these populations effectively, as regards evolutionary history, since it is widely accepted that modern humans evolved in Africa, therefore the simultaneous diverging of populations assumed under the ND model does not hold. It may then be possible to use the residual and population removal diagnostics to highlight problems with the ND model. If these analyses show discrepancies it may be the case that the model does not sufficiently represent correlations between populations, possibly resultant of more complex historical relationships, in which case an extension to the model could help to elucidate these relationships.

2.2 Simulation Methods

Using simulated data is an invaluable way to test the performance of an MCMC algorithm and, in our context, highlight potential improvements or inconsistencies in the model when used in conjunction with the diagnostics proposed in section 2.1.3. A probability model is used to generate data under different scenarios by specifying particular values of parameters designed to answer particular questions. As in many statistical analyses, there are many questions potentially answerable and no attempt is made in our simulation studies to be exhaustive; rather particular parameter configurations under differing models of evolution have been selected to best illustrate discrepancies with the ND model. In this section an extension to the ND model is tentatively proposed in the form of a data simulation procedure which gives flexibility in specifying the labelled history, the rationale being that if data simulated under an alternative evolutionary setting were analysed under the ND model, then discrepancies with the model ought to be detectable.

Let's consider a set of populations and the evolutionary path they may have taken through time. The assumptions underlying the ND model imply a simultaneous divergence from an ancestral population and subsequent independent evolution due to genetic drift (see Figure 2-1). This historical picture is plausible but by no means the only possibility. Figure 2-5 displays the two possible topologies for four populations given that the tree must be bifurcating. If correlations between populations exist and there is reason to believe that the source of the correlation is due to shared ancestry, with subsequent isolated evolution, then the topologies in Figure 2-5 can be used to represent such relationships.

To model either of the topologies in Figure 2-5 using the probability structure under the ND model another layer must be added to the hierarchy, to incorporate the theoretical proto-populations found en route from the MRCAP to the contemporary populations.

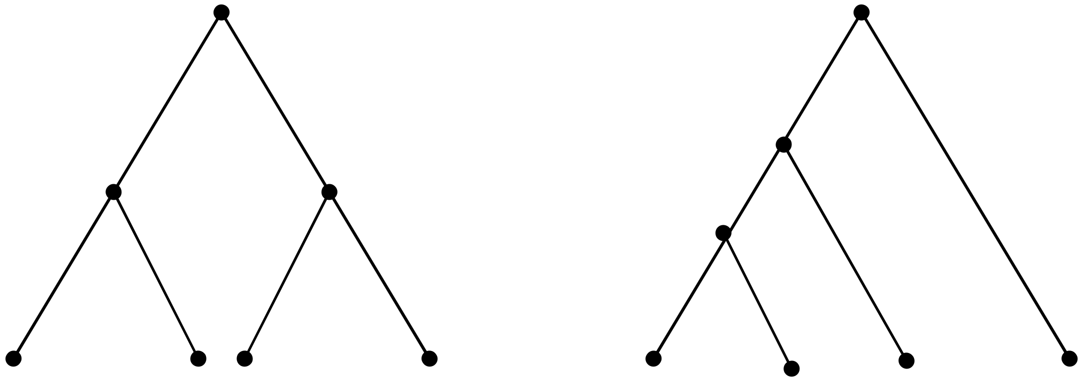


Figure 2-5 Two bifurcating tree topologies for four populations.

Two data simulation procedures are now proposed. The first method simulates under the ND model whereas the second uses the probabilistic assumptions of the ND model while allowing for different topologies and thus labelled histories to be specified. Note that all the simulations were performed in R.

ND Model Simulation Procedure

Let P = number of populations L = number of SNPs. Note that c_i and n_{ij} are fixed in advance.

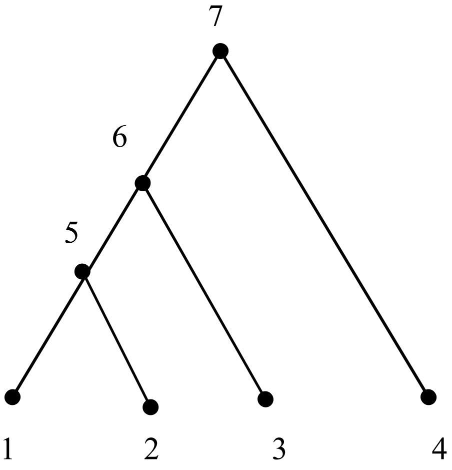
1. Draw π_j independently from a $Be(2, 2)$ where $j = 1, \ldots, L$.

2. Draw $\beta_{ij} | \pi_j, c_i$ independently from a $\mathrm{Normal}(\pi_j, c_i \pi_j (1 - \pi_j))$ where $i = 1, \ldots, P$; $j = 1, \ldots, L$.

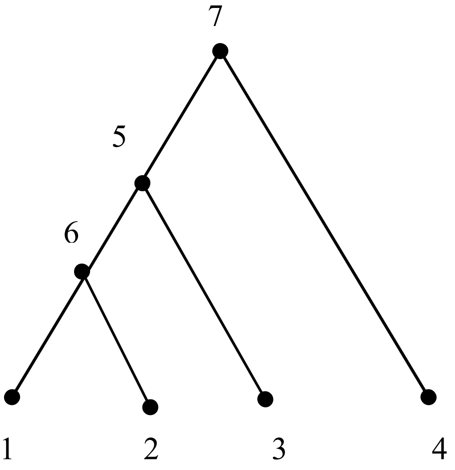
3. Draw $x_{ij} | \beta_{ij}, n_{ij}$ independently from a $\mathrm{Binomial}(n_{ij}, t(\beta_{ij}))$ where $i = 1, \ldots, P$; $j = 1, \ldots, L$.

ND Model + Topology Simulation Procedure

First we stipulate a method for labelling any given bifurcating tree with P contemporary populations. Contemporary populations are labelled from 1, . . . , P, proto-populations are labelled from P+1, . . . , 2P–2 and the MRCAP is labelled 2P–1. Figure 2- illustrates examples of the labelling method. Then we introduce a vector a of length 2P–1, given that the tree is bifurcating, whose kth element is the population ancestral to population k where k = 1, . . . , 2P–1.



a = (5, 5, 6, 7, 6, 7, 0)



a = (6, 6, 5, 7, 7, 5, 0)

Figure 2-6 Two labelled histories for four populations under an alternative evolutionary topology with corresponding a vectors which specify the ancestral relationships.

Figure 2-6 shows two labelled histories for four populations along with the corresponding a vectors. Both labelled histories represent the same model only with different labelling. Populations 5 and 6 are theoretical populations and so the labelling is arbitrary; the only stipulation is that these populations are labelled P+1, . . . , 2P–2, irrespective of order. The MRCAP is population 2P–1 and so a(2P–1) is defined to be zero.

In terms of notation, a distinction between allele frequencies in the sampled populations and in the MRCAP was made in the ND model, the former being labelled π and the latter α. Within the new simulation procedure, a common parameter α is defined and the labelling method described above is used to distinguish between sampled populations, intermediate populations and the MRCAP. The re-parameterisation discussed in section 2.1.2.3 is again used to transform α onto the real-line, and using the function $t(x)$ to define the relationship $\alpha = t(\beta)$.

The simulation procedure is as follows:

1. Draw β_{ij} independently from Beta(2, 2) where $i = 2P–1$ and $j = 1, \dots, L$.

2. Draw $\beta_{ij} \big| \beta_{a(i),j}, c_i$ independently from Normal$\left(t(\beta_{a(i),j}), c_i t(\beta_{a(i),j})\left(1 - t(\beta_{a(i),j})\right)\right)$ where $i = 1, \dots, 2P–2$.

3. Draw $x_{ij} \big| \beta_{ij}, n_{ij}$ independently from a Binomial$\left(n_{ij}, t(\beta_{ij})\right)$ where $i = 1, \dots, P$; $j = 1, \dots, L$.

There is an issue with the order of simulation in step 2 above as the order depends on the tree configuration. For example, in the left of Figure 2-6, populations 4 and 6 would be simulated first followed by populations 3 and 5 and finally populations 1 and 2. This has been handled accordingly when producing simulated data under this model.

2.3 MCMC Estimation – Some Properties

In this section some simulated data will be analysed using the MCMC algorithm discussed in previous sections with the intention of highlighting characteristics of the estimation procedure affected by choices made prior to the analysis, namely the variance of the proposal distribution and its effect on mixing, and data volume on the precision of estimates. Properties of the estimates of parameters describing allele frequencies (π, β) are also illustrated and discussed. These examples are merely illustrative, although representative examples were chosen.

Example 1

These examples illustrate the effect of the proposal variance on efficiency. Data were simulated under the ND model from $P = 4$ populations at $L = 100$ SNPs with sample size $n_{ij} =$ 100 chromosome copies per population. The c parameters were set to distinct values, where $c_1 = 0.05$, $c_2 = 0.10$, $c_3 = 0.15$ and $c_4 = 0.20$. The algorithm was run for 5000 iterations with a burn-in period of 500. The starting values of π and α were set to their true values whereas the c's were all started from the true mean of all the populations, 0.12, which is essentially F_{ST}. This same value was used for the prior mean on ln(c), so that $\mu_c = \ln(0.12) =$ -2.1 and $\sigma_c^2 = 8$, specifying a distribution with large variation, in effect a very uninformative prior. A uniform(0,1) prior on π was used throughout these examples. Note that only the estimates of the c parameters have been considered in these examples as they are the parameters we focus on. Trace plots are a useful way of diagnosing any problems regarding the mixing of the chain and have been used in this section.

The draws in Figure 2-7 appear to be quite strongly correlated, especially for c_1, the consequence of a proposal variance that is too small. When the proposal variance is too small, proposed values tend to be very close to the current value and so are more likely to be accepted. This is undesirable since the chain moves around slowly making small steps each time. Eventually the chain should converge to the target distribution but in a far from efficient manner. The acceptance rates in Table 2 for the proposal variance $\sigma_3^2 = 0.05$ are far too high at around 90%. Gelman (2004a) suggest an acceptance rate of around 40% when parameters are updated in batches; however this is rule of thumb and should be interpreted with caution. The opposite is true of the draws in Figure 2-8. The proposal variance is too high resulting in very low acceptance rates (≈ 6%,Table 2). The trace plot is stationary on the c-axis for long periods then large jumps are made when a proposal is accepted. This is again undesirable as an unsatisfactory portion of the c-space is covered by the chain casting doubt over the representation of the posterior distribution by the simulation draws.

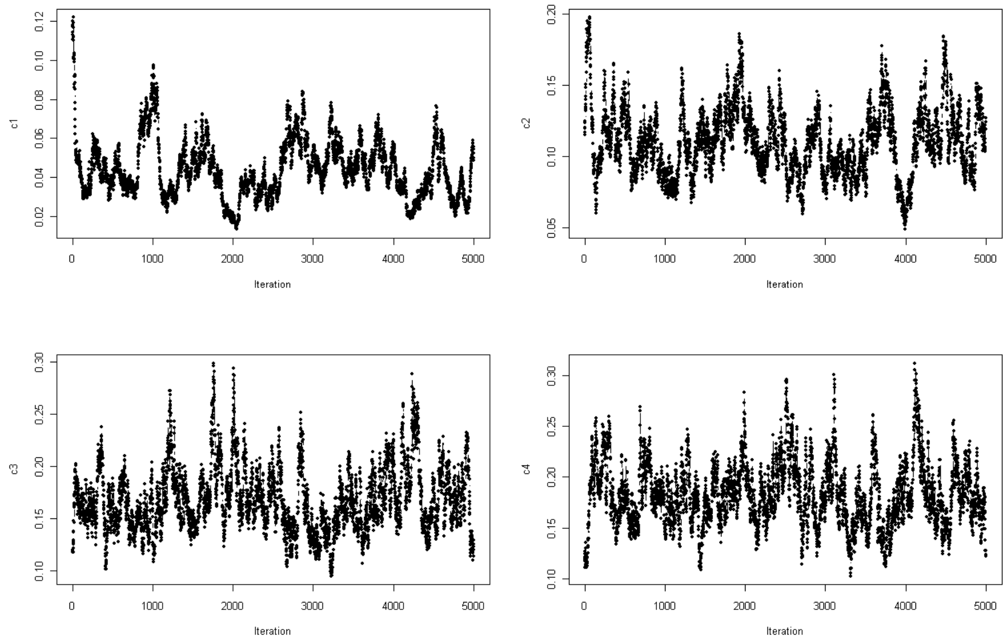


Figure 2-7 Trace plots of an MCMC run of 5000 iterations without removing burn-in, P=4, L=100, n=100, with a proposal variance = 0.05.

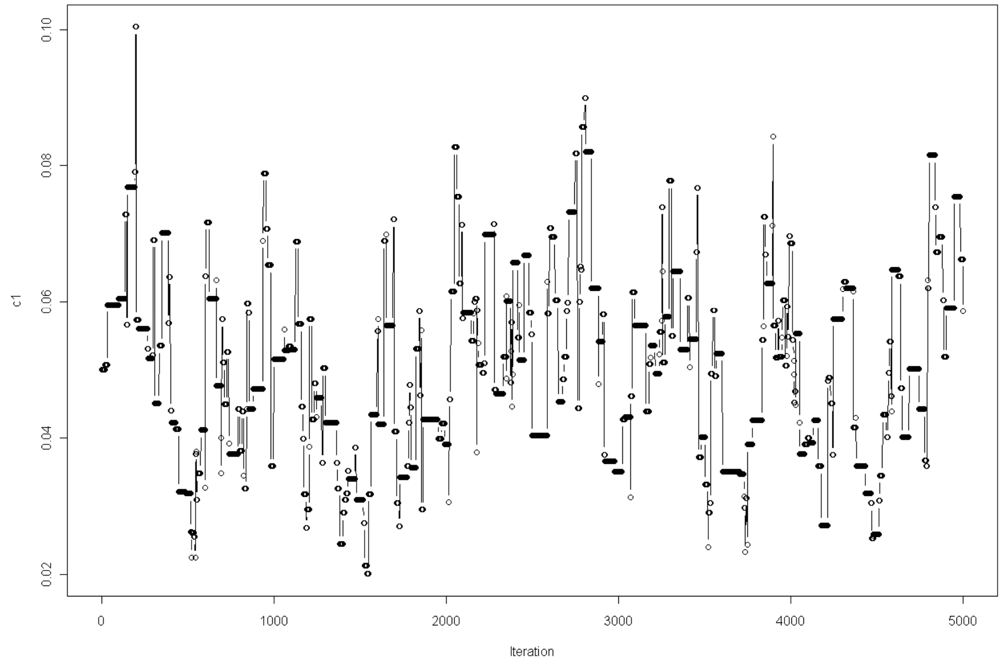


Figure 2-8 Trace plot of an MCMC run of 5000 iterations without removing burn-in, P=4, L=100, with a proposal variance = 3. Only the chain for c_1 is shown.

An ideal proposal variance would strike a balance between the two extremes shown so far. Figure 2-9 illustrates the properties of a simulation where the proposal variance is set at a suitable value. The trace plots are stable in that they do not tend to make regular large jumps, the draws cover a sufficient portion of the parameter space, and proposed values are accepted around 40% of the time (Table 2). Note that finding such a value is essentially a trial and error exercise. Methods have been developed, generally known as adaptive MCMC (Roberts and Rosenthal, 2009), to make the choice of proposal variance (or tuning parameters using their terminology) an automatic process. Table 2 displays the results from the three runs discussed. To summarise these simulations, means after discarding burn-in, acceptance rates for single parameters and 90% credible regions were used. The credible regions were calculated using the 5% and 95% quantiles of the draws not including burn-in. The main point to note is that all the credible regions contain the true value which is reassuring. In fact all the point estimates are close to the true value. Therefore in the two examples where the proposal variance was unsuitable, the location of the inferred posterior distribution was not skewed. However the draws may not have been representative of the shape of the distribution given the same number of iterations compared to the example using a better proposal variance.

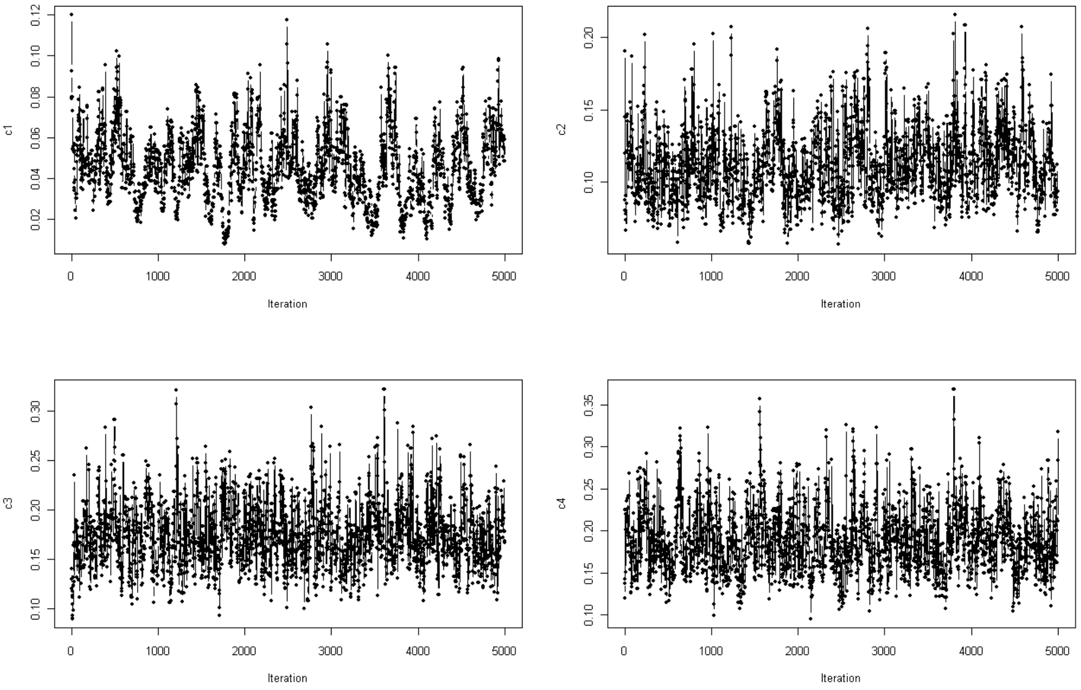


Figure 2-9 Trace plots of an MCMC run of 5000 iterations without removing burn-in, *P*=4, *L*=100, with a proposal variance = 0.4.

Table 2 MCMC results from independent runs varying the proposal variance of the drift parameters.

Parameter	*Actual Value*	σ_3^2 = 0.05			σ_3^2 = 0.4			σ_3^2 = 3		
		Mean	*Acc. Rate*	*90% Cred. Reg*	*Mean*	*Acc. Rate*	*90% Cred. Reg*	*Mean*	*Acc. Rate*	*90% Cred. Reg*
c_1	*0.05*	*0.0438*	*0.8866*	*(0.0226, 0.0708)*	*0.0452*	*0.4038*	*(0.0192, 0.0745)*	*0.0426*	*0.0538*	*(0.0215, 0.0710)*
c_2	*0.10*	*0.1093*	*0.8914*	*(0.0747, 0.1513)*	*0.1113*	*0.3876*	*(0.0766, 0.1565)*	*0.1135*	*0.0588*	*(0.0827, 0.1536)*
c_3	*0.15*	*0.1678*	*0.8936*	*(0.1242, 0.2244)*	*0.1733*	*0.3912*	*(0.1285, 0.2316)*	*0.1755*	*0.0622*	*(0.1234, 0.2304)*
c_4	*0.20*	*0.1826*	*0.8938*	*(0.1355, 0.2378)*	*0.1873*	*0.3974*	*(0.1347, 0.2574)*	*0.1900*	*0.0624*	*(0.1334, 0.2664)*

Note: 5000 simulations – 500 burn-in, P = 4, L = 100, n_{ij}= 100. σ_3^2 is the variance of the normal proposal distribution for c.

Example 2

The second set of examples highlights the effect of data volume on the precision of estimates. Data volume can be varied in the number of SNPs used and the number of individuals in each population. The location and variability of posterior distributions are summarised numerically by the mean and posterior standard deviation (p.s.d) respectively, excluding burn-in and graphically using posterior density plots.

(a) To illustrate the effect of the number of SNPs used, data were simulated under the ND model from P = 4 populations at L = 5, 15, 25, 50, 100, 200 SNPs with sample size n_{ij} = 100 chromosome copies per population. The c parameters were again set to unique values where c_1 = 0.05, c_2 = 0.10, c_3 = 0.15 and c_4= 0.20. The algorithm was run for 5000 iterations with a burn-in period of 500. The same configurations of priors and initial values used in the previous example were used again.

Table 3 shows the results of the three runs with L = 5, 15, 25. The first thing to note is the poor performance of the estimation procedure when L = 5 or 15. When L = 5 the p.s.d's are extremely large resulting in wide credible regions, although in some cases the credible regions do not include the true value. When L = 15 the situation is slightly improved, however, estimates are still fairly poor with p.s.d's that are too large. There is a marked improvement when L = 25 as all credible regions include the true value and the p.s.d's are between 0.02 and 0.08.

Table 3 MCMC results from independent runs varying SNP volume.

Parameter	*Actual Value*	***L = 5***			***L = 15***			***L = 25***		
		Mean	***p.s.d***	***90% Cred.Reg***	***Mean***	***p.s.d***	***90% Cred.Reg***	***Mean***	***p.s.d***	***90% Cred.Reg***
c_1	*0.05*	*0.0966*	*0.2118*	$(1.5\times10^{-7}, 0.0401)$	*0.0331*	*0.0432*	$(2.0\times10^{-5}, 0.0128)$	*0.0170*	*0.0257*	$(1.4\times10^{-5}, 0.0775)$
c_2	*0.10*	*0.2804*	*0.5994*	$(2.1\times10^{-5}, 1.1682)$	*0.2636*	*0.1365*	*(0.1018, 0.5293)*	*0.1082*	*0.0520*	*(0.0414, 0.1980)*
c_3	*0.15*	*0.2483*	*0.6939*	*(0.0007, 0.8004)*	*0.0773*	*0.0554*	*(0.0120, 0.1809)*	*0.1363*	*0.0544*	*(0.0678, 0.2318)*
c_4	*0.20*	*0.2687*	*0.3889*	$(3.2\times10^{-5}, 0.0904)$	*0.4271*	*0.2186*	*(0.1708, 0.8345)*	*0.2069*	*0.0815*	*(0.1090, 0.3559)*

Note: 5000 simulations – 500 burn-in, P = 4, L = number of SNPs = *5, 15, 25,* n_{ij}*= 100.*

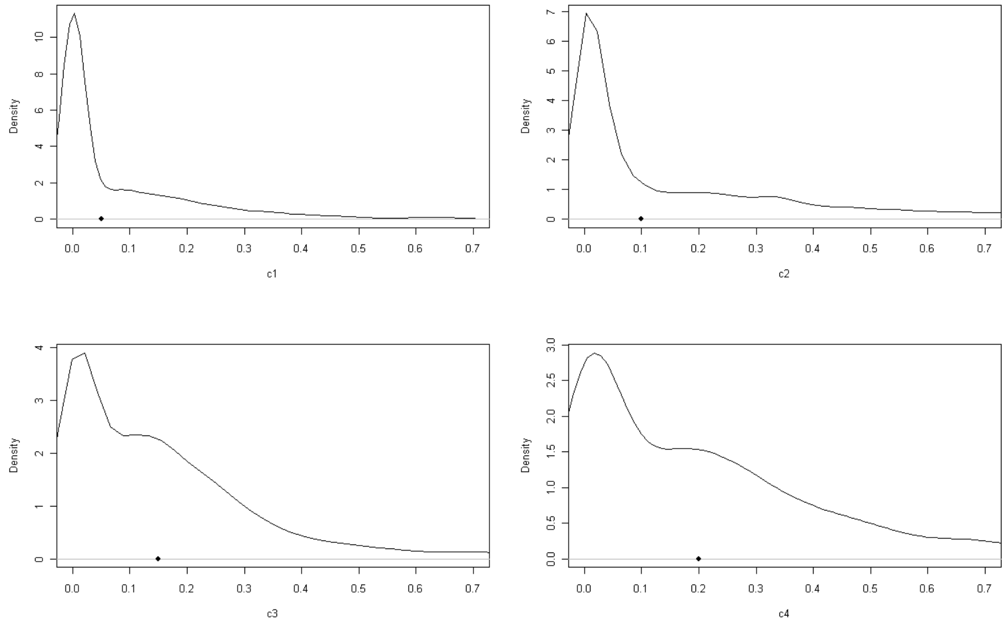


Figure 2-10 Posterior density plots estimated by an MCMC run of 5000 iterations and a burn-in of 500 where P=4, L=5 and n_{ij} = 100**.** Dot indicates the true value.

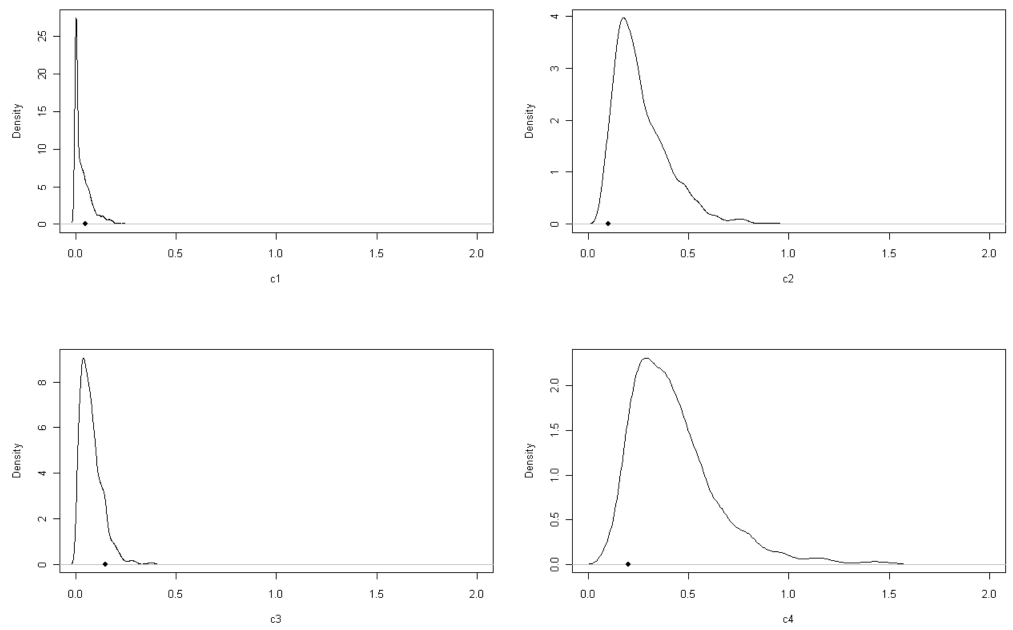


Figure 2-11 Posterior density plots estimated by an MCMC run of 5000 iterations and a burn-in of 500 where P=4, L=15 and n_{ij} = 100**.** Dot indicates the true value.

The density plots in Figure 2-10 (when L = 5) are highly skewed with extremely large variation. In fact the ranges of c_2 and c_3 reach values greater than 10, despite their true values being 0.10 and 0.15 respectively. Again it is observed that the posterior means are quite distant from their true value for all of the c's. In Figure 2-11(when L = 15) the range of values of the c's are less extreme than when L = 5, but there is still a tendency for the chain to reach values distant from the true value and this is reflected in the skewness of the density plots and the large p.s.d's. Overall there is a definite improvement when L is increased from 5 to 25.

The estimated posterior distributions in Figure 2-12 (when L = 25) show much less skewness and extreme values are not found. The means of the distributions are also close to their true values highlighting the improvement made when increasing the number of SNPs to 25. In contrast with Table 3, the p.s.d's in Table 4 (L = 50, 100, 200) are smaller. As we would hope, the credible regions do contain the true values in all cases and most estimates are close to the true value. Also worth noting is that there is still an increase in precision as SNP volume is increased, albeit by smaller increments than with the smaller data sets summarised in Table 3.

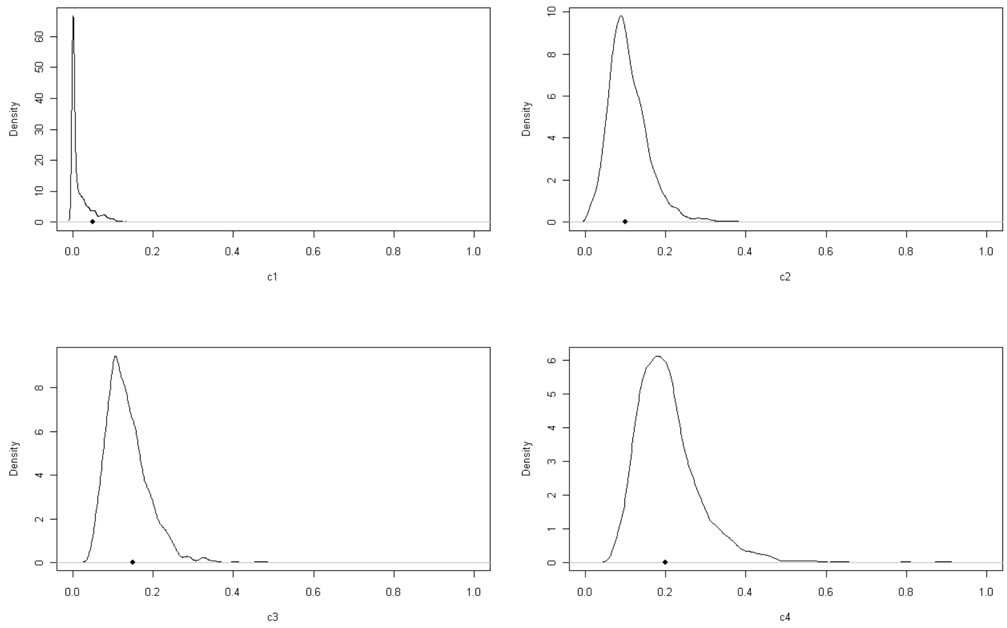


Figure 2-12 Posterior density plots estimated by an MCMC run of 5000 iterations and a burn-in of 500 where P=4, L=25 and n_{ij} = 100. Dot indicates the true value.

Table 4 MCMC results from independent runs varying SNP volume.

Parameter	*Actual Value*	***L = 50***			***L = 100***			***L = 200***		
		Mean	*p.s.d*	*90% Cred.Reg*	*Mean*	*p.s.d*	*90% Cred.Reg*	*Mean*	*p.s.d*	*90% Cred.Reg*
c_1	*0.05*	*0.0368*	*0.0180*	*(0.0129, 0.0716)*	*0.0602*	*0.0166*	*(0.0360, 0.0906)*	*0.0576*	*0.0119*	*(0.0394, 0.0781)*
c_2	*0.10*	*0.0874*	*0.0263*	*(0.0507, 0.1377)*	*0.0755*	*0.0173*	*(0.0516, 0.1066)*	*0.1072*	*0.0175*	*(0.0810, 0.1380)*
c_3	*0.15*	*0.1599*	*0.0425*	*(0.1016, 0.2377)*	*0.1480*	*0.0284*	*(0.1052, 0.1987)*	*0.1507*	*0.0216*	*(0.1183, 0.1870)*
c_4	*0.20*	*0.2207*	*0.0625*	*(0.1399, 0.3396)*	*0.2178*	*0.0417*	*(0.1570, 0.2912)*	*0.2323*	*0.0330*	*(0.1841, 0.2909)*

Note: 5000 simulations – 500 burn-in, P = 4, L = number of SNPs = *50, 100, 200,* n_{ij}= *100.*

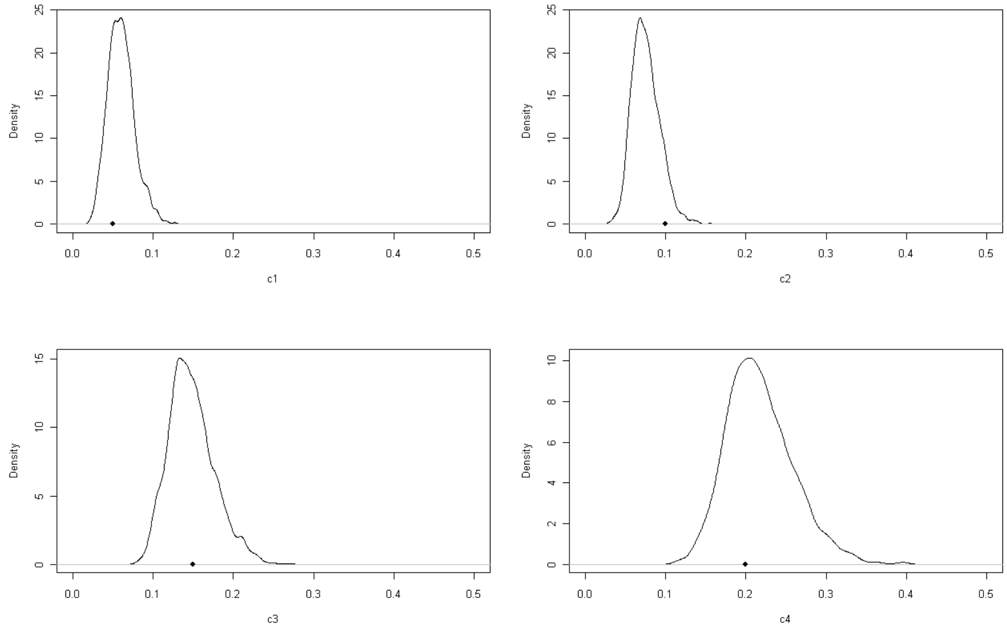


Figure 2-13 Posterior density plots estimated by an MCMC run of 5000 iterations and a burn-in of 500 where P=4, L=100 and n_{ij} = 100. Dot represents the true value.

Only a single example of the characteristics of posterior distributions estimated using a sufficient number of SNPs is shown (Figure 2-13, L = 100). All the distributions resemble the normal curve but exhibit less variation as the number of SNPs increase. Overall, when SNP volume is very small, estimated posterior distributions are skewed with high variability and point estimates of location are unreliable. On the other hand, if the number of SNPs exceeds approximately 50, then estimates are likely to be reliable and posterior distributions without extreme variation should be inferred.

(b) In this example the effect of sample size on precision is explored by altering n_{ij}. Note that since individuals have pairs of chromosomes, n_{ij} must be an even number. SNP data were simulated under the ND model from P = 4 populations at L = 100 SNPs, with sample size n_{ij} = 10, 26, 50 chromosome copies. All other parameters were as before.

In Table 5 the first point to note is that when n_{ij} = 10 the point estimates are in some cases very distant from the true value although having reasonably small p.s.d's. This is unsurprising since, in a simulation setting, when n_{ij} is small, this corresponds to drawing from a binomial distribution with large variability such that the proportion x_{ij} / n_{ij} is potentially less representative of the population proportion α_{ij} (for fixed n_{ij}). The observation that the standard errors are fairly small is due to the adequate number of SNPs included in the simulation. There is a marked improvement when n_{ij} is increased to 26 in both the location and spread of the posterior distribution, and again when $n_{ij} = 50$, as all the credible regions are centred near the true value. There is a slight discrepancy in that the standard error for c_1 is smaller when $n_{ij} = 10$ than when $n_{ij} = 26$ and 50. This is due to the tendency, when the true c is small and the sample size is not sufficient, of the chain to get stuck at very small values without moving very often, resulting in a reduced standard error while the point estimate is deflated from its true value. This property is reflected in the skewed distribution for c_1 in Figure 2-14, and is an example of poor mixing.

Table 5 MCMC results from independent runs varying sample size.

Parameter	*Actual Value*	$n_{ij} = 10$			$n_{ij} = 26$			$n_{ij} = 50$		
		Mean	*p.s.d*	*90% Cred. Reg*	*Mean*	*p.s.d*	*90% Cred. Reg*	*Mean*	*p.s.d*	*90% Cred. Reg*
c_1	*0.05*	*0.0154*	*0.0208*	*(0.0004, 0.0562)*	*0.0631*	*0.0259*	*(0.0286, 0.1036)*	*0.0704*	*0.0217*	*(0.0402, 0.1115)*
c_2	*0.10*	*0.2891*	*0.0834*	*(0.1721, 0.4460)*	*0.0926*	*0.0286*	*(0.0514, 0.1457)*	*0.1037*	*0.0256*	*(0.0681, 0.1495)*
c_3	*0.15*	*0.2063*	*0.0618*	*(0.1166, 0.3133)*	*0.1548*	*0.0372*	*(0.1006, 0.2238)*	*0.1242*	*0.0264*	*(0.0873, 0.1708)*
c_4	*0.20*	*0.0781*	*0.0554*	*(0.0962, 0.2733)*	*0.2044*	*0.0442*	*(0.1424, 0.2876)*	*0.1902*	*0.0395*	*(0.1354, 0.2603)*

Note: 5000 simulations – 500 burn-in, $P = 4$, n_{ij} = chromosome copies per population = 10, 26, 50, $L = 100$.

Again one can see in the plots in Figure 2-14 that the posterior distributions do not have excessive variation but the location of the distributions is not satisfactory. However, as pointed out earlier, there is a definite improvement when the sample size is increased to 26 and then 50 in both the location and the variability of the estimated posterior distributions, as reflected in Figures 2-15 and 2-16.

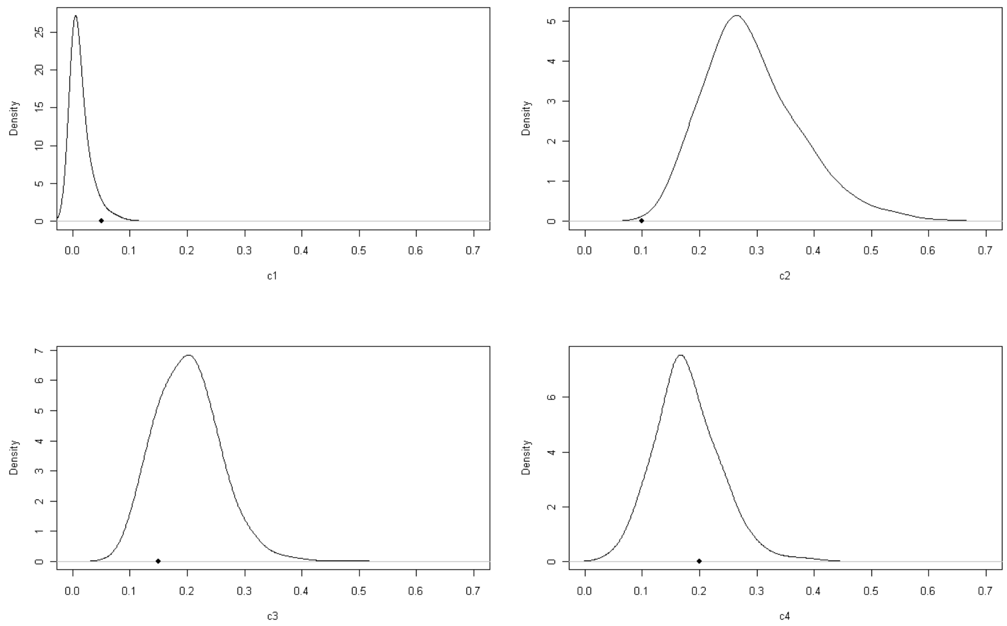


Figure 2-14 Posterior density plots estimated by an MCMC run of 5000 iterations and a burn-in of 500 where P=4, L=100 and n_{ij} = 10. Dot represents the true value.

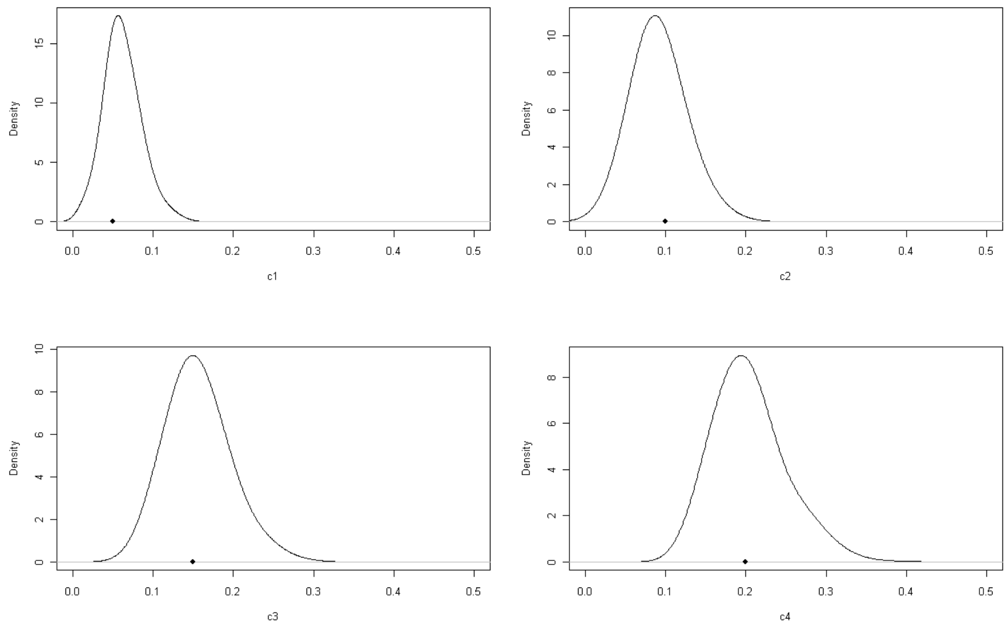


Figure 2-15 Posterior density plots estimated by an MCMC run of 5000 iterations and a burn-in of 500 where P=4, L=100 and n_{ij} = 26. Dot represents the true value.

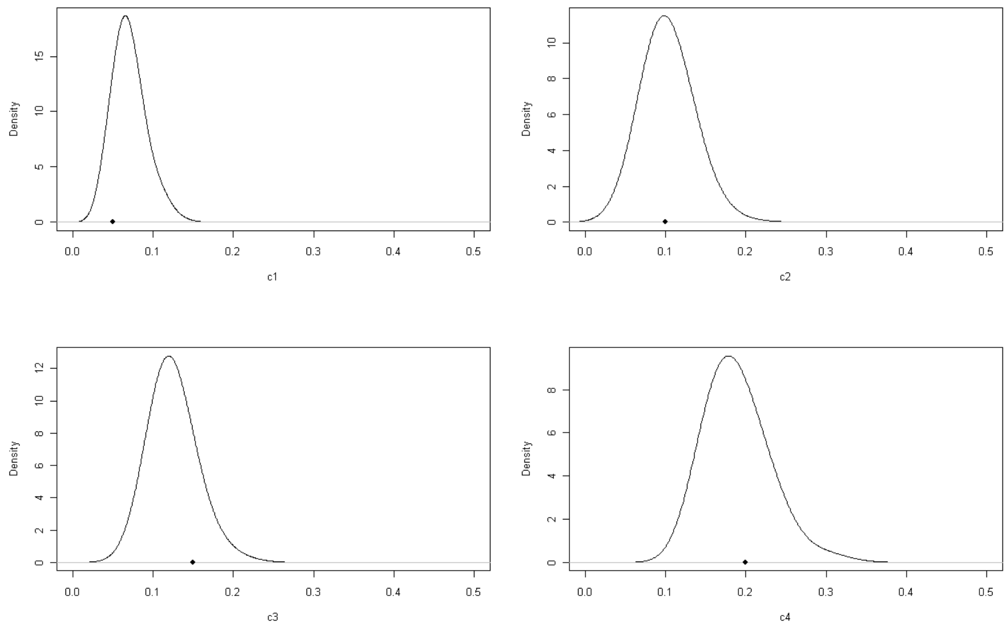


Figure 2-16 Posterior density plots estimated by an MCMC run of 5000 iterations and a burn-in of 500 where P=4, L=100 and n_{ij} = 50. Dot represents the true value.

Table 6 MCMC results from independent runs varying sample size.

Parameter	***Actual Value***	***n_{ij}= 100***			***n_{ij}= 150***			***n_{ij} =200***		
		Mean	***p.s.d***	***90% Cred. Reg***	***Mean***	***p.s.d***	***90% Cred. Reg***	***Mean***	***p.s.d***	***90% Cred. Reg***
c_1	*0.05*	*0.0537*	*0.0158*	*(0.0303, 0.0816)*	*0.0415*	*0.0131*	*(0.0215, 0.0646)*	*0.0447*	*0.0128*	*(0.0258, 0.0680)*
c_2	*0.10*	*0.1108*	*0.0243*	*(0.0754, 0.1548)*	*0.0736*	*0.0166*	*(0.0500, 0.1040)*	*0.0995*	*0.0198*	*(0.0700, 0.1366)*
c_3	*0.15*	*0.1688*	*0.0327*	*(0.1203, 0.2280)*	*0.1646*	*0.0321*	*(0.1172, 0.2238)*	*0.1491*	*0.0273*	*(0.1079, 0.1973)*
c_4	*0.20*	*0.2400*	*0.0451*	*(0.1737, 0.3211)*	*0.2319*	*0.0431*	*(0.1703, 0.3088)*	*0.2399*	*0.0450*	*(0.1785, 0.3201)*

Note: 5000 simulations – 500 burn-in, P = 4, n_{ij}= chromosome copies per population = 100, 150, 200, L = 100.

Another set of analyses are summarised in Table 6 (n = 100, 150, 200) to highlight the plateau reached in precision when the sample is increased to large values. Notice that there is not a clear decreasing trend in p.s.d's for the three runs in Table 6. Of course, if many simulated data sets were analysed and aggregated one would expect to see an increase in precision as the sample size increases but the effect is not pronounced. The plateau in the improvement of precision is simply because most of the information about c is retrieved from the allele frequencies across SNPs. Therefore, once the β's are sufficiently estimated by the sample frequencies, very little extra information is extracted by increasing the sample size further.

Example 3

In this example data simulated under the ND model from $P = 4$ populations at $L = 100$ SNPs with sample size $n_{ij} = 100$ chromosome copies per population were analysed and a single SNP was chosen to illustrate properties of the MCMC algorithm when estimating π and β. The same starting configurations were used as in the previous examples.

Table 7 MCMC results for π and β from a single SNP j

Parameter	*Actual Value*	*Mean*	*p.s.d*	*90% Cred. Reg.*	*Acc. Rate*
π_j	0.4581	0.4106	0.0791	(0.2792, 0.5419)	0.3744
β_{1j}	0.4241	0.3844	0.0474	(0.3068, 0.4653)	0.4116
β_{2j}	0.4661	0.4654	0.0473	(0.3868, 0.5439)	0.4608
β_{3j}	0.4104	0.4320	0.0481	(0.3442, 0.5141)	0.4804
β_{4j}	0.2365	0.2417	0.0408	(0.1780, 0.3121)	0.4576

Note: 5000 simulations – 500 burn-in, $P = 4$, $L = 100$, $n_{ij} = 100$.

From Table 7 notice that β_{ij} ($i = 1, 2, 3, 4$) is estimated with greater precision than π_j i.e. the estimates have a smaller standard error. This is since β_{ij} is estimated well by x_{ij} / n_{ij}, whereas information about π_j comes from the β_{ij}'s, of which there are only four in this case, and these are estimated themselves. This property is also mirrored in the density plots in Figure 2-17 as the distribution for π_j has slightly more variability than those of the β_{ij}'s. Also notice that the estimation procedure does rather well when estimating both the π_j and the β_{ij}'s since all the distributions are centred on the true values.

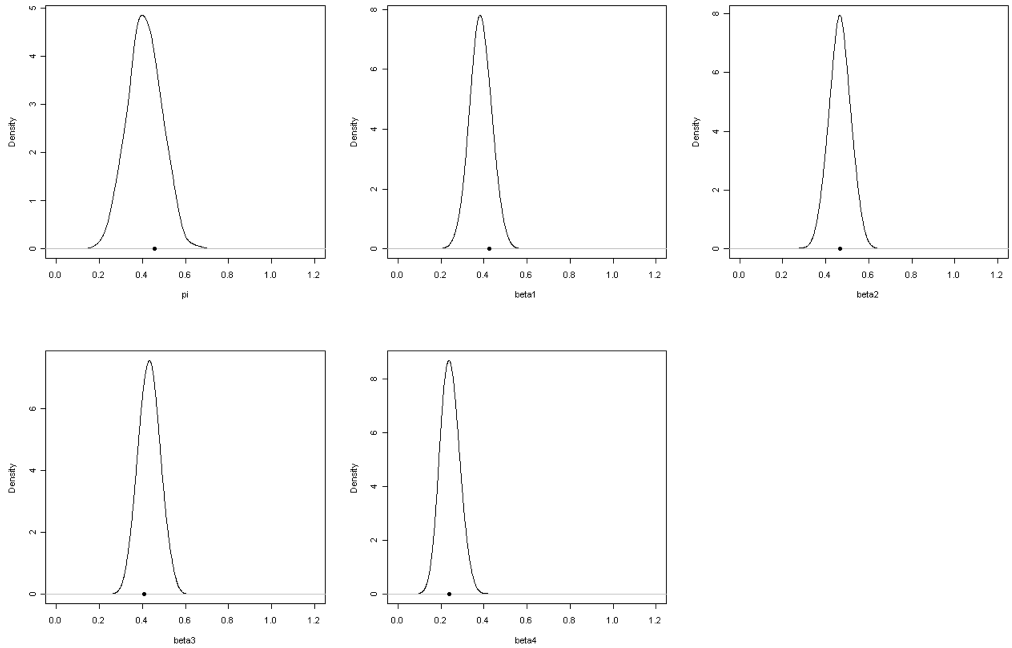


Figure 2-17 Posterior density plots for $\boldsymbol{\pi_j}$ and $\boldsymbol{\beta_{ij}}$ (i = 1, . . 4) for a single SNP j estimated by an MCMC run of 5000 iterations and a burn-in of 500 where P=4, L=100 and n_{ij} = 100. Dot represents the true value.

2.4 ND Model Extension

Likely deviations from the modelling assumptions of the ND model in real data are due to gene-flow or shared ancestry among populations manifest in correlations between population allele frequencies (conditional on π) (Nicholson et al., 2002). If there is reason to believe that correlations are due to shared ancestry, with subsequent isolated evolution, then the topologies in Figure 2- represent some possible historical relationships between contemporary populations. The following extension to the ND model assumes the same probabilistic distributions, for the unobserved population allele frequencies and the observed SNP allele counts, as the ND model, while the hierarchical structure can be varied to capture ancestral relationships between populations.

Consider an equivalent scenario as proposed in section 2.1 where we have a sample of SNP data collected from P populations at L SNPs. Then let n_{ij} be the number of chromosomes typed in the ith population at the jth SNP ($i = 1, . . . , P; j = 1, . . . , L$). The number of copies of the chosen allele in population i at SNP j is x_{ij}, $0 \leq x_{ij} \leq n_{ij}$ ($i = 1, . . . , P; j = 1, . . . , L$).

The unobserved frequency of the chosen allele in the *i*th population at the *j*th SNP is denoted by α_{ij}, $0 \leq \alpha_{ij} \leq 1$. However $i = 1, \ldots, 2P\text{-}1$; $j = 1, \ldots, L$ since α contains contemporary populations, proto-populations and the MRCAP. It follows that $0 < \alpha_{2P-1,j} < 1$ since the MRCAP is assumed to be polymorphic at every SNP.

As in the ND model given n_{ij} and α_{ij} ($i = 1, \ldots, P$; $j = 1, \ldots, L$), x_{ij} is binomially distributed (see equation [4]). Then we introduce a vector a of length $2P–1$, given that the topologies considered are bifurcating, whose *k*th element is the population ancestral to population k, where $k = 1, \ldots, 2P–1$. We also introduce two further vectors o_1 and o_2 both of length $2P–1$ where the *k*th element of o_1 is the first descendant population of population k and the *k*th element of o_2 is the second descendant population of population k. Since the sampled populations have no descendants, $o_i(j) = 0$ where $i = 1, 2$; $j = 1, \ldots, P$. Note that o_1 and o_2 are not unique since corresponding elements can be exchanged.

The allele frequency of populations other than the MRCAP at a given SNP are modelled as

$$\alpha_{ij} \sim \text{Normal}\Big(\alpha_{a(i),j}, c_i\alpha_{a(i),j}\big(1 - \alpha_{a(i),j}\big)\Big), \qquad i = 1, \ldots, 2P - 2; j = 1, \ldots, L, \qquad [14]$$

independently $\forall\ i, j$.

As in the ND model the normal distribution has point masses at the boundaries $\alpha = 0, 1$.

To complete the hierarchy we place independent priors on $\alpha_{2P-1,j}$ and c_i :

$\alpha_{2P-1,1} \ldots, \alpha_{2P-1,L}$ are independent and identically distributed with density f; [15]

$c_1, \ldots, c_{2P-2}$ are independent and identically distributed with density g. [16]

Figure 2-18 gives an example of the hierarchical relationships between four sampled populations within the new model for a particular ancestral configuration. Notice that the c parameters are labelled by the population index at the bottom of a branch.

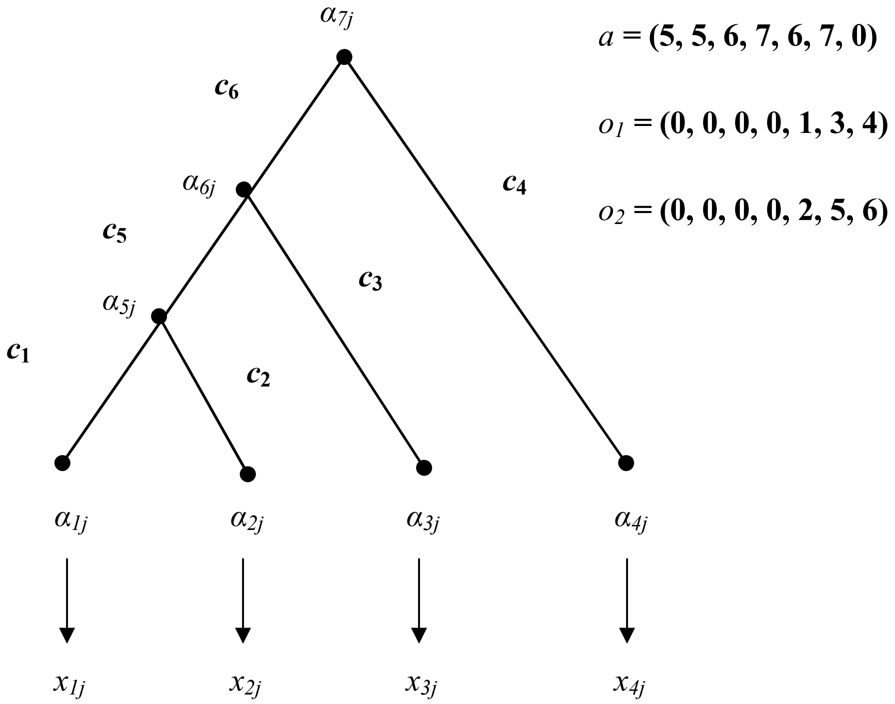


Figure 2-18 A diagrammatic representation of the new model for four populations for a given labelled history with corresponding a, o_1, o_2 vectors at a single SNP j.

2.4.1 Implementation

The Metropolis-Hastings algorithm was employed to sample from posterior distributions. The general form of the algorithm can be found in section 2.1.2; only properties relevant to the new model are discussed here. The prior distributions used are identical to those used for the ND model (see section 2.1.2.4 for details and discussion.)

As before, when calculating the ratio of densities r, useful simplifications can be made to improve the efficiency of the algorithm. Updates are again made in groups of parameters and within these groups the simplifications are made. First note that the full posterior distribution, up to a constant only depending on the data, factored into known conditional distribution is:

$$p(\alpha,c|x)\propto\left\{\prod_{i=1}^{2P-2}\prod_{j=1}^{L}p(\alpha_{ij}|\alpha_{a(i),j},c_i)\right\}\left\{\prod_{i=1}^{P}\prod_{j=1}^{L}p(x_{ij}|n_{ij},\alpha_{ij})\right\}\left\{\prod_{j=1}^{L}p(\alpha_{2P-1,j}|\mu_{\alpha_{2P-1}},\sigma^2_{\alpha_{2P-1}})\right\}\left\{\prod_{i=1}^{2P-2}p(c_i|\mu_c,\sigma_c^2)\right\}.$$

When updating c_i ($i = 1, \ldots, 2P$-2),

$$r=\frac{\left\{\prod_{j=1}^{L}p(\alpha_{ij}|\alpha_{a(i),j},c_i^*)\right\}\left\{\prod_{i=1}^{P}\prod_{j=1}^{L}p(x_{ij}|n_{ij},\alpha_{ij})\right\}\left\{\prod_{j=1}^{L}p(\alpha_{2P-1,j}|\mu_{\alpha_{2P-1}},\sigma^2_{\alpha_{2P-1}})\right\}p(c_i^*|\mu_c,\sigma_c^2)}{\left\{\prod_{j=1}^{L}p(\alpha_{ij}|\alpha_{a(i),j},c_i)\right\}\left\{\prod_{i=1}^{P}\prod_{j=1}^{L}p(x_{ij}|n_{ij},\alpha_{ij})\right\}\left\{\prod_{j=1}^{L}p(\alpha_{2P-1,j}|\mu_{\alpha_{2P-1}},\sigma^2_{\alpha_{2P-1}})\right\}p(c_i|\mu_c,\sigma_c^2)}$$

$$=\frac{\left\{\prod_{j=1}^{L}p(\alpha_{ij}|\alpha_{a(i),j},c_i^*)\right\}p(c_i^*|\mu_c,\sigma_c^2)}{\left\{\prod_{j=1}^{L}p(\alpha_{ij}|\alpha_{a(i),j},c_i)\right\}p(c_i|\mu_c,\sigma_c^2)},$$

where c_i^* is the proposed value and c_i is the current value. The parameters in α can be split into three groups: the MRCAP, the remaining proto-populations and contemporary populations.

When updating $\alpha_{k,j}$, $k = 2P{-}1$ (MRCAP),

$$r=\frac{p\left(\alpha_{o_1(k)}\middle|\alpha^*_{kj},c_{o_1(k)}\right)p\left(\alpha_{o_2(k)}\middle|\alpha^*_{kj},c_{o_2(k)}\right)\left\{\prod_{\substack{i=1\\ i\neq o_1(k)\\ i\neq o_2(k)}}^{2P-2}p\left(\alpha_{ij}\middle|\alpha_{a(i),j},c_i\right)\right\}\left\{\prod_{i=1}^{P}p\left(x_{ij}\middle|n_{ij},x_{ij}\right)\right\}p\left(\alpha^*_{kj}\middle|\mu_k,\sigma_k^2\right)\left\{\prod_{i=1}^{2P-2}p\left(c_i\middle|\mu_c,\sigma_c^2\right)\right\}}{p\left(\alpha_{o_1(k)}\middle|\alpha_{kj},c_{o_1(k)}\right)p\left(\alpha_{o_2(k)}\middle|\alpha_{kj},c_{o_2(k)}\right)\left\{\prod_{\substack{i=1\\ i\neq o_1(k)\\ i\neq o_2(k)}}^{2P-2}p\left(\alpha_{ij}\middle|\alpha_{a(i),j},c_i\right)\right\}\left\{\prod_{i=1}^{P}p\left(x_{ij}\middle|n_{ij},x_{ij}\right)\right\}p\left(\alpha_{kj}\middle|\mu_k,\sigma_k^2\right)\left\{\prod_{i=1}^{2P-2}p\left(c_i\middle|\mu_c,\sigma_c^2\right)\right\}}$$

$$=\frac{p\left(\alpha_{o_1(k)}\middle|\alpha^*_{kj},c_{o_1(k)}\right)p\left(\alpha_{o_2(k)}\middle|\alpha^*_{kj},c_{o_2(k)}\right)p\left(\alpha^*_{kj}\middle|\mu_k,\sigma_k^2\right)}{p\left(\alpha_{o_1(k)}\middle|\alpha_{kj},c_{o_1(k)}\right)p\left(\alpha_{o_2(k)}\middle|\alpha_{kj},c_{o_2(k)}\right)p\left(\alpha_{kj}\middle|\mu_k,\sigma_k^2\right)},$$

where α^*_{kj} is the proposed value and α_{kj} is the current value.

When updating α_{kj}, $k = P$+1, . . . , $2P{-}2$ (proto-populations),

$$r=\frac{p(\alpha_{kj}^{*}|\alpha_{a(k),j},c_k)p(\alpha_{o_1(k),j}|\alpha_{kj}^{*},c_{o_1(k)})p(\alpha_{o_2(k),j}|\alpha_{kj}^{*},c_{o_2(k)})\left\{\prod_{\substack{i=1\\i\neq k\\i\neq o_1(k)\\i\neq o_2(k)}}^{2P-2}p(\alpha_{ij}|\alpha_{a(i),j},c_i)\right\}\left\{\prod_{i=1}^{P}p(x_{ij}|n_{ij},\alpha_{ij})\right\}}{p(\alpha_{kj}|\alpha_{a(k),j},c_k)p(\alpha_{o_1(k),j}|\alpha_{kj},c_{o_1(k)})p(\alpha_{o_2(k),j}|\alpha_{kj},c_{o_2(k)})\left\{\prod_{\substack{i=1\\i\neq k\\i\neq o_1(k)\\i\neq o_2(k)}}^{2P-2}p(\alpha_{ij}|\alpha_{a(i),j},c_i)\right\}\left\{\prod_{i=1}^{P}p(x_{ij}|n_{ij},\alpha_{ij})\right\}}$$

$$\times\frac{p(\alpha_{2P-1,j}|\mu_{\alpha_{2P-1}},\sigma_{\alpha_{2P-1}}^{2})\left\{\prod_{i=1}^{2P-2}p(c_i|\mu_c,\sigma_c^2)\right\}}{p(\alpha_{2P-1,j}|\mu_{\alpha_{2P-1}},\sigma_{\alpha_{2P-1}}^{2})\left\{\prod_{i=1}^{2P-2}p(c_i|\mu_c,\sigma_c^2)\right\}}$$

$$=\frac{p(\alpha_{kj}^{*}|\alpha_{a(k),j},c_k)p(\alpha_{o_1(k),j}|\alpha_{kj}^{*},c_{o_1(k)})p(\alpha_{o_2(k),j}|\alpha_{kj}^{*},c_{o_2(k)})}{p(\alpha_{kj}|\alpha_{a(k),j},c_k)p(\alpha_{o_1(k),j}|\alpha_{kj},c_{o_1(k)})p(\alpha_{o_2(k),j}|\alpha_{kj},c_{o_2(k)})}$$

where α_{kj}^{*} is the proposed value and α_{kj} is the current value.

When updating α_{kj} (k =1, . . . , P, contemporary populations),

$$r=\frac{p(\alpha_{kj}^{*}|\alpha_{a(k),j},c_k)\left\{\prod_{\substack{i=1\\i\neq k}}^{2P-2}p(\alpha_{ij}|\alpha_{a(i),j},c_i)\right\}p(x_{kj}|n_{kj},\alpha_{kj}^{*})p(\alpha_{2P-1,j}|\mu_{\alpha_{2P-1}},\sigma_{\alpha_{2P-1}}^{2})\left\{\prod_{i=1}^{2P-2}p(c_i|\mu_c,\sigma_c^2)\right\}}{p(\alpha_{kj}|\alpha_{a(k),j},c_k)\left\{\prod_{\substack{i=1\\i\neq k}}^{2P-2}p(\alpha_{ij}|\alpha_{a(i),j},c_i)\right\}p(x_{kj}|n_{kj},\alpha_{kj})p(\alpha_{2P-1,j}|\mu_{\alpha_{2P-1}},\sigma_{\alpha_{2P-1}}^{2})\left\{\prod_{i=1}^{2P-2}p(c_i|\mu_c,\sigma_c^2)\right\}}$$

$$=\frac{p(\alpha_{kj}^{*}|\alpha_{a(k),j},c_k)p(x_{kj}|n_{kj},\alpha_{kj}^{*})}{p(\alpha_{kj}|\alpha_{a(k),j},c_k)p(x_{kj}|n_{kj},\alpha_{kj})},$$

where α_{kj}^{*} is the proposed value and α_{kj} is the current value.

Under the ND model and consequently the new model, when an allele is lost within a population it cannot then return, as mutations are not permitted. Another assumption of the ND model is that the ancestral population was polymorphic for a given SNP. Therefore a situation could not arise when an ancestral population exhibited no variation at a given SNP. However under the new model, ancestral populations other than the MRCAP can drift to fixation meaning that all descendant populations must be monomorphic at that particular

SNP. To deal with these complexities, the modelling assumptions are altered and a set of conditions are added to the MCMC algorithm when updating ancestral populations other than the MRCAP, to reflect these changes.

We introduce the quantity $\beta_{ij} \in \Re$ $(i = 1, \ldots, 2P-2, j = 1, \ldots, L)$ such that $\alpha_{ij} = t(\beta_{ij})$,

$$\beta_{ij} \sim \text{Normal}\big(t(\beta_{a(i),j}), c_i t(\beta_{a(i),j})(1 - t(\beta_{a(i),j}))\big) \qquad [17]$$

and

$$x_{ij} \sim \text{Binomial}\big(n_{ij}, t(\beta_{ij})\big). \qquad [18]$$

Now suppose β_{ij} $(i = P+1, \ldots, 2P-2; j = 1, \ldots, L)$ is being updated, with a proposal to move to β^*_{ij}. We introduce the following additional conditions to ensure that invalid parameter configurations cannot occur:

1. If $\beta^*_{ij} \leq 0$ and $\beta_{\text{o}_k(i),j} \neq 0\ (k = 1, 2)$, reject β^*_{ij}.

2. If $\beta^*_{ij} \geq 1$ and $\beta_{\text{o}_k(i),j} \neq 1\ (k = 1, 2)$, reject β^*_{ij}.

3. If $\beta^*_{ij} \in (0,1)$ and $\beta_{a(i),j} \notin (0,1)$, reject β^*_{ij}.

4. If $\beta^*_{ij} \in (0,1)$ and $\beta_{o_k(i),j} \notin (0,1)\ (k = 1, 2)$, reject β^*_{ij}.

2.4.2 **Proposal Distributions**

In this section the particular proposal distributions used to draw new values in the MCMC algorithm are defined. A detailed discussion of the effect proposal distributions have on the MCMC algorithm can be found in section 2.1.2.3 and illustrations can be found in section 2.3

The proposal distributions $J_t(.|.)$ at iteration *t* for β and ln(c) are:

$$J_t\left(\beta_k^t \middle| \beta_k^{t-1}\right) \sim \text{Normal}\left(\beta_k^{t-1}, \sigma_4^2\right), k = 2P - 1, \qquad [19]$$

$$J_t\left(\beta_k^t \middle| \beta_k^{t-1}\right) \sim \text{Normal}\left(\beta_k^{t-1}, \sigma_5^2\right), k = P + 1, \ldots, 2P - 2, \qquad [20]$$

$$J_t\left(\beta_k^t \middle| \beta_k^{t-1}\right) \sim \text{Normal}\left(\beta_k^{t-1}, \sigma_6^2\right), k = 1, \ldots, P, \qquad [21]$$

$$J_t\left(\ln c^t \middle| \ln c^{t-1}\right) \sim \text{Normal}\left(\ln c^{t-1}, \sigma_7^2\right). \qquad [22]$$

Notice that an additional distribution is used for ancestral populations other than the MRCAP, as the new model contains three groups of β parameters.

Chapter 3

Results

In this section both simulated and real SNP data will be analysed under the ND model of SNP allele frequencies and under the newly developed model discussed in section 2.4. What follows may be split into two parts. In the first part, data will be analysed under the ND model in both situations where it is an accurate and an inaccurate representation of the process responsible for the data. Informal diagnostics will then be used to highlight whether the ND model fits the data well in both situations. The second part focuses on the new model, discussing the difficulties one encounters when fitting the model and the proposed solutions to such problems. Finally, simulated and real data are analysed with the aim of retrieving information regarding the most appropriate tree topology for a set of populations.

3.1 Simulation under ND Model

As previously mentioned, Nicholson et al. (2002) found that in some situations, the estimates of the *c*'s in the ND model were unstable when a population was removed and the model re-fitted. This is clearly an undesirable property and so in this section we investigate whether it is inherent in the ND model in the case where the modelling assumptions are fulfilled. This exercise is effectively a preliminary to what follows but necessary to ensure that the population removal strategy can be used to highlight departures from the modelling assumptions.

In order to assess whether estimates of *c* were stable under the ND model, 100 independent data sets were analysed under the ND model and then re-analysed, removing an arbitrary population from the data set. It was of interest to see whether estimates of the *c*'s differed significantly when the full data set was considered compared to the reduced data set. The comparisons were made by calculating the difference between the draws from the two analyses at every step in the chain for corresponding c parameters, excluding burn-in, and computing 90% credible regions for the differences. This process was repeated twice, removing a different population each time. A significant difference was declared if a credible region did not contain zero. Since we were keen only to flag potential violation of the model, a large nominal Type I error rate was chosen (10%), thus increasing the power.

On inspection, one would be surprised to see the estimates of the c's affected by a population being removed, given that the modelling assumptions are valid under simulation, since the majority of the information in the data regarding c_i comes from the variation across SNPs within population i. Therefore removing a single population j ($j \neq i$) from the data should not significantly affect the estimate of c_i.

3.1.1 **Analysis**

First we simulated 100 independent SNP data sets under the ND model, each data set containing P=4 populations, typed at L=100 SNPs with sample sizes n_{ij}=100. Adhering to the simulation procedure for the ND model outlined in section 2.2, each of the 100 sets was simulated as follows:

1. Draw π_j from a $Be(2, 2)$ where $j = 1, \ldots, 100$.

2. Draw $\beta_{ij} | \pi_j, c_i$ from a Normal$(\pi_j, c_i\pi_j(1-\pi_j))$ where $i = 1, \ldots, P;\ j = 1, \ldots, L;$ $c = (0.05, 0.10, 0.15, 0.20)$.

3. Draw $x_{ij} | \beta_{ij}, n_{ij}$ from a Binomial$(n_{ij}, t(\beta_{ij}))$ where $i = 1, \ldots, P;\ j = 1, \ldots, L.$

The particular choice of c was made to reflect a situation where all populations show differing amounts of genetic drift. To put the numbers into context, a value of c = 0.05 is similar to estimates for European populations, whereas c = 0.20 would correspond to African populations (Nicholson et al., 2002).

For all MCMC analyses discussed in this section the initial values of the c's, π's and β's were set to their true values. An uninformative prior on the π's was used, namely the Un(0, 1) distribution, with a log-normal prior on the c's. Each chain was run for 10000 iterations with a burn-in of 1000.

The first analysis compares estimates of c when population 4, the most differentiated, was removed to corresponding estimates from the full data set. Therefore 3 sets of 100 intervals are calculated and are illustrated in Figure 3-1, 3-2 and 3-3. Out of the 300 credible regions in Figure 3-1, 3-2 and 3-3, only one does not contain zero, found in Figure 3-1 (one interval in Figure 3-2 only just contains zero). Therefore none of the estimates of c appear to be significantly affected by the removal of population 4 from the data set, as one would hope. This result is rather surprising since the test should, on average, reject the null hypothesis 10% of the time (given a type I error rate of 0.1), suggesting that the test may not be particularly powerful, or is not particularly well calibrated.

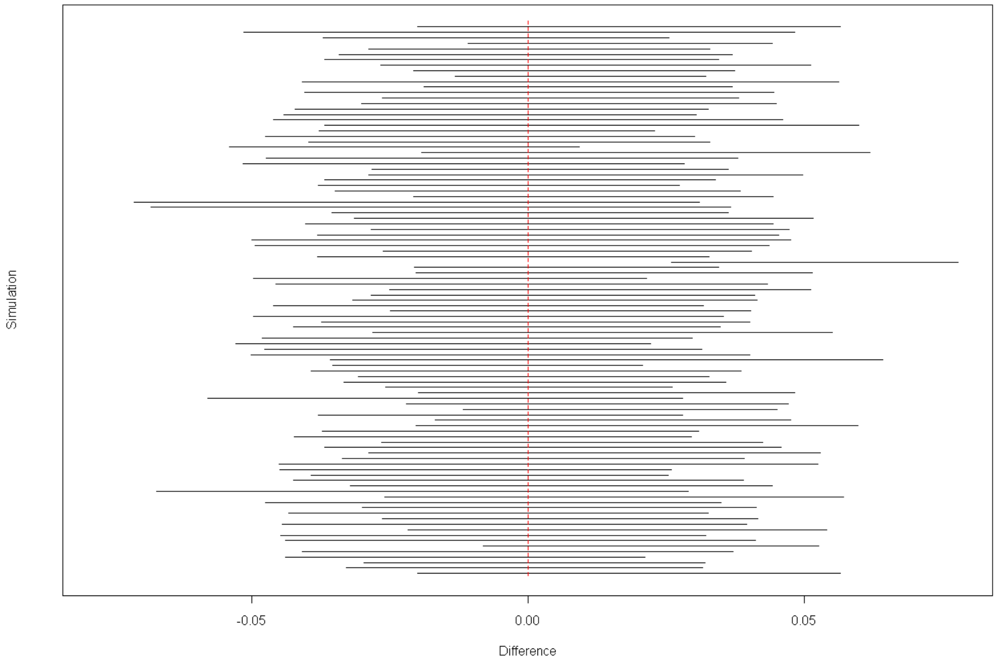


Figure 3-1 100 90% credible regions for the difference between draws of two separate MCMC analyses; one with full data set, one with population four removed from the data set. Each interval compares the estimates c_1 between the two analyses.

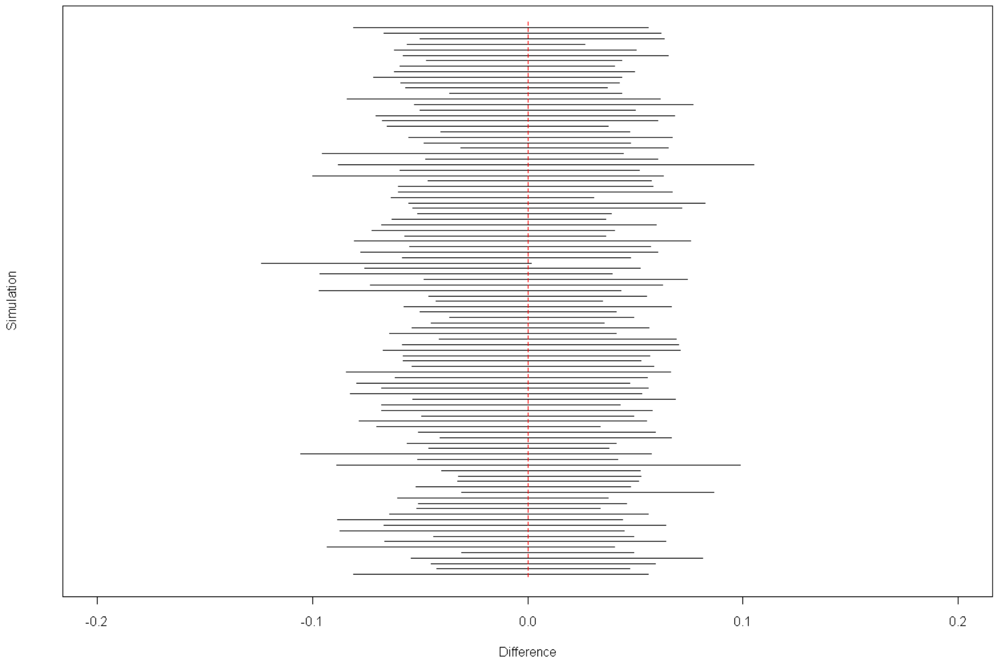


Figure 3-2 100 90% credible regions for the difference between draws of two separate MCMC analyses; one with full data set, one with population four removed from the data set. Each interval compares the estimates of $c_2 = 0.10$ between the two analyses.

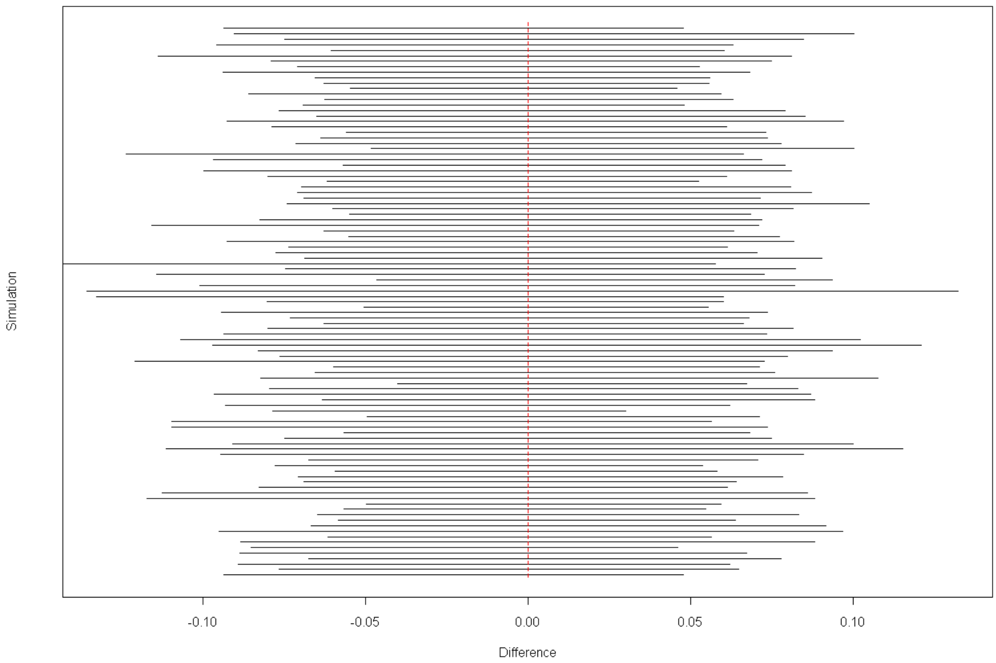


Figure 3-3 100 90% credible regions for the difference between draws of two separate MCMC analyses; one with full data set, one with population four removed from the data set. Each interval compares the estimate of $c_3 = 0.15$ between the two analyses.

In the second analysis population 1, the least differentiated, was removed from the data set and then the data re-analysed, making the appropriate comparisons. All of the credible regions in Figure 3-4, 3-5 and 3-6 contain zero and so again there is evidence to suggest that estimates of c are stable under the ND model when populations are removed from the data. When both analyses are considered in conjunction it can be concluded that estimates are robust to population removal under the ND model when the modelling assumptions are satisfied. This is particularly satisfying as it provides an informal way of testing the adequacy of the ND model.

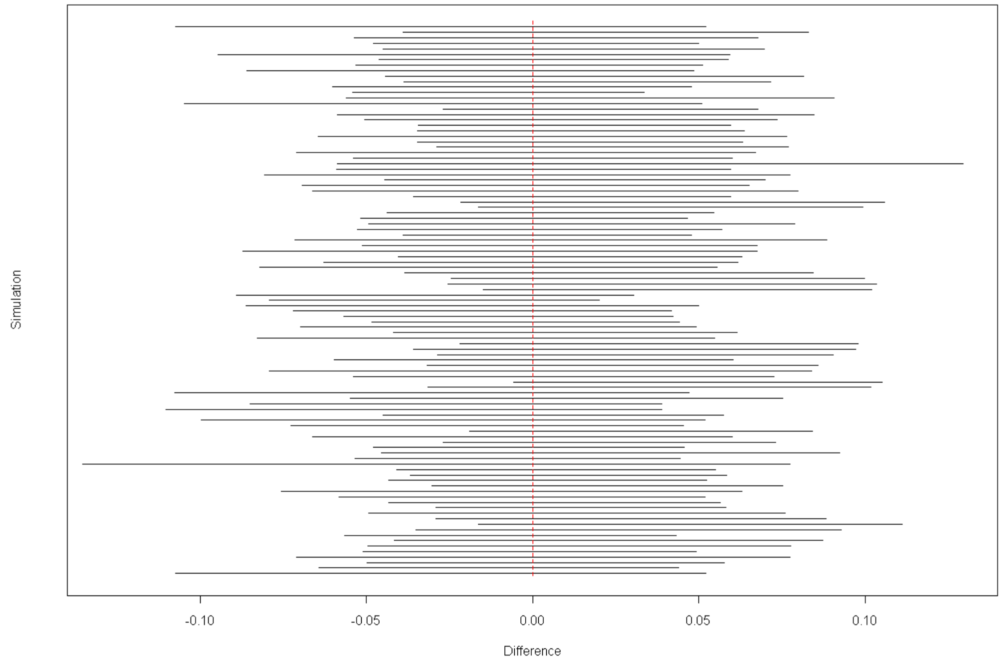


Figure 3-4 100 90% credible regions for the difference between draws of two separate MCMC analyses; one with full data set, one with population one removed from the data set. Each interval compares the estimate of $c_2 = 0.20$ between the two analyses.

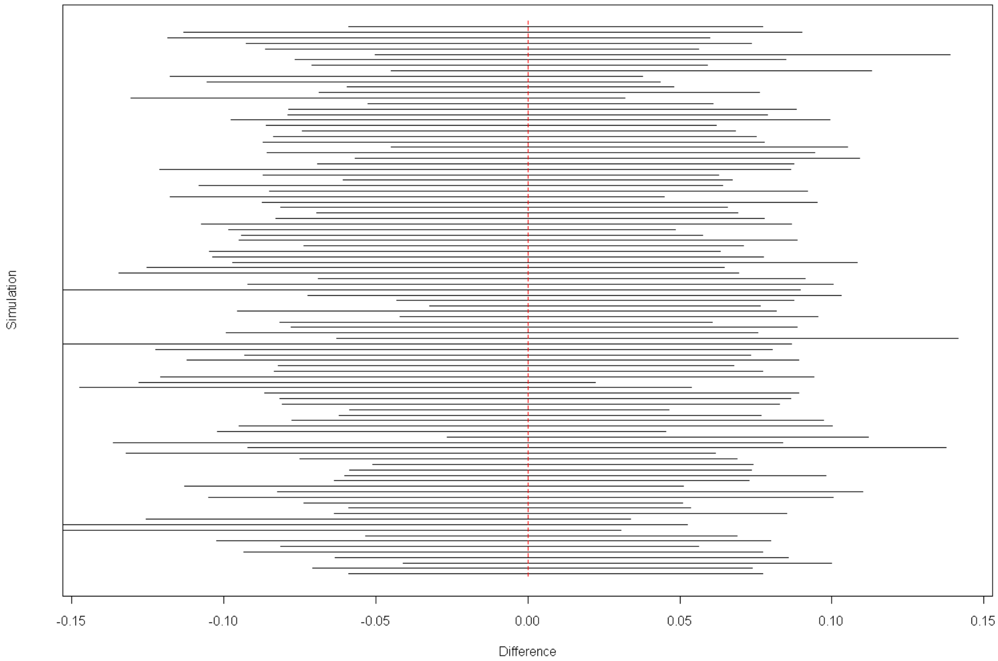


Figure 3-5 100 90% credible regions for the difference between draws of two separate MCMC analyses; one with full data set, one with population one removed from the data set. Each interval compares the estimate of $c_3 = 0.15$ between the two analyses.

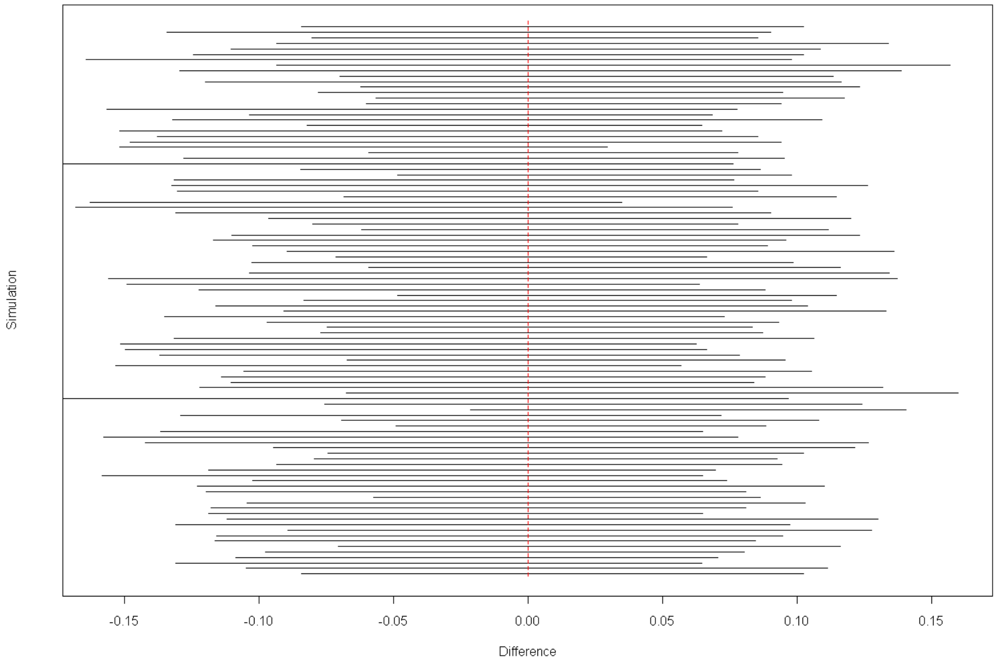


Figure 3-6 100 90% credible regions for the difference between draws of two separate MCMC analyses; one with full data set, one with population one removed from the data set. Each interval compares the estimate of $c_4 = 0.20$ between the two analyses.

Another tool available to assess the fit of the model is the set of standardised residuals (see equation [13]), used to validate the assumption of normality and to assess the variance structure (see expression [8]). For each SNP there are P residuals and these are plotted against the fitted values of π to give the plot in Figure 3-7 (a). Note that the illustrated residuals were calculated using estimates from a single but representative analysis on the full data set. The residuals suggest that both constant variance and zero mean are reasonable assumptions regarding the standardised noise term. The residuals appear to be, at least approximately, normally distributed with zero mean and unit variance, as shown in Figure 3-7 (b). The residual distribution is slightly light-tailed, although not to an extent where normality is implausible.

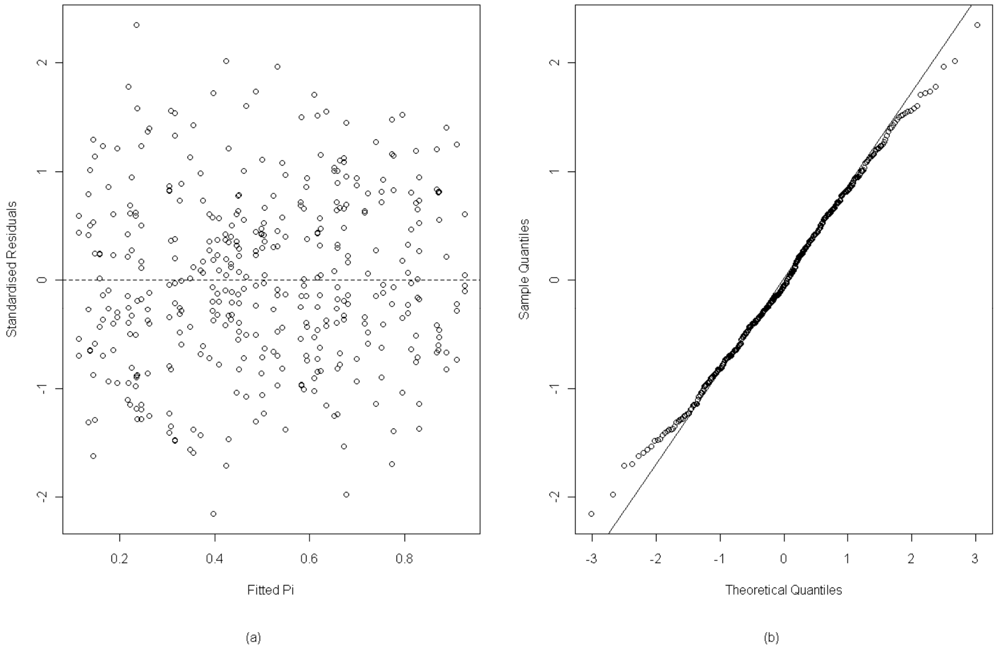


Figure 3-7 (a) Standardised residuals vs fitted π (b) Normal Q-Q plot - ordered standardised residuals vs theoretical quantiles from a standard normal distribution.

In summary, when data are simulated under the ND model, estimates are robust to data removal and the residual analysis plots suggest that the model fits the data well, as ought to be the case when using simulated data.

3.2 European Populations

Here a SNP data set is presented and analysed from the Human Genome Diversity Panel (HGDP-CEPH), sampled from four European populations: a French, an Italian, a Russian and a Scottish population. The French sample was from a Basque population found in the south-west of France, the Italian sample from Sardinia in the Mediterranean Sea, the Russian sample from a location north-east of Moscow (GR 61N, 39-41E) and the Scottish sample from the Orkney Isles. The four populations were assessed at 194 SNP loci under the ND model. SNPs were sampled at widely spaced intervals along an arbitrary chromosome to ensure independence. Samples sizes from the French, Italian, Russian and Scottish populations are 24, 28, 25 and 16 individuals, respectively.

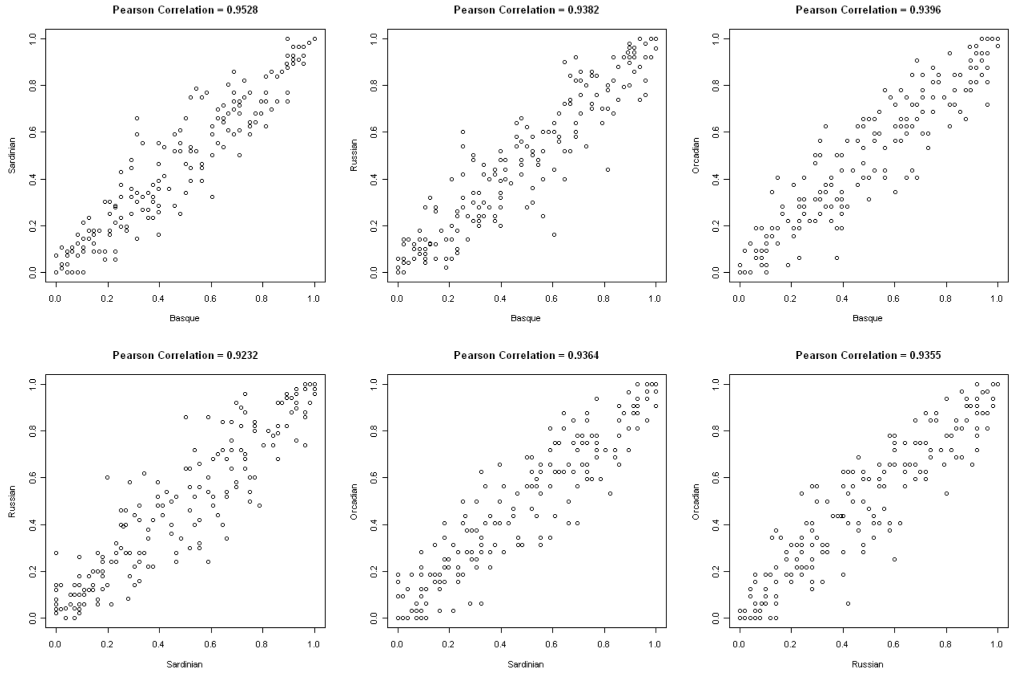


Figure 3-8 Sample allele frequencies at every SNP for all pair-wise population combinations.

Studies of human genetic diversity have found that Europe is the most genetically homogeneous of all the continents (Cavalli-Sforza, 1993), for reasons that are not fully understood, but may be related to continuous gene flow between populations. The plots in Figure 3-8 show highly correlated frequencies for all pairs of populations which suggests little differentiation has occurred between populations, corresponding to small values of *c*. The French-Basque and Sardinian allele frequencies have the highest sample correlation coefficient and the Sardinian and Russian frequencies are the least correlated, as might be expected based on geographical separation.

An MCMC analysis was performed on the European data set with a run length of 10000 iterations. β_{ij} was started from its corresponding x_{ij}/n_{ij}, initial π_j's were drawn from a Beta(2, 2) distribution and the c_i's were started from $F_{ST} = 0.0069$, calculated using all populations (Consortium, 2005). The same prior distributions were used as in section 3.1.

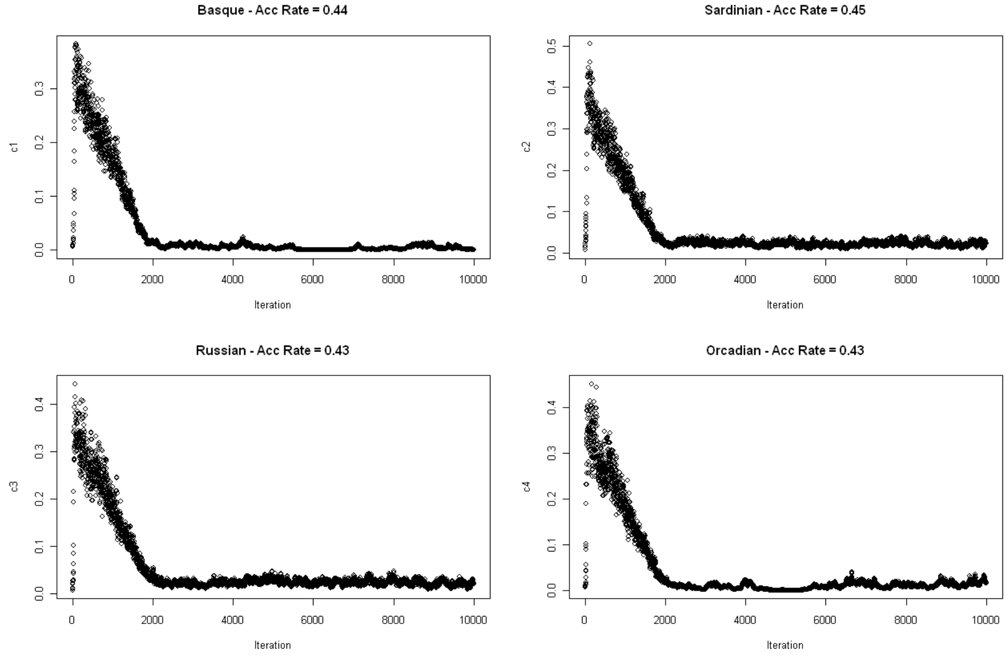


Figure 3-9 Trace plots of *c* parameters from an MCMC run with 10000 iterations for the European data set.

From the trace plots in Figure 3-9 we can see that the chain settles down after around 2000 iterations and so a burn-in period of 2000 was used. The acceptance rates, after adjusting the proposal standard deviation, also seem to suggest that the chain is mixing sufficiently. Since mixing is difficult to see clearly in Figure 3-9, a chain was plotted with the burn-in period removed (Figure 3-10). This confirms that the chain is moving around the parameter space in a satisfactory manner.

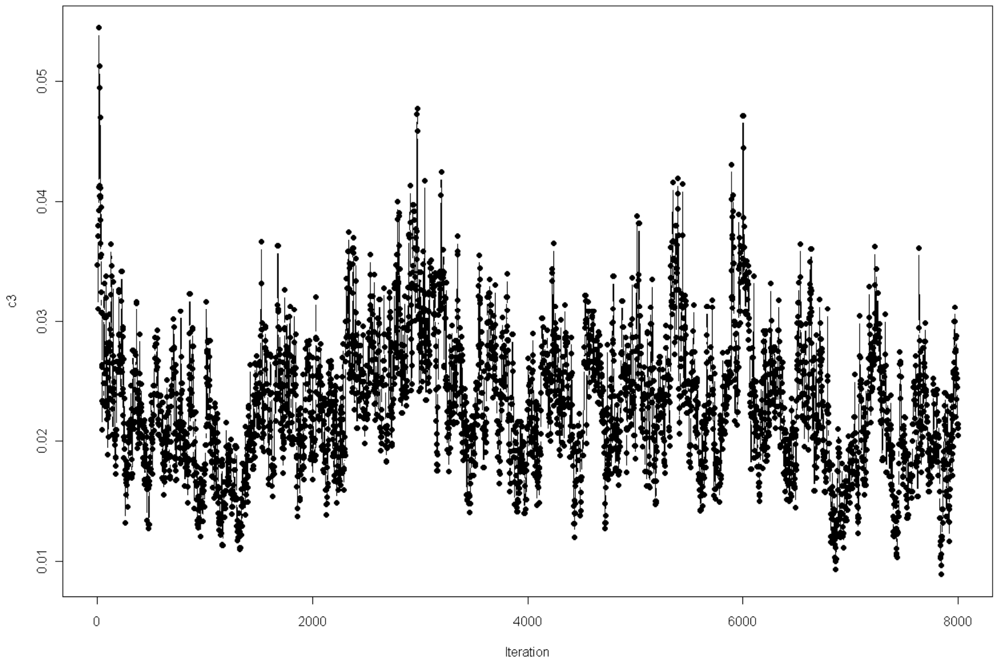


Figure 3-10 Trace plot of c for the Russian population after removing burn-in.

Table 8 Summaries from MCMC Results for European Data

Population	*Parameter*	*Mean*	*Posterior Standard Deviation*	*90% Credible Region*	*Acceptance Rate*
French-Basque	c_1	*0.0025*	*0.0039*	*(8.2×10^{-5}, 0.0097)*	*0.4401*
Sardinian	c_2	*0.0220*	*0.0050*	*(0.0148, 0.0308)*	*0.4355*
Russian	c_3	*0.0249*	*0.0053*	*(0.0169, 0.0342)*	*0.4295*
Orcadian	c_4	*0.0103*	*0.0067*	*(0.0009, 0.0222)*	*0.4340*

Overall the results in Table 8 suggest that very little differentiation has occurred between these populations, the French-Basque population having undergone the least genetic drift by some margin, and the Russian and Sardinian populations showing the most genetic drift, almost equal in fact. That the French-Basque population is the least differentiated may be due to other populations being sampled from either remote parts of Europe (Russian) or islands (Orkney, Sardinia) but this is far from clear. Also, for these data, it appears that F_{ST} gives a fairly good idea of the magnitude of single-population differentiation.

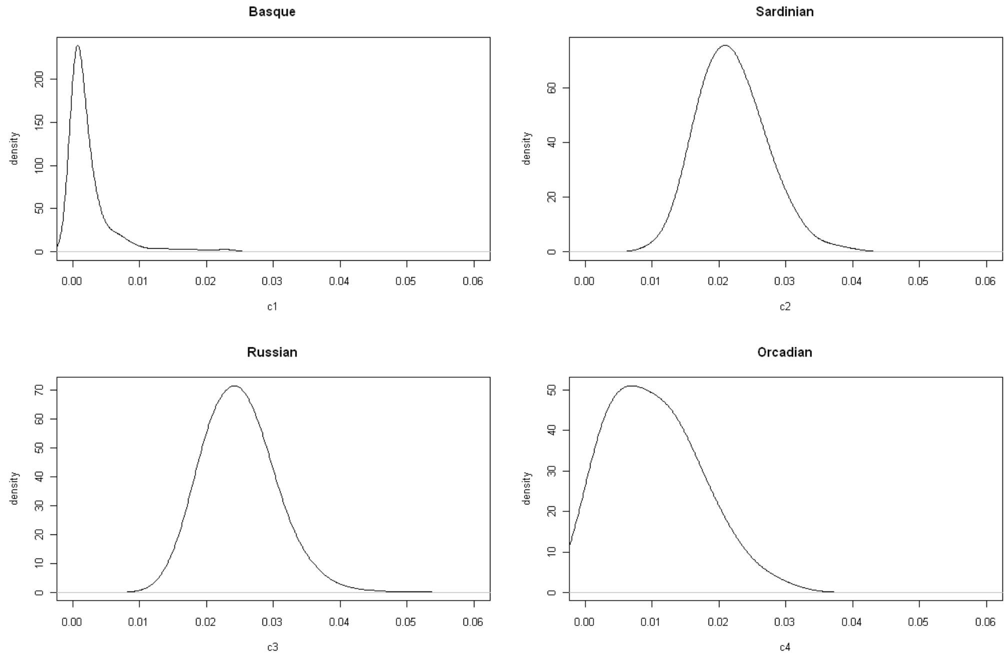


Figure 3-11 Posterior density plots of the *c* parameters for the European data.

The posterior density plots in Figure 3-11 for the French-Basque and Orcadian drift parameters show positive skew in both cases, with the French-Basque being more skewed. The posterior density plots for the Sardinian and Russian populations both resemble the bell-curve of the normal distribution. In all cases the distributions do not have large amounts of variation suggesting that the drift parameters are estimated quite well.

The residual plots in Figure 3-12 do not indicate any problems with the assumptions regarding normality and variance structure of the population allele frequencies; in fact it appears that the ND model fits the data rather well. Table 9 shows credible regions, calculated as before, where every population has been removed and the estimates compared to those from the full data set. Of the 12 intervals, one does not contain zero, when the Sardinian population is removed. Any suggestion regarding the cause of this discrepancy would be speculative, but it at least flags a potential underlying lack of fit.

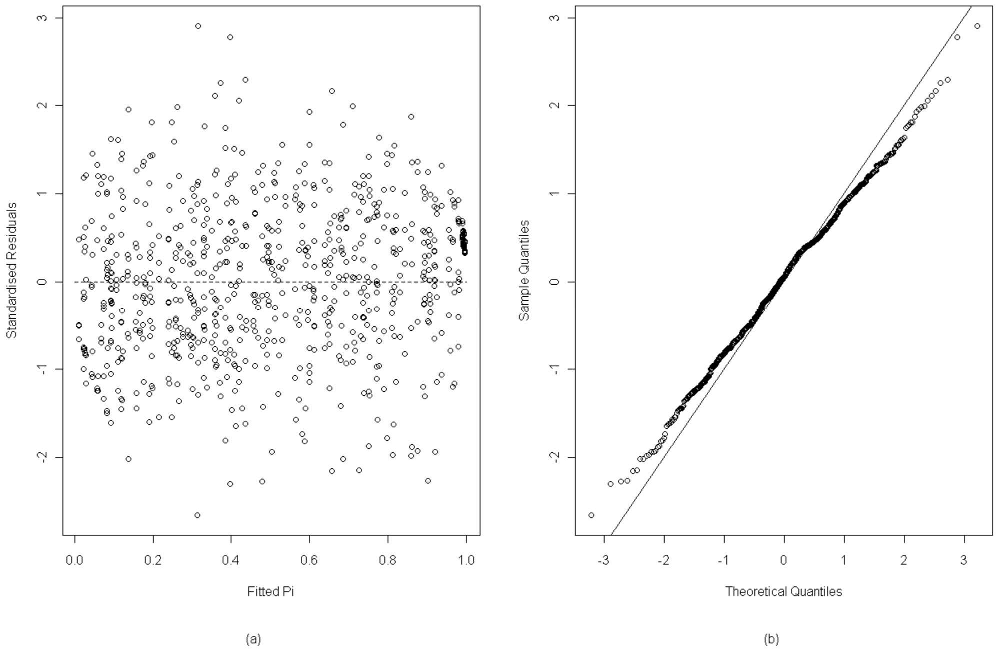


Figure 3-12 (a) Standardised residuals vs fitted π's (b) Normal Q-Q plot - ordered standardised residuals vs theoretical quantiles from a standard normal distribution.

Table 9 90% Credible regions for differences between estimates from full and reduced data sets.

Population	***Parameter***	***Removed Population***			
		French-Basque	***Sardinian***	***Russian***	***Orcadian***
French-Basque	c_1	-	*(-0.0893, -0.0020)*	*(-0.1221, 0.0065)*	*(-0.0949, 0.0070)*
Sardinian	c_2	*(-0.1057, 0.0039)*	-	*(-0.1069, 0.0130)*	*(-0.106, 0.0129)*
Russian	c_3	*(-0.0991, 0.0138)*	*(-0.0857, 0.0156)*	-	*(-0.0930, 0.0034)*
Orcadian	c_4	*(-0.0999, 0.0168)*	*(-0.0893, 0.0129)*	*(-0.1138, 0.0074)*	-

Note: differences calculated after removing burn-in.

The estimates of c for the European populations in this analysis are consistent with the consensus that Europe is the most genetically homogeneous continent, since all estimates suggest very little differentiation. The distributional assumption of normal population allele frequencies appears to hold and the variance structure defined in expression [5] seems to be realistic for these data. It is probably the case that gene flow has occurred between the sampled populations and so the assumption of independent evolution of populations is not likely to hold here. However, in general, regardless of whether the assumptions underlying it are entirely valid, a statistical model that fits some data well remains a useful tool. In our context a model for the joint distribution of allele frequencies across populations can be useful in association studies for common human diseases (Nicholson et al., 2002). Therefore the most notable observation from this analysis is that the ND model appears to fit the data remarkably well.

3.3 Simulation under New Tree Model

In section 3.1 estimates of *c* were shown to be stable when using simulated data under the ND model and the residual diagnostics reflected that the model fitted the data well, as would be expected. It is of interest to see whether the same diagnostics highlight lack of fit and instability when data resulting from more complex patterns of ancestry are analysed under the ND model. One would expect to see lack of fit manifest in the residuals when using the incorrect model to analyse the data. However the property of instability also offers insight since Nicholson et al. (2002) reported extremely unstable estimates of *c* when highly correlated populations were included in the sample.

To answer these questions 100 independent data sets were simulated using the new model for a given ancestral configuration and analysed using the ND model. In assessing stability the same approach was taken as before where an arbitrary population was removed and estimates compared by computing credible regions of differences at every step of the chain after removing burn-in. The process was again repeated twice, removing a different population each time.

3.3.1 Analysis

First we simulated 100 independent SNP data sets under the new model using the ancestor vector *a* to define a common evolutionary history, with each data set containing P=4 populations, typed at L=100 SNPs, with sample sizes n_{ij}=100. Adhering to the simulation procedure for the new model outlined in section 2.2, each of the 100 data sets were simulated as follows:

1. Draw β_{ij} from Beta(2, 2) where $i = 2P–1$ and $j = 1, . . . , L$.
2. Draw $\beta_{ij} \big| \beta_{a(i),j}, c_i$ from Normal$\big(t(\beta_{a(i),j}), c_i t(\beta_{a(i),j})(1 - t(\beta_{a(i),j}))\big)$ where $i = 1, . . . , 2P–2$, c = (0.40, 0.32, 0.02, 0.02, 0.18, 0.20), a = (7, 6, 5, 5, 6, 7, 0).
3. Draw $x_{ij} \big| \beta_{ij}, n_{ij}$ from Binomial$\big(n_{ij}, t(\beta_{ij})\big)$ for $i = 1, \ldots, P,\ j = 1, \ldots, L$.

The c's were configured in this way in an attempt to represent a real data set including Europeans and some other populations continentally separated from Europe.

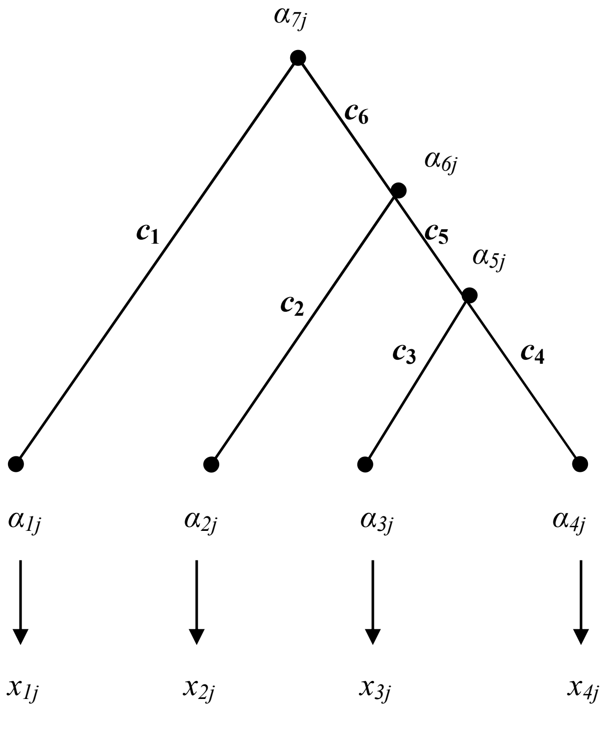


Figure 3-13 A diagrammatic representation of the model used to simulate the data for a single SNP j.

For all MCMC analyses discussed in this section the prior configurations were exactly the same as those used in section 3.1.1. The 100 simulated data sets were analysed under the ND model for four populations for the full data sets and three populations for the reduced data sets and the estimates compared using the method previously described. In the first analysis population 1 was removed, the data re-analysed and corresponding c estimates were compared. Referring to Figure 3-13 it is difficult to make any prior judgement as to the results of this analysis other than it seems likely that estimates will be unstable given that the data are analysed under the incorrect model.

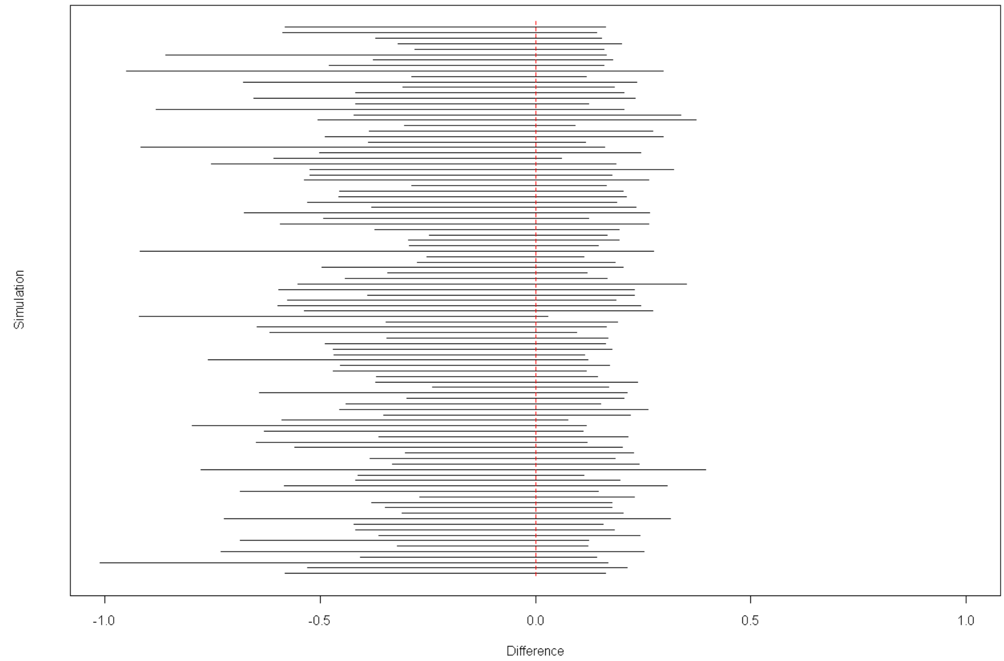


Figure 3-14 100 90% credible regions for the difference between draws of two separate MCMC analyses under the ND model; one with the full data set, one with population 1 removed from the data set. Comparison 1 refers to the difference between the estimate of c_2 for the full data and c_1 for the reduced data set.

All the intervals in Figure 3-14 (where population 1 is removed) contain zero, although most are not centred on zero but on negative values, suggesting that it is more likely that the estimate is larger when using the reduced data set compared to the full data set, given the sign of the difference. But since all of the intervals contain zero it must be concluded that the estimates appear to be stable in this case. Looking at Figure 3-15, many more of the intervals do not contain zero suggesting that these estimates are unstable. In fact, of the 100 intervals, 12 do not contain zero. Another observation is that many of the intervals only just contain zero and all are centred on positive values. Therefore estimates tend to be larger when the full data set is considered (significantly so for 12% of the data sets). Figure 3-16 is much the same as Figure 3-15, only out of the 100 intervals, 19 do not contain zero this time.

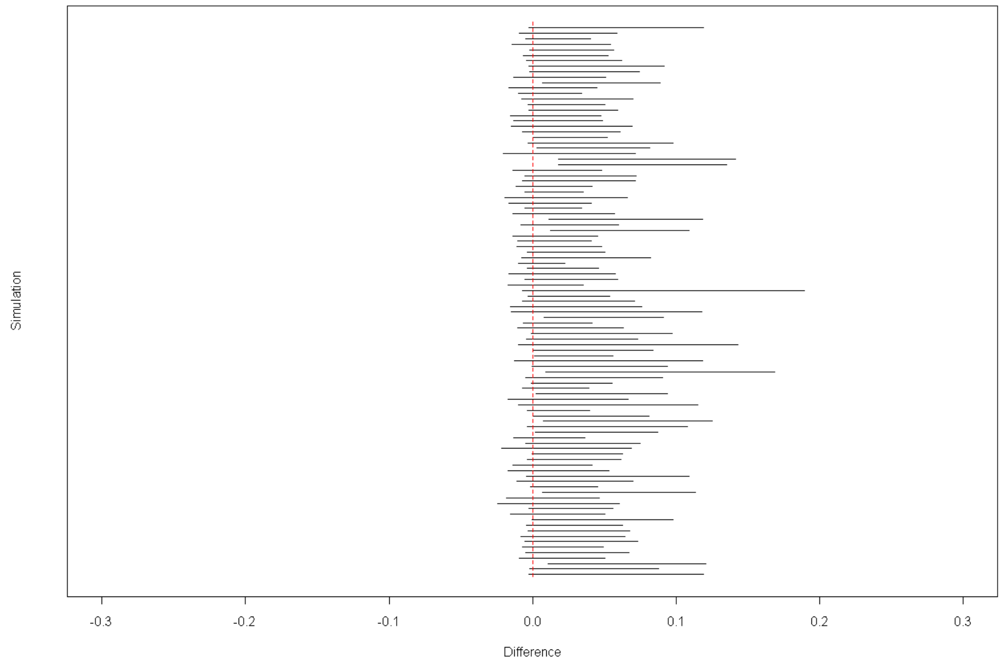


Figure 3-15 100 90% credible regions for the difference between draws of two separate MCMC analyses under the ND model; one with full data set, one with population 1 removed from the data set. Comparison 2 refers to the difference between the estimate of c_3 for the full data and c_2 for the reduced data set.

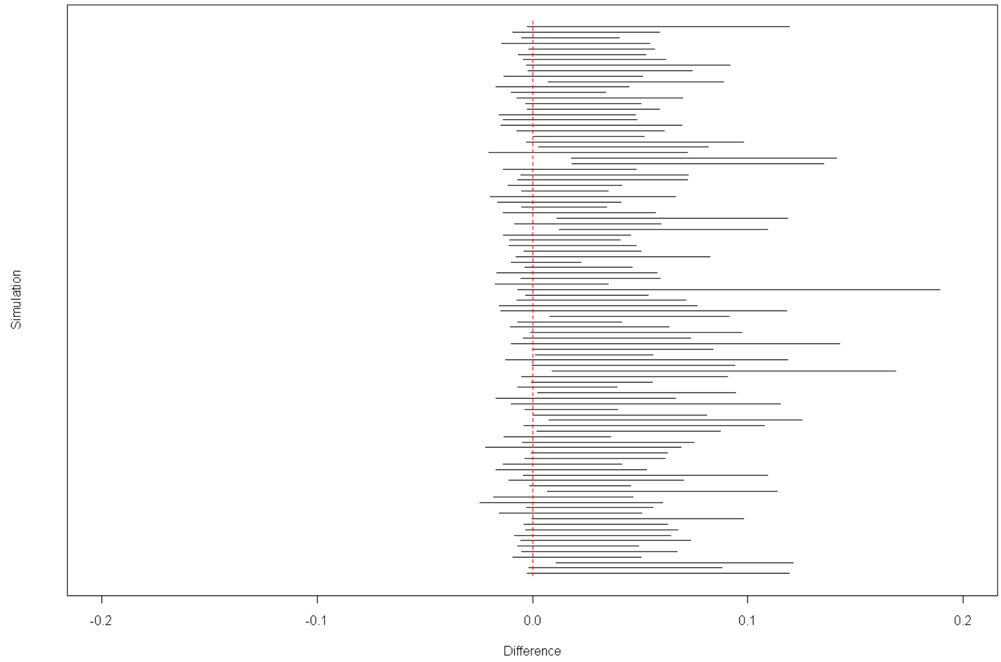


Figure 3-16 100 90% credible regions for the difference between draws of two separate MCMC analyses under the ND model; one with full data set, one with population 1 removed from the data set. Comparison 3 refers to the difference between the estimate of c_4 for the full data and c_3 for the reduced data set.

From these analyses it is observed that estimates of c are not robust to population removal (in this case the population connected to the MRCAP) when using data simulated under a given bifurcating tree topology and then subsequently analysed under the ND model.

In the second analysis population 4 was removed from the data set, the data re-analysed and corresponding c estimates were compared. Once again, one would expect to see instability in the estimates since the incorrect model is used to analyse these data.

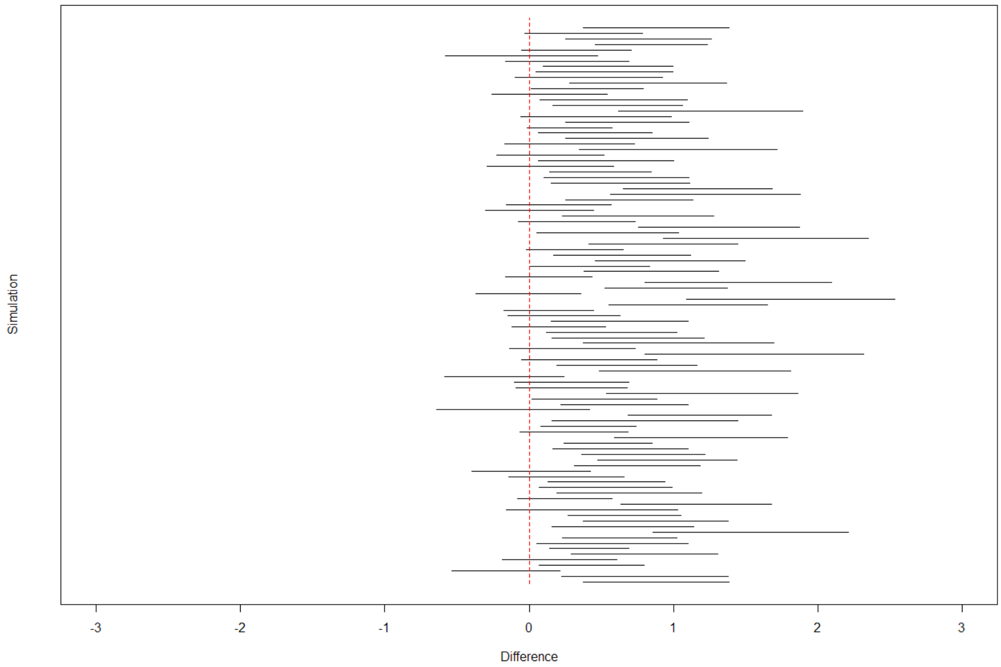


Figure 3-17 100 90% credible regions for the difference between draws of two separate MCMC analyses under the ND model; one with full data set, one with population one removed from the data set. Comparison 1 refers to the difference between the estimate of c_1 for the full data and c_1 for the reduced data set.

When population 4 is removed from the data set c_1 is extremely unstable, as illustrated in Figure 3-17. Of the 100 intervals, 67 did not contain zero. The intervals tend to be wholly positive, revealing the tendency of the estimate of c_1 from the full data set to be larger than c_1 from the reduced data set. Looking at Figure 3-18, the estimates of c_2 are again unstable although less so than for c_1. Of the 100 intervals, 35 do not contain zero. Figure 3-19 exhibits the most intervals that do not contain zero, at 95 out of 100. These simulations highlight the definite instability in the estimates of c when populations are removed.

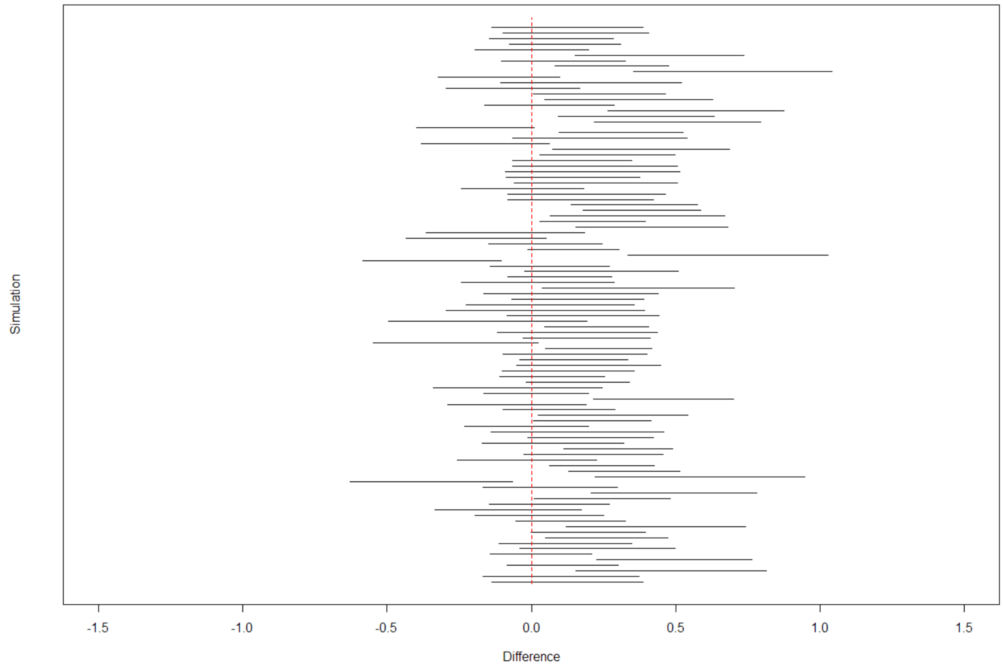


Figure 3-18 100 90% credible regions for the difference between draws of two separate MCMC analyses under the ND model; one with full data set, one with population one removed from the data set. Comparison 2 refers to the difference between the estimate of c_2 for the full data and c_2 for the reduced data set.

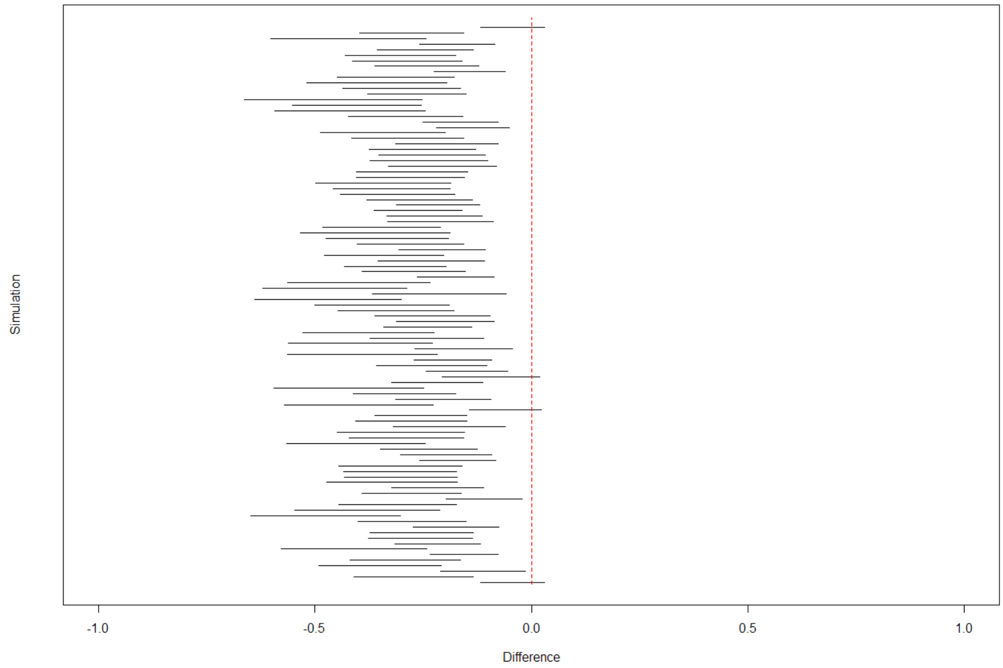


Figure 3-19 100 90% credible regions for the difference between draws of two separate MCMC analyses under the ND model; one with full data set, one with population one removed from the data set. Comparison 3 refers to the difference between the estimate of c_3 for the full data and c_3 for the reduced data set.

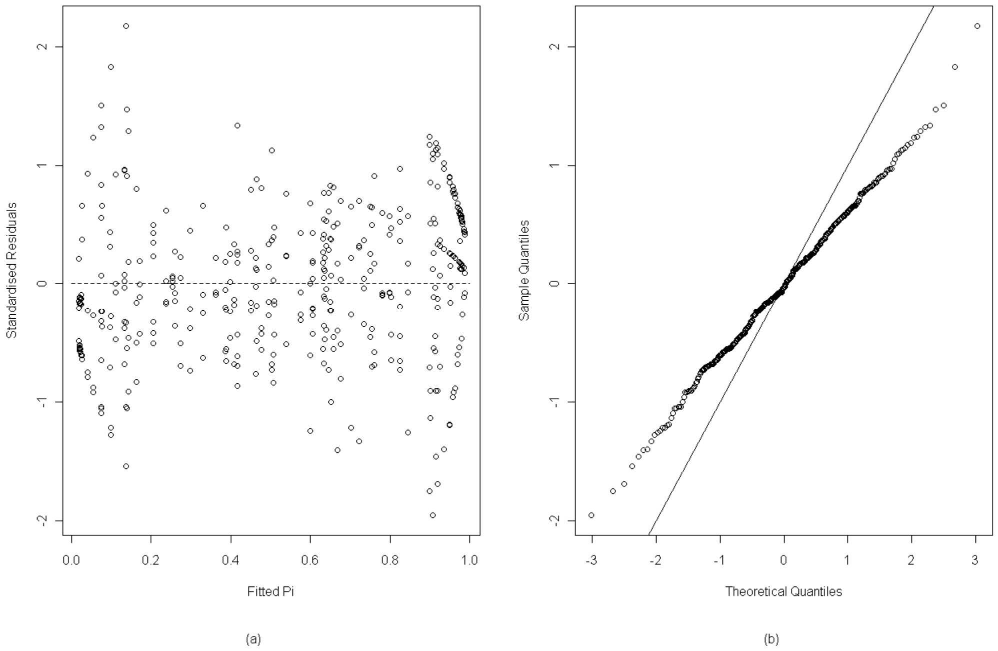


Figure 3-20 (a) Standardised residuals vs fitted π (b) Normal Q-Q Plot - ordered standardised residuals vs theoretical quantiles from a standard normal distribution. Note that the residuals were calculated using estimates from a single but representative analysis on the full data set under the ND model using data simulated under the new model.

Looking at the residual diagnostics in Figure 3-20, there is evidence that the model does not fit the data well. The assumption of constant variance appears to be violated as the variation in the residuals tends to be greater for values of π near the extremes. The assumption of normality also appears to be strongly violated for these data.

This example demonstrates that if a bifurcating tree topology is the correct representation of the evolutionary history of a set of populations, then the inadequacy of the ND model can be highlighted using the residual and population-removal diagnostics. This result is encouraging since the extension to the ND model was used to simulate these data, while Nicholson et al. (2002) found similar results when using real data. In the next section the fit of the ND model will be explored for data whose topology likely deviates from the ND model.

3.4 Global Populations 1

Here a SNP data set is presented and analysed from the HGDP-CEPH panel for four populations sampled from Africa, Cambodia and Mexico. The data set includes a North African Mozabite population, a Biaka Pygmy population from sub-Saharan Africa, a Cambodian population and a Native American Pima population from Mexico. The four populations were assessed at the same 194 SNP loci as in the previous example (section 3.2). Samples sizes of the Biaka, Mozabite, Cambodian and Pima populations are 32, 30, 11 and 25 individuals, respectively.

For these data, the assumption of independent evolution is more plausible than for the European data set due to the geographical distances between populations; however the simultaneous divergence of all sampled populations assumed under the ND model is unlikely to hold here. It was shown in section 3.3 that both the residual and population removal diagnostics can be used to highlight departures from the simple topology under the ND model. If it is the case that the populations under examination do in fact have a more complex evolutionary history, as seems likely, then one would expect lack of fit when the data are analysed under the ND model.

Of the relationships illustrated in Figure 3-21, that the Mozabite and Biaka populations are fairly strongly correlated (Pearson correlation coefficient, P.C.C = 0.7469) is the least surprising since both are African populations, albeit rather geographically separated. The highest sample correlation is between the Pima and Cambodian populations (P.C.C = 0.7688, Figure 3-21). Given the geographic distance between these two populations a strong relationship does not seem to make intuitive sense. However it is generally accepted that the Americas were populated by East Asians during the last ice age (Atkinson, Gray and Drummond, 2008), when it was possible to travel from Siberia to Alaska on foot, due to the ice coverage (Olson, 2002). With this in mind, the association between the Pima and Cambodian populations is much less puzzling. There is also a fairly strong relationship between the Mozabite and Cambodian populations (P.C.C = 0.7463, Figure 3-21) although no obvious interpretation presents itself.

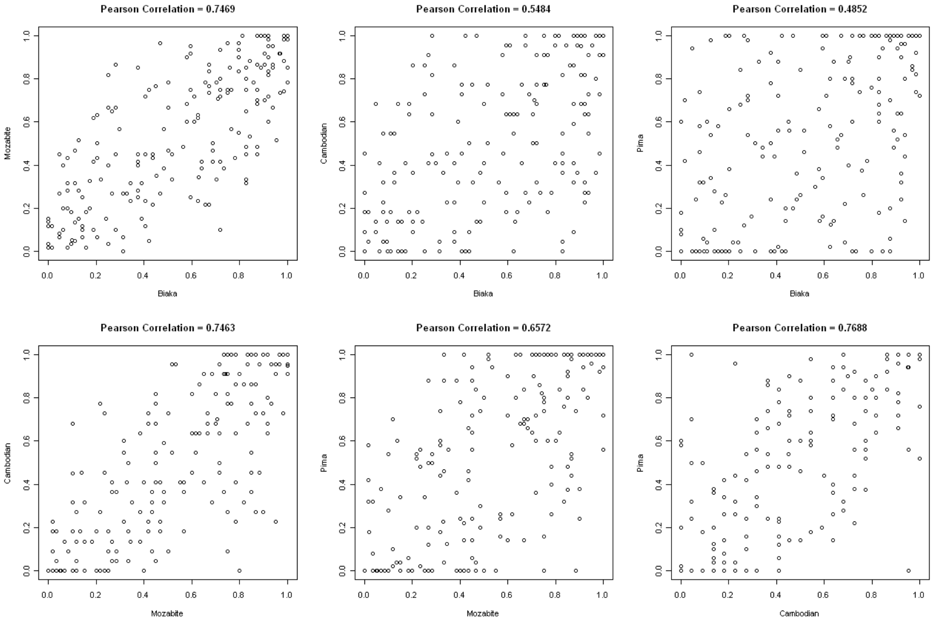


Figure 3-21 Sample allele frequencies at every SNP for all pair-wise population combinations with corresponding sample Pearson correlation coefficients.

An MCMC analysis was performed with a run length of 10000 iterations. β_{ij} was started from its corresponding x_{ij}/n_{ij}, initial π_j's were randomly drawn from a Beta(2, 2) distribution and the c_i's were started from $F_{ST} = 0.0594$, calculated using all populations. The same prior distributions were used as in previous sections.

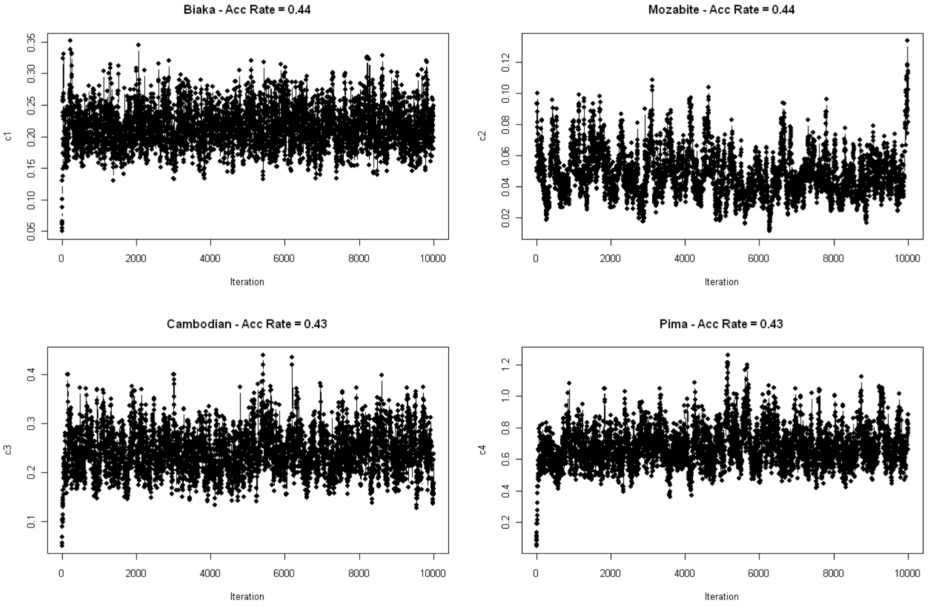


Figure 3-22 Trace plots of *c* parameters from an MCMC run with 10000 iterations for the World 1 data set.

Looking at Figure 3-22, the chains appear to settle to the target distribution after around 1000 iterations, possibly quicker than this. The proceeding results are presented after removing a burn-in of 1000. The chains also show satisfactory mixing resulting in acceptance rates of approximately 40%.

Table 10 Summaries of MCMC Results for World 1 Data Set

Population	*Parameter*	*Mean*	*Posterior Standard Deviation*	*90% Credible Region*	*Acceptance Rate*
Biaka	c_1	*0.2151*	*0.0310*	*(0.1662, 0.2690)*	*0.4367*
Mozabite	c_2	*0.0479*	*0.0149*	*(0.0271, 0.0747)*	*0.4373*
Cambodian	c_3	*0.2420*	*0.0433*	*(0.1762, 0.3164)*	*0.4347*
Pima	c_4	*0.6792*	*0.1181*	*(0.5109, 0.8886)*	*0.4342*

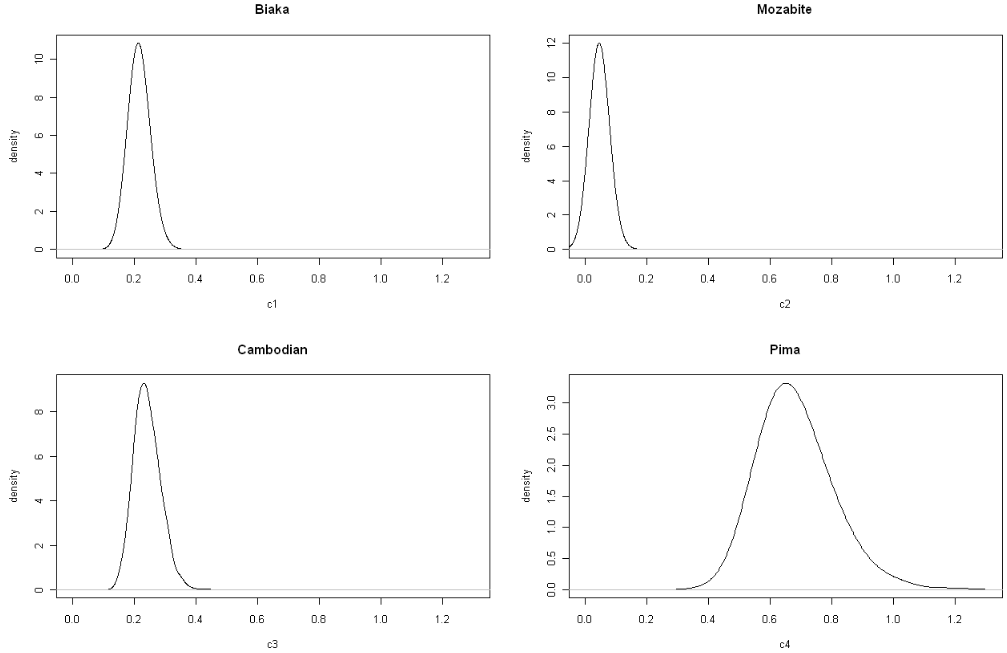


Figure 3-23 Posterior density plots of the *c* parameters for Global data set 1.

The population with the largest value of *c* is the Pima population ($\hat{c}_4$ = 0.6792, Table 10). This is possibly due to the small numbers of immigrants thought to have populated the Americas from East Asia (Atkinson et al., 2008); recall that *c* is inversely proportional to population size. An alternative explanation is that the large *c* represents an old population, since genetic drift is proportional to time. However, Native Americans are not thought to be particularly old populations, and so the population size interpretation seems more feasible. The Mozabite population has the smallest value of c ($\hat{c}_2$ = 0.0479, Table 10), a likely reflection of common origin with Europeans. In fact, if one considers Europeans and North Africans they tend to resemble one another in many phenotypes. The estimates of c for the remaining two populations, the Biaka pygmies and the Cambodians, are more challenging to interpret. The Biaka pygmies are thought to be a very old population which would suggest a higher value of c, although population size may have contributed to its relatively moderate value ($\hat{c}_1$ = 0.2151, Table 10). The value of c for the Cambodians ($\hat{c}_3$ = 0.2420, Table 10) is higher than one might expect for an East Asian population (Nicholson et al., 2002). It is notable that this sample is the smallest studied and may not be representative. It is also worth highlighting the discrepancy between F_{ST} and the estimates of c for these data. F_{ST} = 0.0594 suggests that approximately 6% of the overall variation in allele frequencies is between-population variation, whereas the estimates of c for single populations suggest relatively large differentiation for all populations.

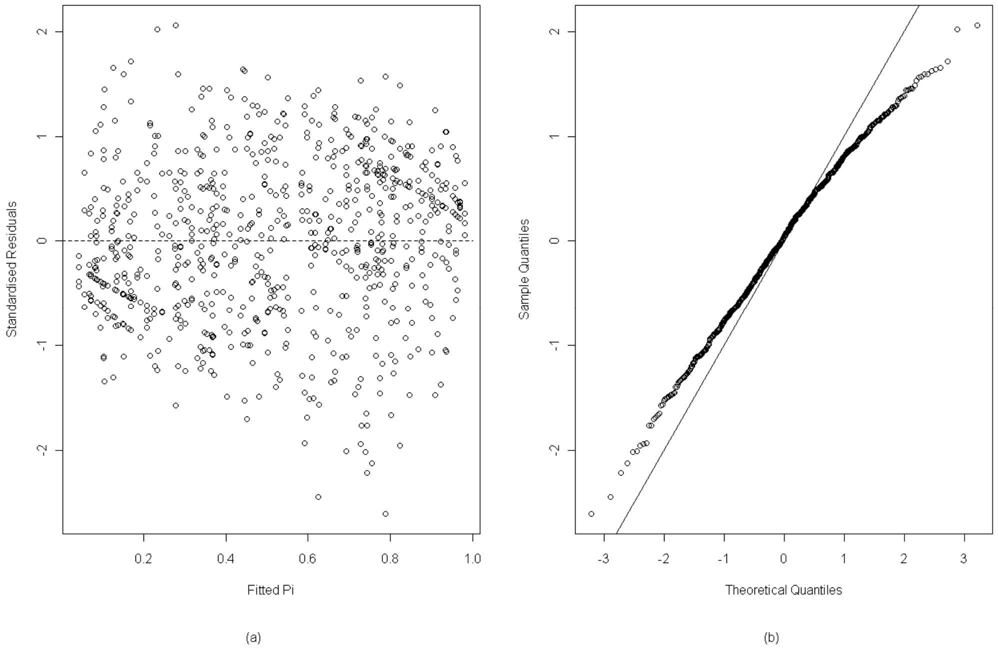


Figure 3-24 (a) Standardised residuals vs fitted π's (b) Normal Q-Q plot - ordered standardised residuals vs theoretical quantiles from a standard normal distribution.

The residual plots in Figure 3-24 highlight that the ND model does not fit these data at all well. The noise does not appear to have constant variance since the range of the residuals is not constant across all the fitted values of π. Also the distribution of the residuals does not appear to resemble a standard normal since there is skewness suggested in Figure 3-24 (b).

Table 11 90% Credible Regions for Differences between Estimates from Full and Reduced Data Sets.

Population	*Parameter*	*Removed Population*			
		Biaka	*Mozabite*	*Cambodian*	*Pima*
Biaka	c_1	-	*(-0.1797, 0.0452)*	*(-0.0360, 0.1202)*	*(-0.0712, 0.0896)*
Mozabite	c_2	*(-0.1011, 0.0059)*	-	*(-0.0452, 0.0313)*	*(-0.0226, 0.0387)*
Cambodian	c_3	*(-0.0011, 0.1871)*	*(-0.0468, 0.1490)*	-	*(-0.1385, 0.0674)*
Pima	c_4	*(-0.0712, 0.3960)*	*(-0.0925, 0.4033)*	*(-0.4672, 0.1243)*	-

Note: differences calculated after removing burn-in.

The intervals in Table 11 are somewhat surprising since they all contain zero, suggesting that the estimates are robust to population removal. The estimates of the amount of genetic drift for the Pima population appear to change the most when a population is removed since the intervals are centred on values quite distant from zero, but again stability must be concluded. It may be the case that increasing the number of SNPs yields significant differences, but this avenue has not been pursued here. This example highlights that the leave-one-out diagnostic is not infallible, since it does not highlight any discrepancies for these data when it is likely that the ND model is not an accurate representation. Nevertheless, there is evidence from the residual analysis that the ND model does not represent these data adequately, and the tree topology under the ND model may be the source of the disagreement.

3.5 Global Populations 2

Here another data set taken from the HGDP-CEPH panel is presented and analysed. These data include two populations from sub-Saharan Africa: Mbuti pygmies and Mandenka; and two from Europe: a French and a Tuscan population. The four populations were assessed at 194 SNP loci under the ND model. Samples sizes of the Mandenka, Mbuti, French and Tuscan populations are 24, 15, 29 and 8 individuals, respectively.

As with the populations analysed in section 3.4, the populations comprising this data set are not likely to be represented well by the ND model, as regards their evolutionary past, and therefore it should be the case that a lack of fit be manifest in the diagnostics. The analysis in section 3.3.1 showed that the diagnostics are able to detect departures from the ND model given that the data reflect an alternative model, and so similar results in the proceeding analysis would provide evidence that these populations are described by the model in section 3.3.1, or a model of similar structure.

As would be expected, the European populations are highly correlated (P.C.C = 0.9305, Figure 3-25), which reflects the genetic homogeneity found in European populations. The African populations are also exhibit a strong correlation (P.C.C = 0.8381, Figure 3-25) probably due to being in fairly close geographic proximity. The Mandenka appear to be more closely related to the two European populations than the Mbuties are to the Europeans, although there is not an obvious interpretation for this relationship.

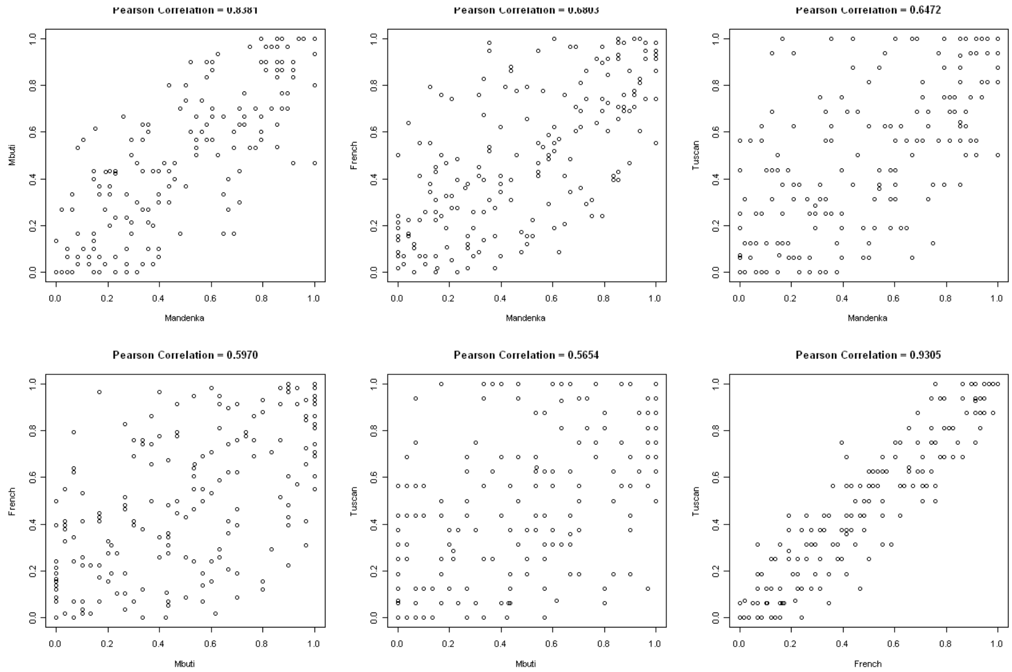


Figure 3-25 Sample allele frequencies at every SNP for all pair-wise population combinations with corresponding sample Pearson correlation coefficients.

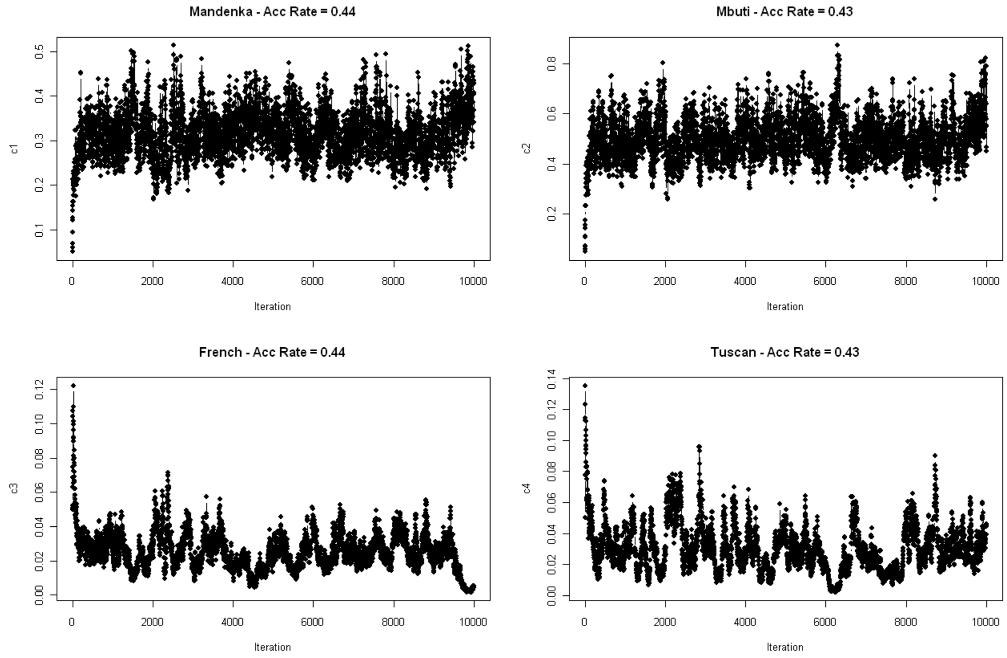


Figure 3-26 Trace plots of *c* parameters from an MCMC run with 10000 iterations for the World 2 data set.

An MCMC analysis was performed on these data, using identical initial configurations as were used in section 3.2 and 3.4, with $F_{ST} = 0.0425$. The chains from this analysis, presented in Figure 3-26, do not highlight any problems with mixing and the acceptance rates are all approximately 40%. The chains for the European populations appear to move around the parameter space in smaller steps, which appears to be a feature of estimation procedure when the *c*'s are small, but this does not present any immediate problems.

Both of the African populations have large estimates of *c* (Mandenka, $\hat{c}_1 = 0.3160$; Mbuti, $\hat{c}_2 = 0.4962$; Table 12) which may be reflecting the age of these populations relative to the Europeans. The Mbuti having a considerably larger value may be a consequence of a smaller population size or that they are in fact older. The Europeans again have small values of *c* (French, $\hat{c}_3 = 0.0250$; Tuscan, $\hat{c}_4 = 0.0300$; Table 12) reflecting a relatively small amount of genetic drift. Again, when compared to the estimates of *c*, F_{ST} appears to under-estimate the proportion of between-population variation relative to the total variation, since F_{ST} is essentially the mean of the *c*'s.

Table 12 Summaries of MCMC results for global data set 2.

Population	*Parameter*	*Mean*	*Posterior Standard Deviation*	*90% Credible Region*	*Acceptance Rate*
Mandenka	c_1	*0.3160*	*0.0531*	*(0.2338, 0.4071)*	*0.4408*
Mbuti	c_2	*0.4962*	*0.0833*	*(0.3764, 0.6434)*	*0.4345*
French	c_3	*0.0250*	*0.0108*	*(0.0092, 0.0427)*	*0.4401*
Tuscan	c_4	*0.0300*	*0.0150*	*(0.0101, 0.0580)*	*0.4340*

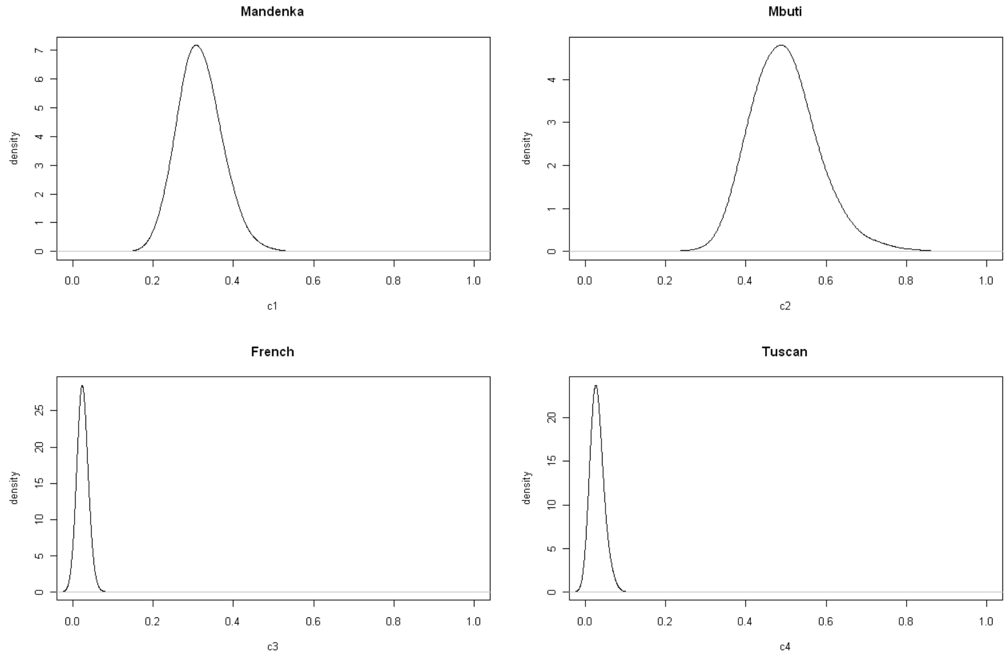


Figure 3-27 Posterior density plots of the *c* parameters for the Global data set 2.

The plot in Figure 3-28 (a) suggests that the noise does not have constant variance, since the range of the residuals is not constant across the fitted values of π. The assumption of normality also appears to be violated in this case. The leave-one-out diagnostic also suggests a lack of fit since half of the intervals in Table 13 do not contain zero. These instances occur when either of the European populations is removed from the data set. This is particularly interesting since instability was found in section 3.3.1 when populations with small values of *c*, located below the most recent population split, were removed. As previously discussed, either of the topologies in Figure 2- is likely to be fairly accurate for these data. The fact that both diagnostics suggest that the ND model does not fit the data well in this instance is evidence that the source of the discrepancy is the incorrect topology of the ND model.

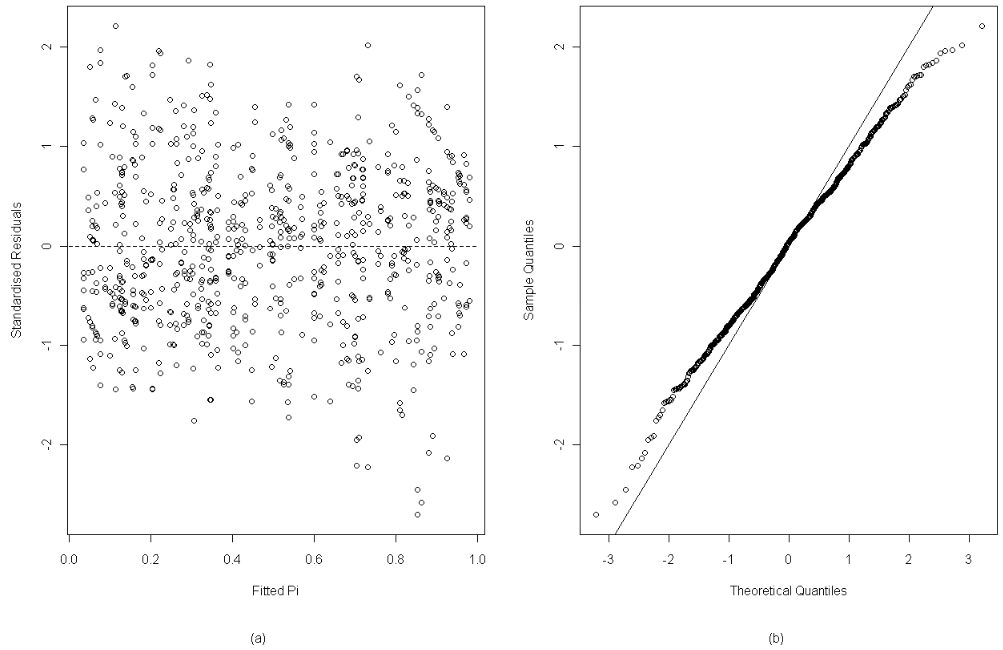


Figure 3-28 (a) Standardised residuals vs fitted π's (b) Normal Q-Q plot - ordered standardised residuals vs theoretical quantiles from a standard normal distribution.

Table 13 90 % Credible regions for differences between estimates from full and reduced data sets.

Population	***Parameter***	***Removed Population***			
		Mandenka	***Mbuti***	***French***	***Tuscan***
Mandenka	c_1	-	*(-0.2252, 0.0301)*	*(0.1772, 0.3505)*	*(0.1746, 0.3448)*
Mbuti	c_2	*(-0.3527, 0.0469)*	-	*(0.2305, 0.5043)*	*(0.2194, 0.4970)*
French	c_3	*(-0.0024, 0.0352)*	*(-0.0021, 0.0335)*	-	*(-0.3101, -0.1701)*
Tuscan	c_4	*(-0.0010, 0.0469)*	*(-0.0137, 0.0393)*	*(-0.4625, -0.2458)*	-

Note: differences calculated after removing burn-in.

3.6 Identifiability

When fitting the new model there arises an issue with identifiability, particularly with the *c*'s, meaning that there is insufficient information in the data to estimate parameters independently. To illustrate the problem, some data were simulated under the new model (see section 2.2 for details) for a labelled history defined by $a = (7, 6, 5, 5, 6, 7, 0)$ and $c = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$. An MCMC analysis was performed under the new model, specifying the correct labelled history, with a run length of 5000 iterations and prior distributions identical to those used in previous analyses.

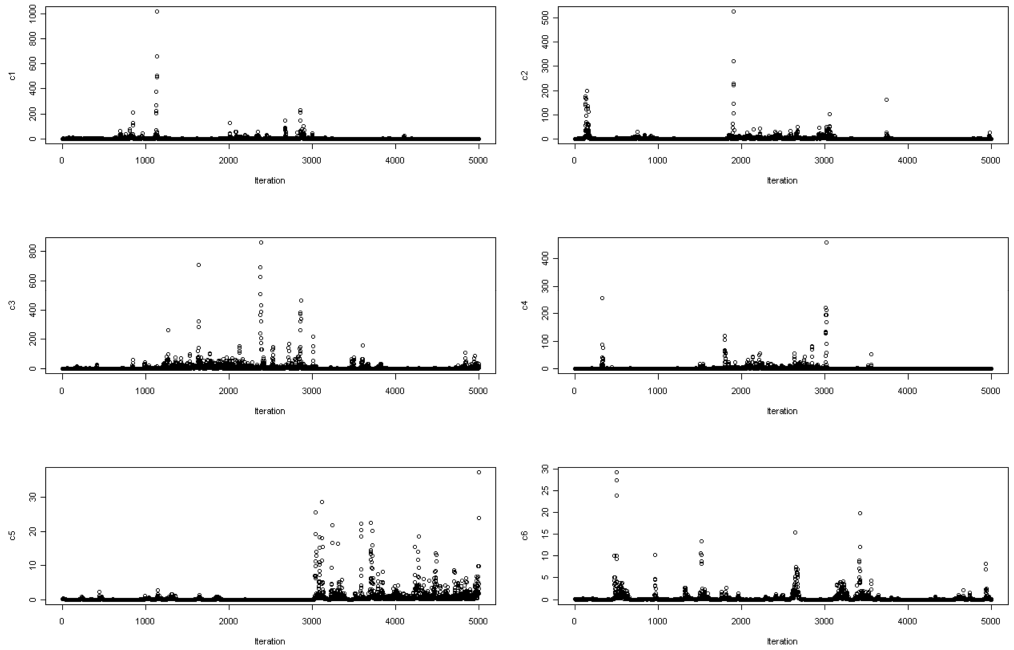


Figure 3-29 Trace plots of an MCMC run of 5000 iterations without removing burn-in, P=4, L=100, n=100, c = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1), a = (7, 6, 5, 5, 6, 7, 0).

The chains for the *c*'s in Figure 3-29 are clearly unsatisfactory. Previous analyses have yielded values no greater than 0.7 for a highly differentiated population. That the chains reach values in the order 10^2, and in one case 10^3, suggests that there might be an issue with identifiability. Note also that many simulated data sets were considered under various topologies and using different configurations of *c*, with similar results.

To investigate the source of the problem it is helpful to consider a simple example. Since the new model is, under any given topology, a series of bifurcating branches, let us consider the simplest case of two populations. Note that for two populations, the new model is exactly the ND model with $P = 2$. To motivate the concept of identifiability in a similar, but not identical, context to the two population model, consider two random walks, X_n and Y_n, with variance σ^2 and τ^2 respectively, and with identical initial values such that $X_o = Y_o$. Then further suppose that both processes follow Brownian motion. In relation to our model, the initial value corresponds to the ancestral frequency π at a single SNP locus; the random walk reflects the Markov chain used to derive the probabilistic properties of ND model under the Wright-Fisher model; Brownian motion reflects the Normal distributions used to characterise allele frequencies and the variance terms are simplified such that they do not depend on the mean. The theory of Brownian motion then states that after time t, $X_n \sim \text{Normal}(X_{o,}\ n\sigma^2)$ and $Y_n \sim \text{Normal}(Y_{o,}\ n\tau^2)$. To then make an inference about the individual variance terms, a first step might be to compute the difference between X_n and Y_n and one could then proceed with a likelihood-based argument, using the result that $X_n - Y_n \sim \text{Normal}(0,\ n(\sigma^2 + \tau^2))$. Without following through the mathematics of a likelihood argument, it is still clear that independent estimates of σ^2 and τ^2 cannot be found in this case; only their sum is identifiable. This would still be the case if any number of independent pairs of random walks were considered, allowing the initial values of each pair to vary, which corresponds to sampling at numerous SNP loci. However the situation is slightly different when using the two population model, since differences between allele frequencies are not directly calculated. There is some information about c_i in the variation of the distribution of allele frequencies in population i across all SNP loci.

Figure 3-30 shows the chains from an analysis on two populations where the c's are identical and relatively small at 0.01. In this case there is insufficient information in the distribution of allele frequencies and the chains for the individuals c's do not behave at all well, even though point estimates are fairly accurate. Notice the negative correlation between the chains for c_1 and c_2, reflected in the individual chains and also in the plot of c_2 against c_1. The most interesting observation is that the behaviour of the chains mirror the properties suggested in the identifiability example; namely that the chain for the sum of the c's is well determined but not individually and that the relationship between c_1 and c_2 is described well by the line $c_1 + c_2 = k$, where k is a constant. This suggests that differences between allele frequencies are implicitly considered during the MCMC estimation procedure under the ND and the new model. However, the identifiability issue does not arise in all cases, particularly when the c's are large. For example in Figure 3-31 the chains mix well in both cases for c's of 0.1 and 0.5.

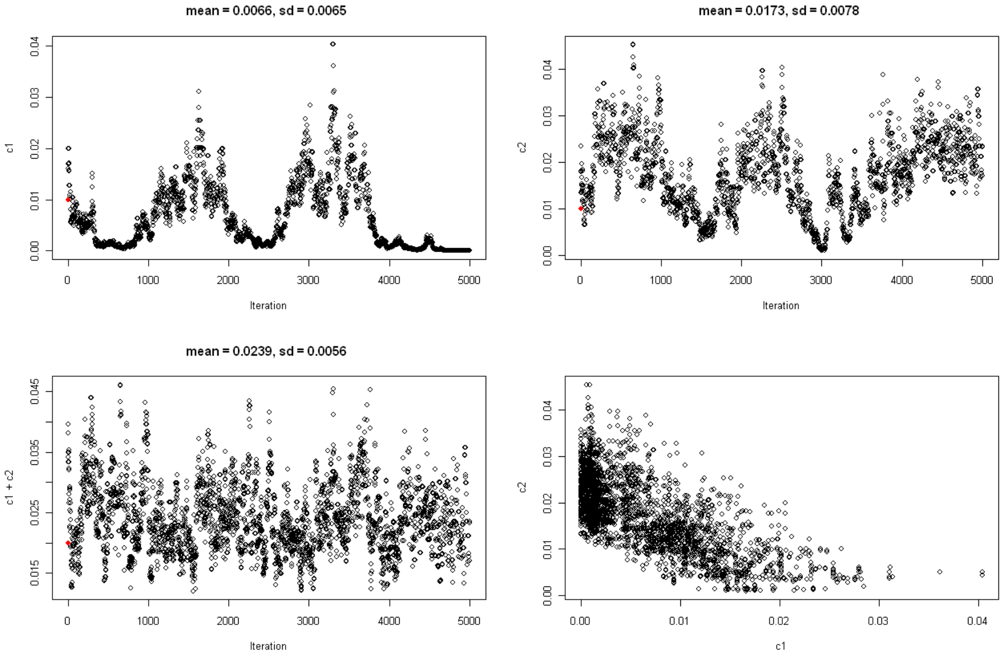


Figure 3-30 Trace plots from an MCMC analysis for two populations, with a run length of 5000 iterations. Red dots indicate true values ($c_1 = c_2 = 0.01$).

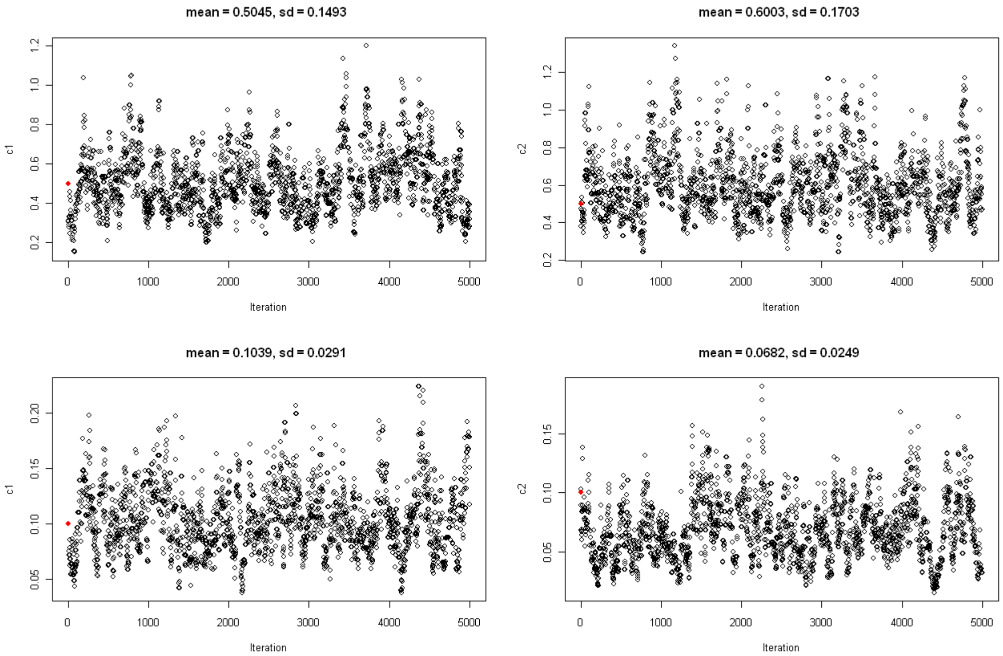


Figure 3-31 Trace plots from two MCMC analyses for two populations, with run lengths of 5000 iterations. Red dots indicate true values ($c_1 = c_2 = 0.5$ and $c_1 = c_2 = 0.1$)

Returning to the new model and the cause of the identifiability, it is clear that the example highlights a contributing factor, since the new model is a series of bifurcating branches. However the model for two populations is formally identifiable, only there is very little information in the data to estimate the c's, particularly when there is low variation across SNPs. This means that there must be further causes of non-identifiability in the new model.

The novelty of the new model is its use of ancestral populations other than MRCAP, called internal nodes, whose allele frequency can be zero or one. Using such theoretical populations potentially offers a more realistic model in a historical sense than the ND model, but this added complexity appears to be more of a burden, since the difficulties that arise and the methods used to facilitate them seem to render the model non-identifiable. The most likely reason for the difficulties one encounters when fitting the new model is that the variance of β_{ij} is dependent on the mean, $\beta_{a(i),\ j}$. This is only a problem because the internal node frequencies are allowed to vary on the real line. The factor of $\beta_{a(i),j}\ (1-\beta_{a(i),j})$ in the variance means that $\beta_{a(i),\ j}$ must be in the range (0, 1) to avoid a negative variance. The truncation function facilitates this requirement, but in doing so distributions with zero variance are frequently considered in the likelihood calculations. The set of conditions in section 2.4.1 ensure that undefined quantities do not occur when calculating r. In short, the conditions make sure that any descendant of a population, whose frequency at a given SNP is zero or one, is also zero or one at that SNP. This must be the case as mutation and migration are not permitted. The result is that at some SNP loci, allele frequency distributions occur with infinite spikes at the boundaries. Remembering that information about c_i comes from the distribution of β across SNPs in population i, and that frequencies will tend to reach the boundaries when the c's are large, there appears to be a contradiction, since the c's being large will tend to move the frequencies towards the boundaries, which in turn will produce distributions of the β's with infinitely small variation at a particular boundary value. In the previous example for two populations, identifiability was an issue when the c's were small; in this case a contradiction occurs when the c's are large. Taken in conjunction, these examples suggest some possibilities as to the cause of the problems when fitting the new model.

As a solution and also to clarify the cause of the non-identifiability, it was decided to simplify the new model such that the variance of the distribution of allele frequencies does not depend on the mean. Although the accuracy of the model, in a population genetics sense, may suffer, it was considered worthwhile, since the fit of various models can be assessed, given that the simpler model is identifiable.

To define the simplified version of the new model, expressions [15], [16] and [18] remain unchanged, but expression [17] is reduced to

$$\begin{aligned}&\beta_{ij} \sim \text{Normal}\left(\beta_{a(i),j}, c_i\right), i = 1, \ldots 2P-2; j = 1, \ldots, L, \\ &\text{independently } \forall\, i, j.\end{aligned} \qquad [23]$$

Since the variance term in expression [23] does not depend on $\beta_{a(i),j}$, the truncation function is not needed to ensure negative variances do not occur. This model is not only easier to implement using MCMC, but information is not lost when using the truncation function. A similar model is implemented and fitted using restricted maximum likelihood (REML) using the CONML option within the phylogenetic package PHYLIP (Felsenstein, 1993), which can also be used to construct trees.

When fitting the simplified model using MCMC, particularly the Metropolis-Hastings algorithm, identical simplifications can be made when calculating the ratio r as were made for the full extension to the ND model (see section 2.4.1 for details); only the distribution in [23] is substituted where appropriate. The model is more straightforward to fit since the conditions in section 2.4.1 are not implemented and it is also hoped that removing such restrictions ameliorates the identifiability issue.

3.7 Simplified Model – An Example

Presented in this section are some graphical summaries from an MCMC analysis under the simplified model using simulated data. Data were simulated from four populations at 100 SNP loci under the simplified model as follows:

1. Draw β_{ij} from Beta(2, 2) where $i = 2P–1$ and $j = 1, \ldots, L$.

2. Draw $\beta_{ij} \left| \beta_{a(i),j}, c_i\right.$ from $\text{Normal}\left(\beta_{a(i),j}, c_i\right)$ where $i = 1, \ldots, 2P–2, j = 1, \ldots, L$, $\boldsymbol{c} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$, $\boldsymbol{a} = (5, 5, 6, 6, 7, 7, 0)$.

3. Draw $x_{ij} \left| \beta_{ij}, n_{ij}\right.$ from $\text{Binomial}\left(n_{ij}, t\left(\beta_{ij}\right)\right)$ for $i = 1, \ldots, P, j = 1, \ldots, L$.

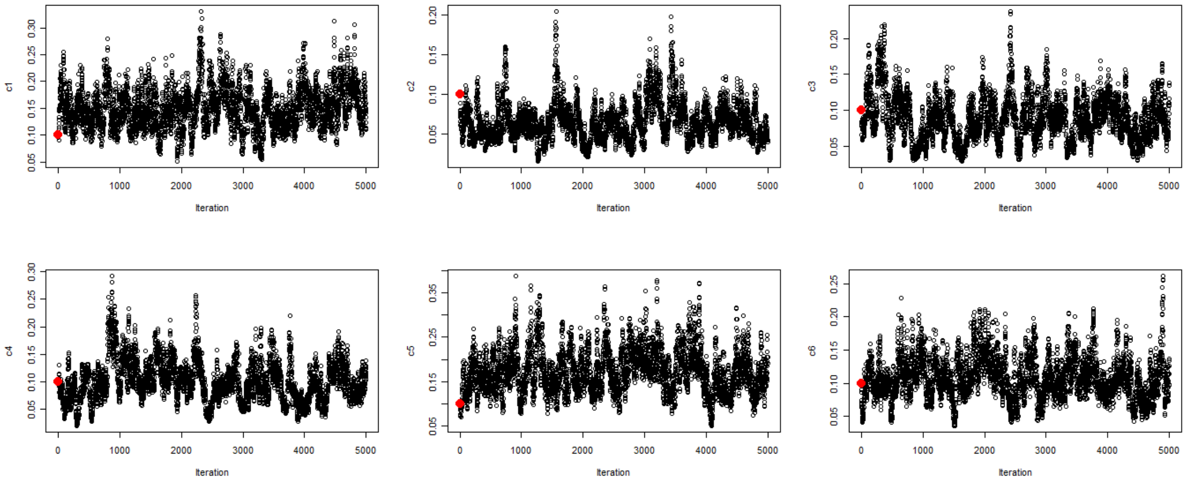


Figure 3-32 Trace plots of the *c*'s from the simplified model fitted using MCMC, without removing burn-in with a run length 5000. Red dots indicate true values.

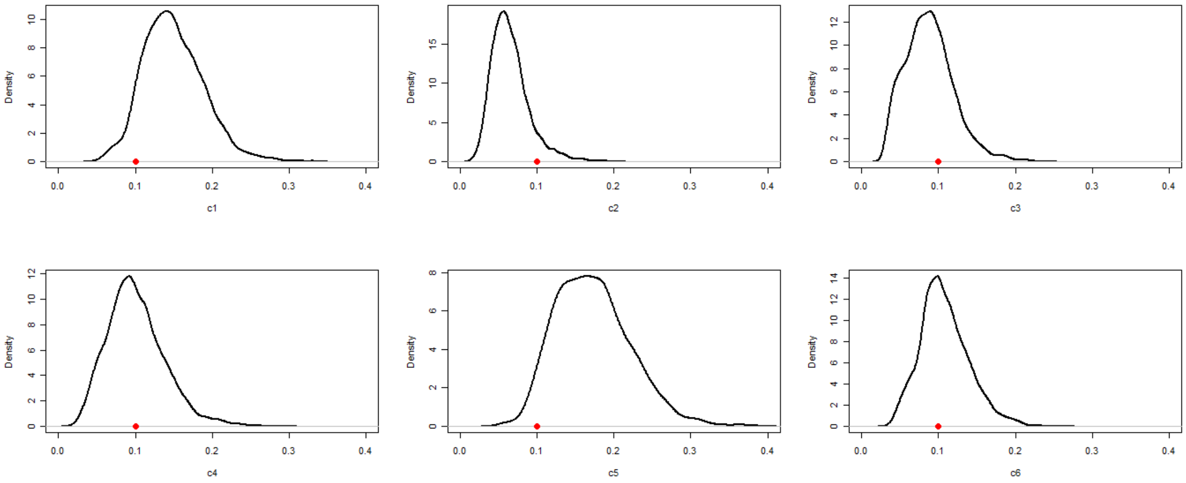


Figure 3-33 Posterior density plots of the *c*'s from the simplified model fitted using MCMC, without removing burn-in with a run length of 5000. Red dots indicate true values.

An MCMC analysis was performed with a run length of 5000 using the standard initial configurations and prior distributions. From the plots in Figure 3-32 it appears the problems that arose when fitting the new model are not encountered for the simplified model. All the chains for the *c*'s mix sufficiently and extreme values are not accepted. It is also clear from Figure 3-33 that the MCMC estimation procedure does rather well as regards the location of the estimated posterior distribution, since the true values are well within an acceptable range from the centre of the estimated distributions.

The plots in Figures 3-34 and 3-35 indicate that the allele frequency parameters are estimated well and the chains for the individual parameters mix well. Another characteristic of the β's is the increased precision when estimating contemporary frequencies, relative to ancestral frequencies; a property observed when fitting the ND model (see section 2.3, example 3).

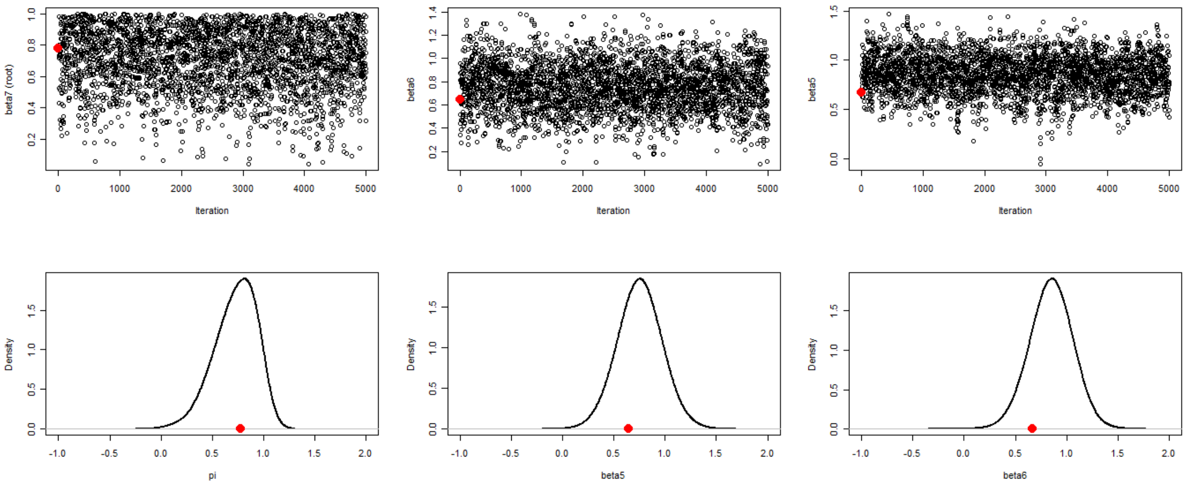


Figure 3-34 Trace plots and posterior densities of the ancestral allele frequencies (β_5, β_6, β_7) from an arbitrarily chosen SNP under the simplified model. Model fitted using MCMC, with a run length of 5000 without removing burn-in. The red dots indicate true values.

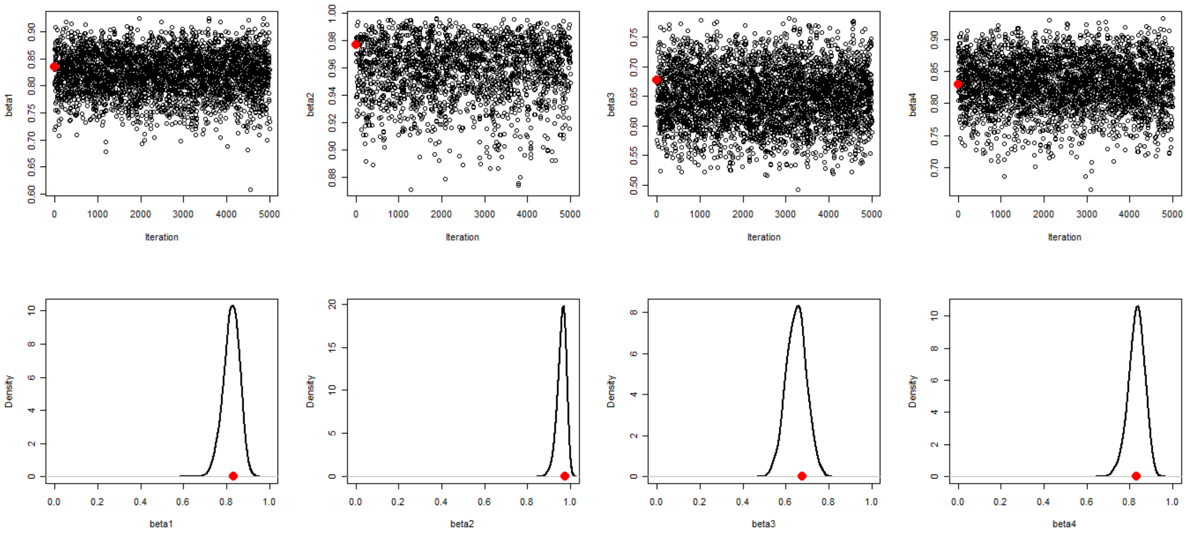


Figure 3-35 Trace plots and posterior density plots of the contemporary allele frequencies (β_1, β_2, β_3, β_4) from an arbitrarily chosen SNP under the simplified model. Model fitted using MCMC, with a run length 5000 without removing burn-in. Red dots indicate true values.

From the analysis reported in this section and numerous others performed under the simplified model, it can be concluded that identifiability is not an issue. This presents an opportunity to fit various topologies and labelled histories and investigate which is the most appropriate using residual diagnostics. These tasks constitute the remainder of this thesis.

3.8 Simulation under Simplified Model

In this section, the potential for inferring the most likely labelled history for a set of populations using residual diagnostics is explored. The leave-one-out diagnostic is unsuitable when considering bifurcating topologies of more than two populations, since in most instances the interpretation of a particular branch changes when a population is removed, and so one would expect to see instability in parameter estimates, even when the correct labelled history is specified. For this reason it was decided to rely solely on residuals for making judgements. Although this method does not provide a quantitative model selection criterion, its simplicity over methods such as Bayes factors made it appealing for our purposes.

The hierarchical structure of the ND model stipulates *P* populations descending from a single ancestral population and is reflected in the set of standardised residuals (see expression [13]). The generic form of a residual is the difference between some true value and the estimate of the true value provided by fitting the model, standardised by the standard deviation of the estimate. In the case of the ND model, the true values are the contemporary allele frequencies and the estimates are the corresponding ancestral frequencies. But since the contemporary frequencies are themselves estimated, albeit relatively well, the standardisation factor must be inflated to account for the added uncertainty (see proof in Appendix A). The situation becomes more complex when bifurcating topologies are considered, since internal proto-populations are both ancestral to some populations and descendants of other proto-populations.

For the simulation studies in this section, the set of standardised residuals are defined by:

$$\frac{\beta_{ij} - \hat{\beta}_{a(i),j}}{\sqrt{\hat{c}_i}}; i = 1, \dots, 2P - 2; j = 1, \dots, L. \qquad [24]$$

Two important points must be made in relation to the formula in expression [24]. In the following simulation analyses, the contemporary allele frequencies are considered fixed and known. This is reasonable since given a large enough sample, the contemporary allele frequencies are well estimated by the sample frequencies. Therefore the binomial sampling step in the hierarchy is removed, meaning that an inflated variance is not needed in the denominator of the residual formula. The second point only applies to the residuals for the internal nodes. It is clear that the allele frequencies at internal nodes are estimated during the MCMC procedure but using expression [24], are considered known; a similar situation as in the ND model. It was decided to substitute the mean of the appropriate chain from the MCMC analysis for the true ancestral frequencies, bearing in mind when making any judgements, that the standard deviations of the residuals from the internal nodes have probably been under-estimated.

All simulated data sets in this section include four populations and have been simulated under the topology shown in Figure 3-36, where there are two pairs of populations with shared ancestry. Under this particular topology there are three possible labelled histories, which allows a sufficient but manageable number of model comparisons to be made. The alternative topology has 12 potential labelled histories (4 MRCAPs × 3 orderings for each MRCAP) and so only the topology in Figure 3-36 was considered. This topology is by no means the most accurate and it is potentially more interesting to infer the most likely topology. However it was decided that the capability of the model selection process must be assessed first, using the most practically convenient topology, before proceeding with more detailed analyses.

The first data set was simulated under the new simplified model (see section 3.6) with a labelled history defined by a = (5, 5, 6, 6, 7, 7, 0) and c = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1). Three MCMC analyses were performed using the same data set; only changing the labelled history. Each analysis corresponds to a particular labelled history from

Figure 3-36, with a run length of 10000 iterations and prior distributions identical to those used in previous analyses.

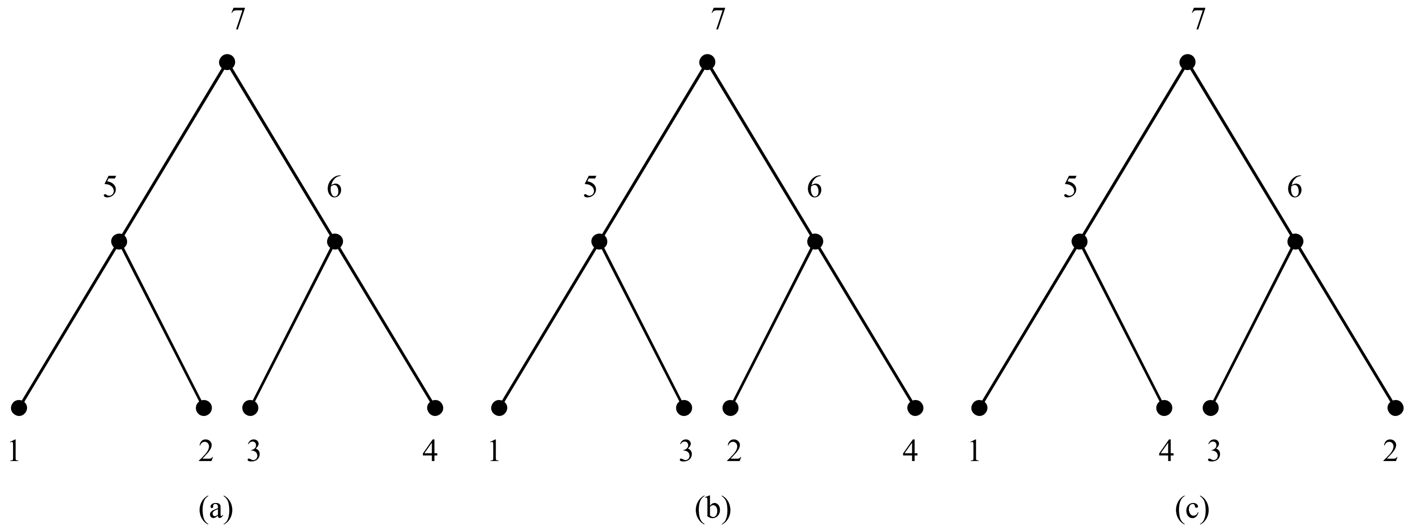


Figure 3-36 Labelled histories of the three MCMC analyses. The data were simulated under (a), and subsequently analysed under all three topologies.

To investigate whether there is sufficient information in the data to infer the correct labelled history (a) out of (a), (b) and (c) (Figure 3-36), residuals were compared between the three analyses; the rationale being that given that the data were simulated under (a) the residuals for the incorrect models (b) and (c), should show lack of fit, or at least a worse fit than (a). As a preliminary to the residual analysis, the performance of the model was checked for each of the analyses. Figure 3-37 does not indicate any problems when fitting the model using the correct labelled history, since all the chains appear to mix well and the estimated posterior densities all contain the true value. Figures 3-38 and 3-39 highlight the effects of specifying the incorrect labelled history. When populations are re-arranged, the model attempts to accommodate this by adjusting the estimates of the c's. The c's corresponding to outer branches (1-4) are over-estimated, since the data are congruent with population 1 and 2 and populations 3 and 4 being closely related. The c's for the internal branches (5-6) are under-estimated for exactly the same reason; the model is attempting to reduce the distance between population pairs 1 and 2, and 3 and 4. Ultimately, the model does not manage to compensate for the alternative labelled histories and the actual distances between populations are not recovered. The question is then: can the residuals recover these discrepancies and distinguish the correct labelling? It is also worth noting that the re-arrangements do not affect the performance of the MCMC algorithm, since the chains still appear to mix adequately.

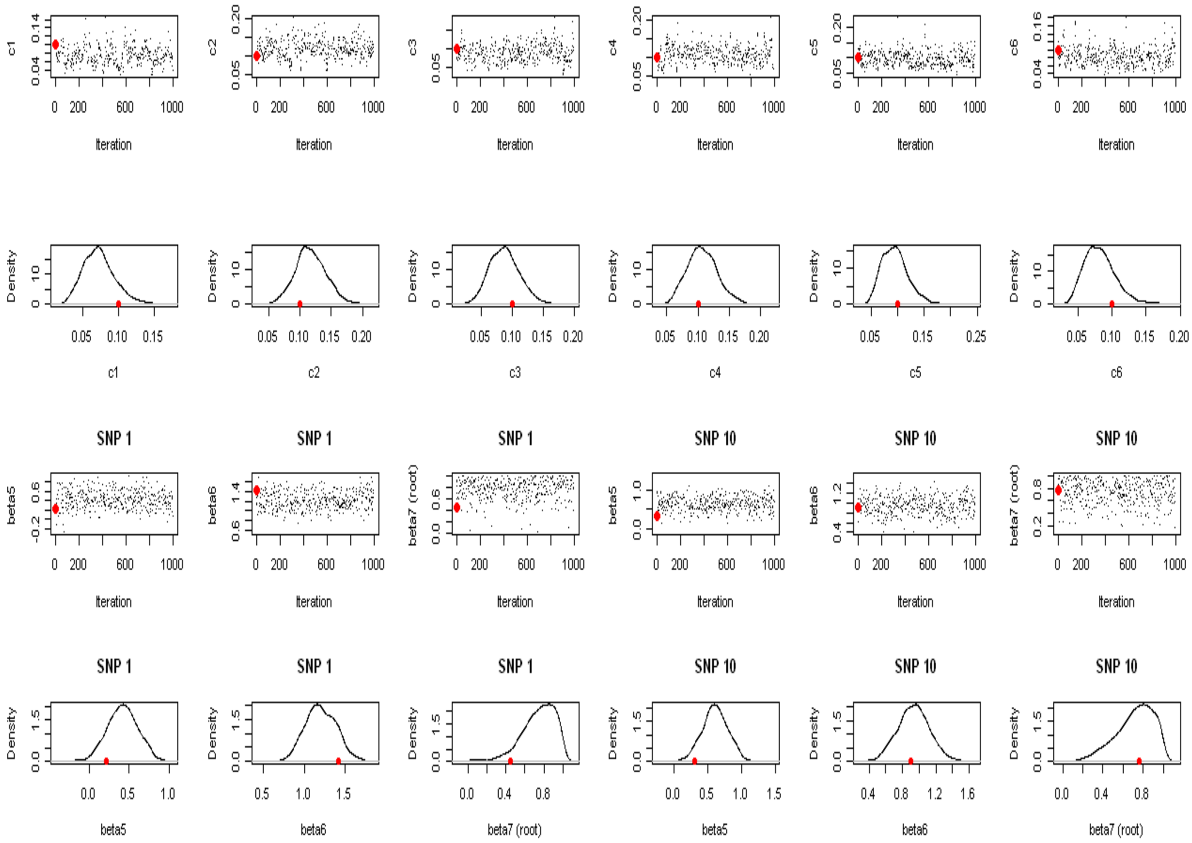


Figure 3-37 Graphical summaries from MCMC analysis under the correct labelled history (a) (Figure 3-36). Note that SNPs 1 and 10 were arbitrary choices.

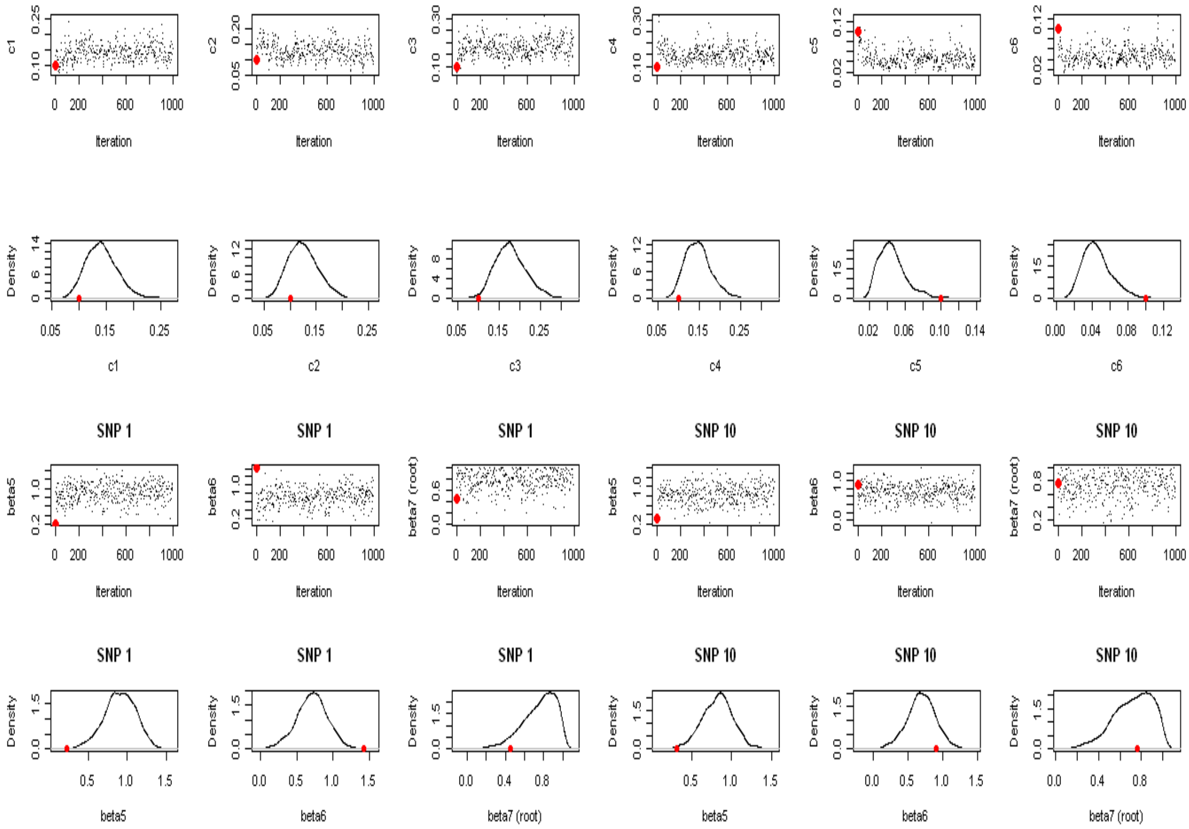


Figure 3-38 Graphical summaries from MCMC analysis under incorrect labelled history (b) (Figure 3-36). Note that SNPs 1 and 10 were arbitrary choices.

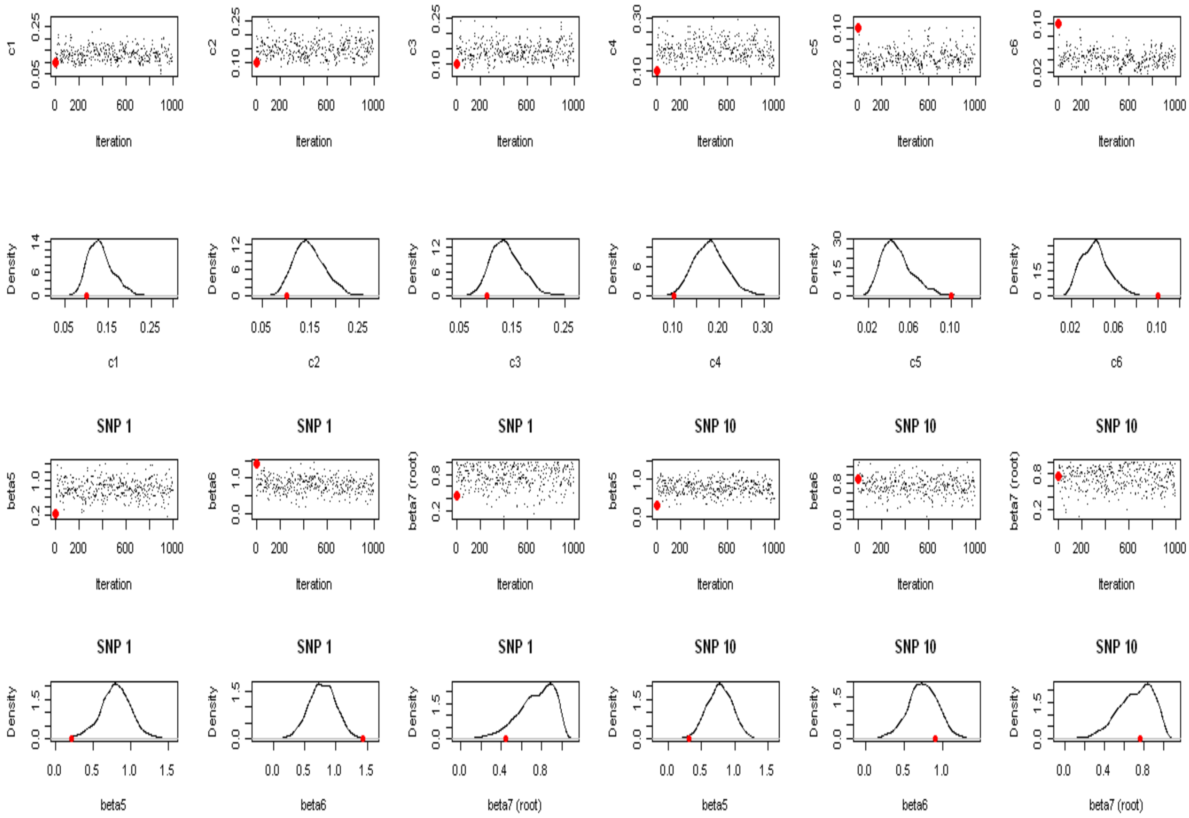


Figure 3-39 Graphical summaries from MCMC analysis under incorrect labelled history (c) (Figure 3-36). Note that SNPs 1 and 10 were arbitrary choices.

Looking at the plots in Figure 3-40, the first point to note is that, for all three analyses, the residuals appear to have mean zero. However there is a suggestion that constant variance is violated in all three cases. The sample Pearson correlation between the residuals and fitted values for analyses (a), (b) and (c) are 0.3836, 0.3842 and 0.3797 respectively, suggesting slight positive correlation. Importantly, all are very similar meaning that without knowing which residuals correspond to the correct analysis, it would be very difficult to make any definite assertions as to the correct labelled history. The same can be said for the QQ plots in Figure 3-41, that the correct labelled history is not distinguishable. In all three cases normally distributed noise does not seem implausible.

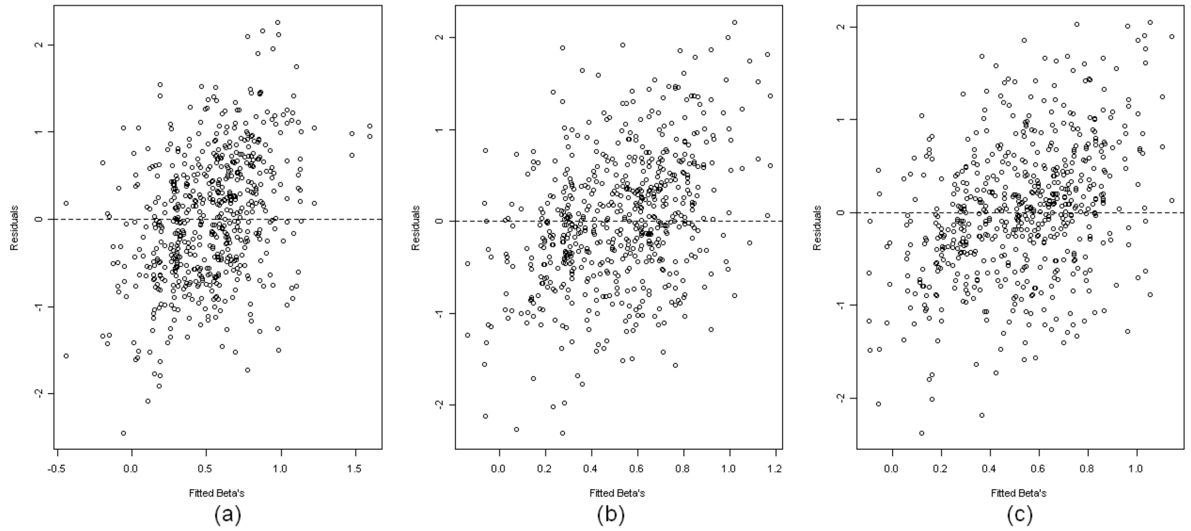


Figure 3-40 Residual plots from three MCMC analyses on the same data set. The residuals in (a) are from the analysis using the correct labelled history; (b) and (c) are both from analyses under incorrect labelled histories.

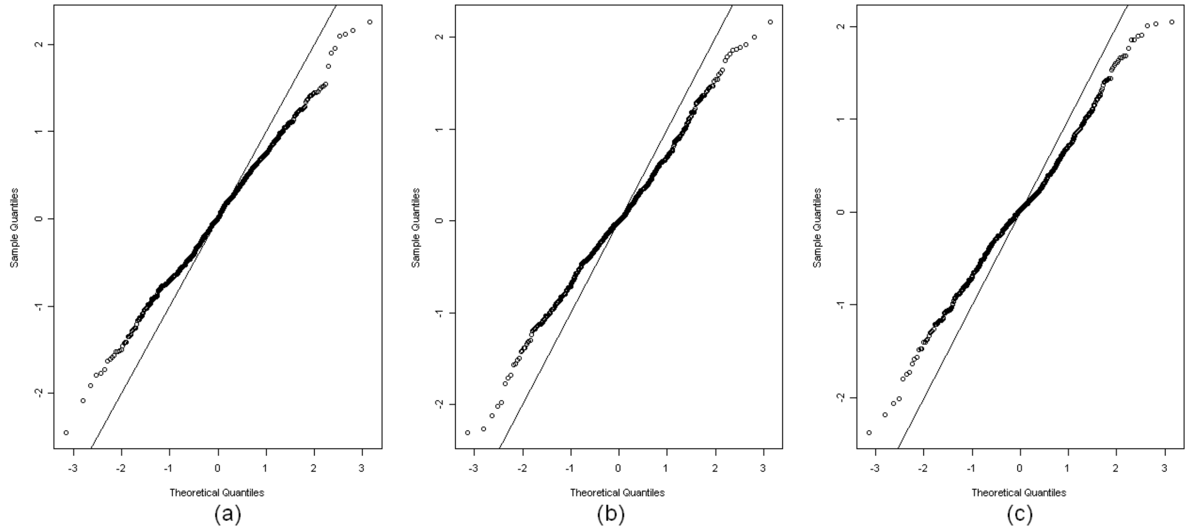


Figure 3-41 Normal QQ plots using residuals from three MCMC analyses on the same data set. The residuals used in (a) are from the analysis using the correct labelled history; (b) and (c) are both from analyses under incorrect labelled histories.

This example suggests that the residuals are not informative when attempting to choose the correct labelled history. This may well reflect the inability of the residuals in general to recover the required information or it may be due to the rather unrealistic *c* configuration used in this example. When all the *c*'s are equal, re-arranging the labelling does have an effect on the estimates of the *c*'s but not a particularly profound one. To investigate a more realistic scenario, data were simulated under the same topology but the *c*'s were allowed to vary throughout the tree.

The second data set was simulated under the new simplified model with the same labelled history as before, only this time *c* = (0.05, 0.05, 0.3, 0.3, 0.1, 0.1). Three MCMC analyses were performed as before, changing the labelled history, and with the same prior distributions.

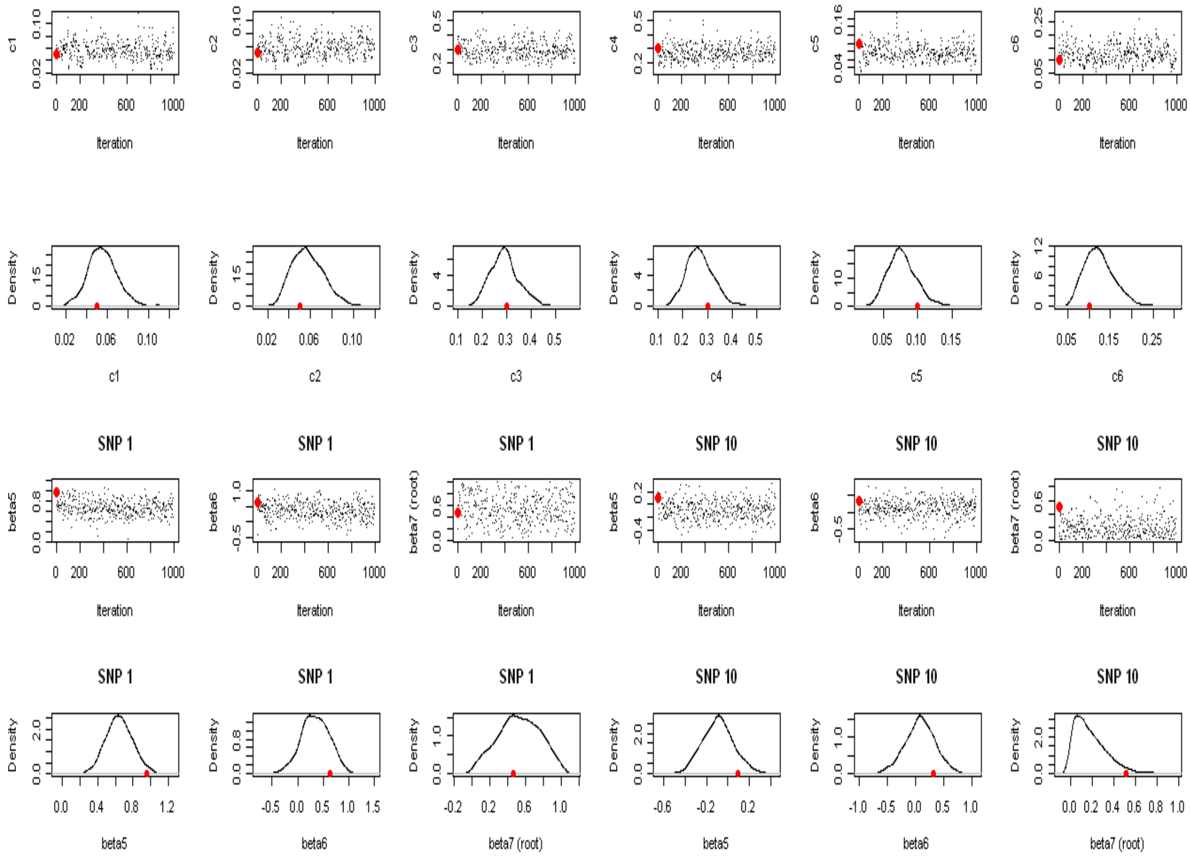


Figure 3-42 Graphical summaries from MCMC analysis under correct labelled history (a) (Figure 3-36).

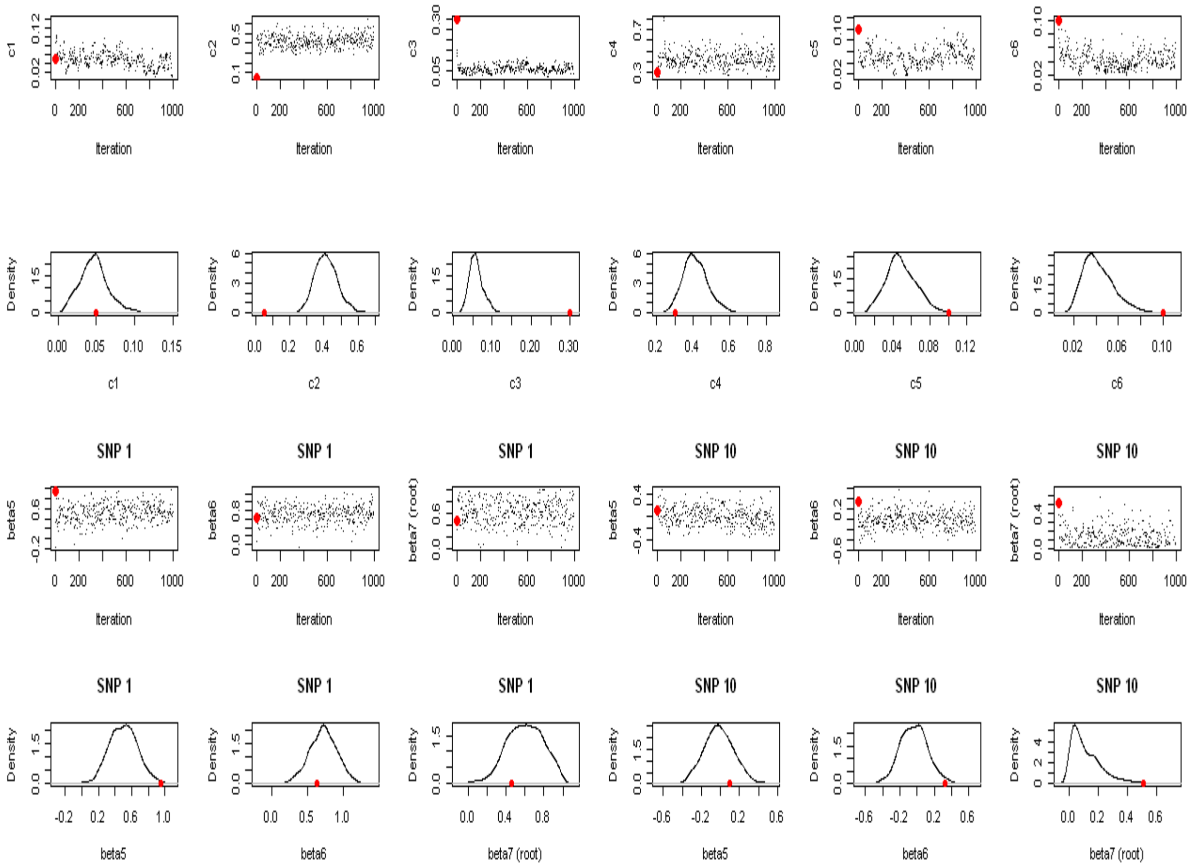


Figure 3-43 Graphical summaries from MCMC analysis under incorrect labelled history (b) (Figure 3-36).

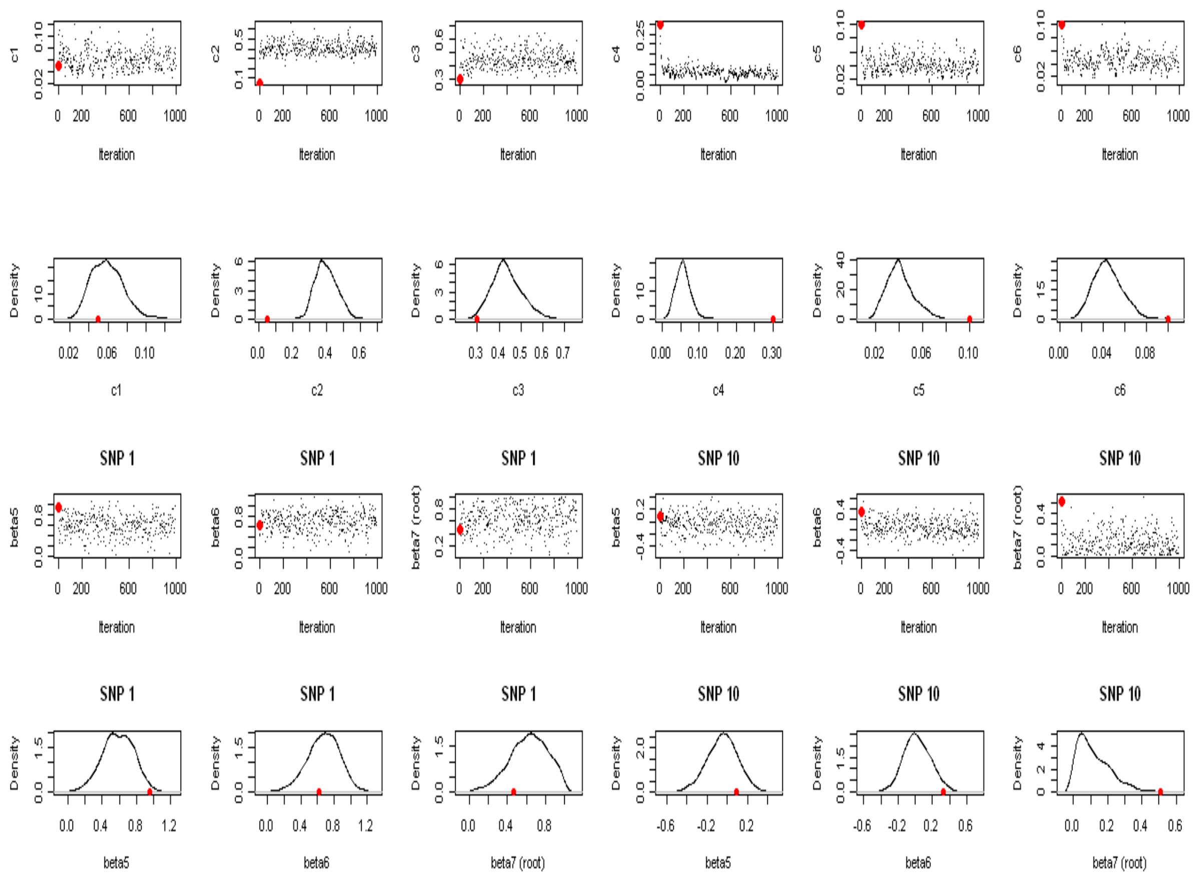


Figure 3-44 Graphical summaries from MCMC analysis under incorrect labelled history (c) (Figure 3-36).

As in the previous example, the MCMC algorithm appears to perform well when using the correct labelled history, as indicated by the plots in Figure 3-42. The estimates of the c's again undergo adaptation to the re-arrangement of populations. In the previous example the variation of allele frequencies across SNPs was the same in every population since the c's were all the same. Therefore, when the labelled history was changed, the model implicitly used differences between population allele frequencies to recover information. If this wasn't the case, the estimates would not have changed since the variation does not change, just the labelling. In this example, the c's are different throughout the tree so the data carries additional information. This becomes clear when the labelled history is changed. The value of c_2 in Figure 3-43 is very over-estimated, simply because, given the particular labelling, the data reflects a population with much higher variation in allele frequencies (population 3, c = 0.3). The opposite occurs for the estimate of c_3 since population 2 has been assigned in its place, which has a small c (population 2, c = 0.05) and therefore very little variation in allele frequencies across SNPs. These same properties can be seen in Figure 3-44. The values of the internal c's (c_5 and c_6) are also under-estimated in both incorrect analyses, which is likely a compromise made to minimise the discrepancies between the differences between population frequencies, resulting from the re-arrangement of populations.

Considering the residual plots in Figure 3-45, the analysis under the correct labelled history does appear to produce residuals more consistent with the modelling assumptions than those from the incorrect analyses. Plots (b) and (c) suggest some positive correlation between the residuals and the fitted frequencies; although not particularly strong correlations, clearly stronger than in plot (a). Sample Pearson correlations of 0.0393, 0.2777, 0.2921 for analyses (a), (b) and (c) respectively, confirm this. The QQ plots in Figure 3-46 are less encouraging since all three plots are fairly similar, and all indicate that normality is not implausible.

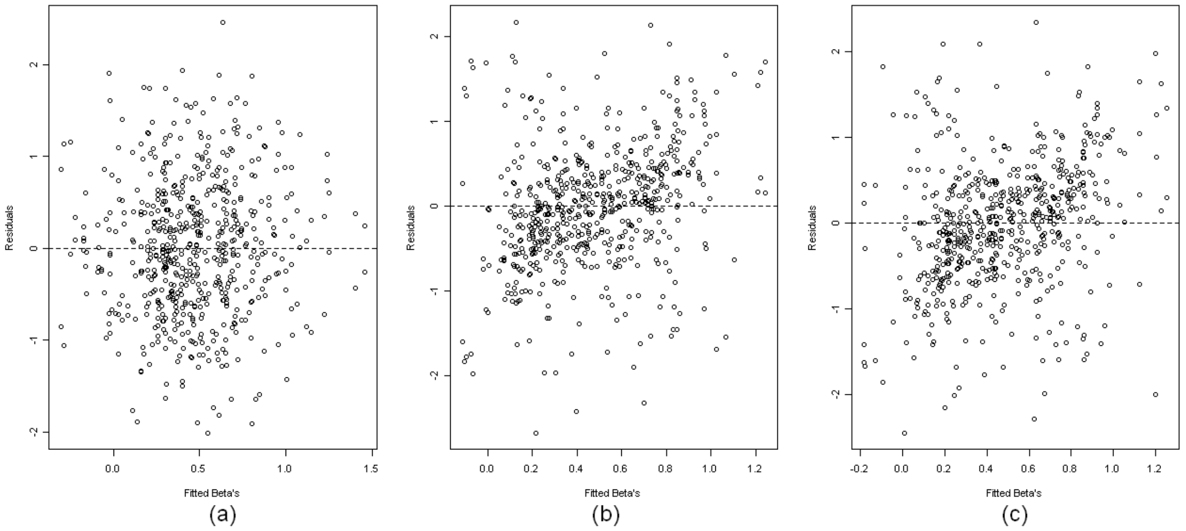


Figure 3-45 Residual plots from three MCMC analyses on the same data set. The residuals in (a) are from the analysis using the correct labelled history; (b) and (c) are both from analyses under incorrect labelled histories.

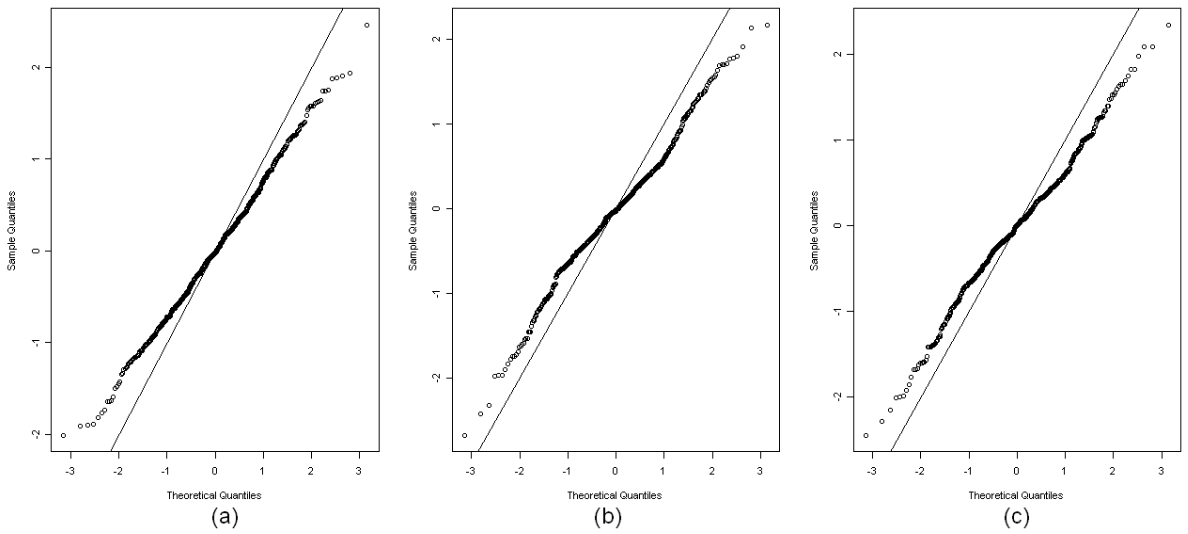


Figure 3-46 Normal QQ plots using residuals from three MCMC analyses on the same data set. The residuals used in (a) are from the analysis using the correct labelled history; (b) and (c) are both from analyses under incorrect labelled histories.

This second example is much more encouraging than the first and suggests that not only is there information in the data to infer the correct labelled history, but that residuals are capable of recovering it.

3.9 Global Data Set 1 - Analysis under Simplified Model

In this section the data set analysed in section 3.4 under the ND model is re-analysed under the new simplified model in an attempt to infer the most likely labelled history for these populations. As a reminder, global data set 1 included a Biaka pygmy population, a North African Mozabite population, a Cambodian population and a Native American Pima population. Figure 3-47 shows the three possible labelled histories using the balanced topology. One might expect the African populations to be grouped together, with the Cambodian and Pima populations forming the other group (Figure 3-47 (a)) due to the close geographic proximity of the Africans and the aforementioned settlement of the Americas from South East Asia. These groupings are supported by the correlations in Figure 3-21, but the correlation between the Mozabite and Cambodian populations may affect the inference. The balanced topology was used for purely practical reasons as it provides a more manageable number of groupings, as compared to the unbalanced alternative, but is not *a priori* a more accurate description.

Each model in Figure 3-47 was fitted to global data set 1. As in previous analyses a log-normal prior on the *c*'s and a uniform prior on (0, 1) on the MRCAP frequencies was used. The *c*'s were started from $F_{ST} = 0.0594$, the frequencies at the tips of the tree (β_1, β_2, β_3 and β_4) from the sample frequencies and the ancestral frequencies (β_5, β_6 and β_7) at 0.5. Each MCMC chain was run for 25000 iterations and an appropriate burn-in removed.

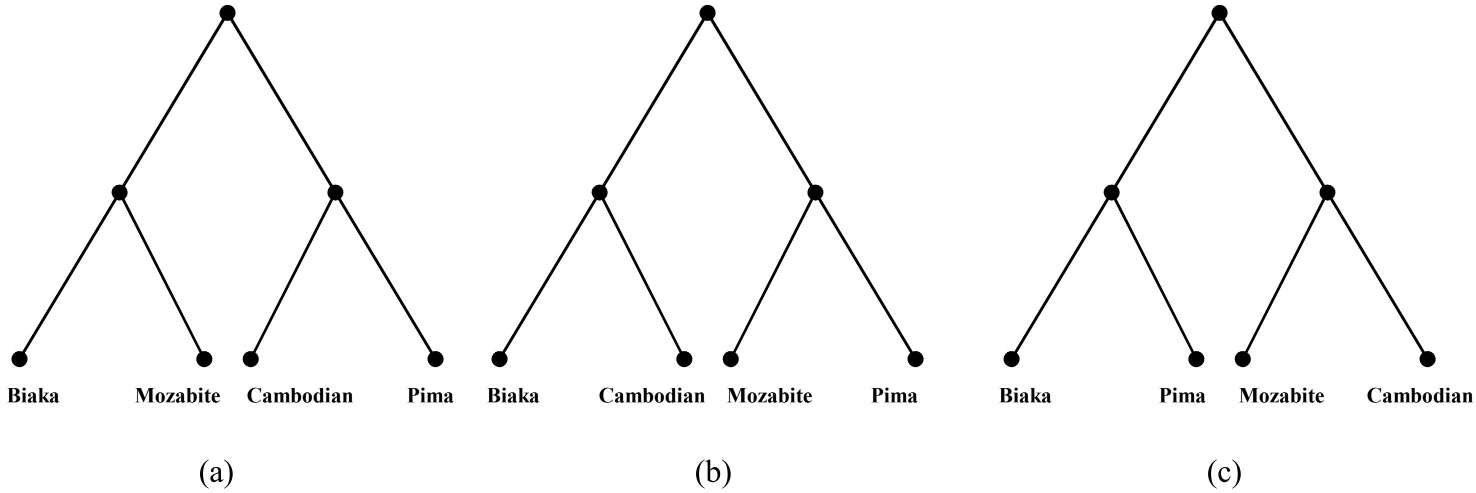


Figure 3-47 Three labelled histories used under the new simplified model fitted to global data set 1.

The three sets of residuals in Figure 3-48 are calculated using the formula in [25], which is slightly different from expression [24], since the sample frequencies are used instead of the population frequencies. Formally, the denominator should then contain a factor to inflate the variance, but since the sample frequencies are approximately the population frequencies, this formula is sufficient.

$$\frac{x_{ij}/n_{ij} - \hat{\beta}_{a(i),j}}{\sqrt{\hat{c}_i}}; i = 1, \dots, 2P - 2; j = 1, \dots, L. \qquad [25]$$

The residual plots in Figure 3-48 seem to suggest that the model using labelled history (a) from Figure 3-47 fits the data best. Plots (b) and (c) exhibit less variance towards the boundary values of the fitted β's, which violates the assumption of constant variance. This feature is also found in plot (a) but to a lesser extent. Sample Pearson correlations between the residuals and fitted values from analyses (a), (b) and (c) are 0.0600, 0.0977 and 0.1048 respectively, which again suggests that analysis (a) produces residuals most consistent with the modelling assumptions. The QQ plots in Figure 3-49 do not appear to offer any insight into the most appropriate labelled history since all are fairly similar. In all three cases normality does not seem plausible. Therefore from Figure 3-48 and the sample correlations, labelled history (a), which groups the African populations and the Cambodian and Pima populations together, appears to be the most likely of the three for these data. Although the signal in the residuals is not particularly strong, it is satisfying that a rather informal model selection procedure produces results consistent with current knowledge for a real data set.

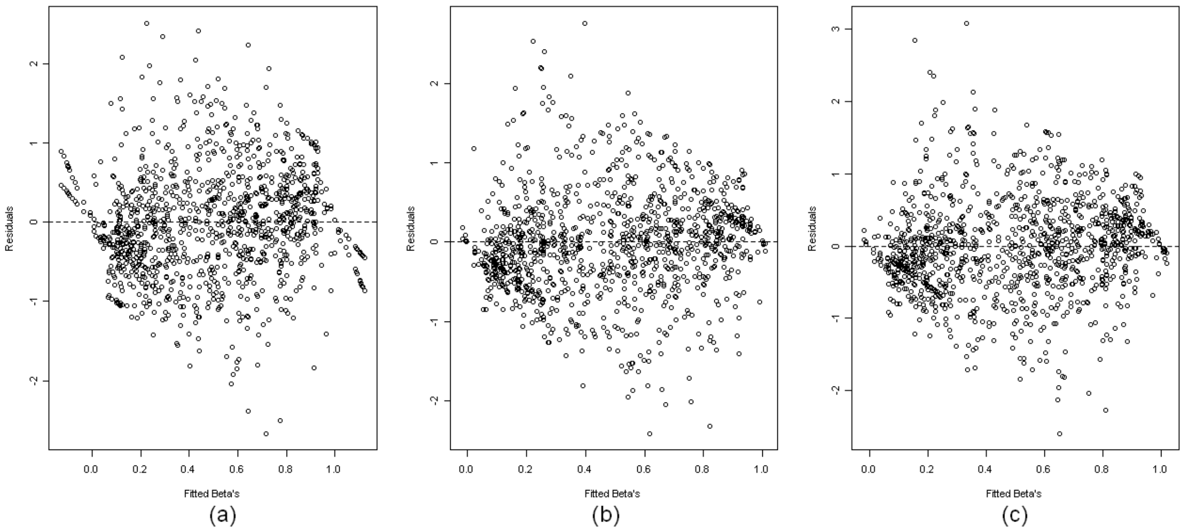


Figure 3-48 Residual plots from three MCMC analyses on global data set 1. The labels (a), (b) and (c) refer to the labelled histories in Figure 3-47.

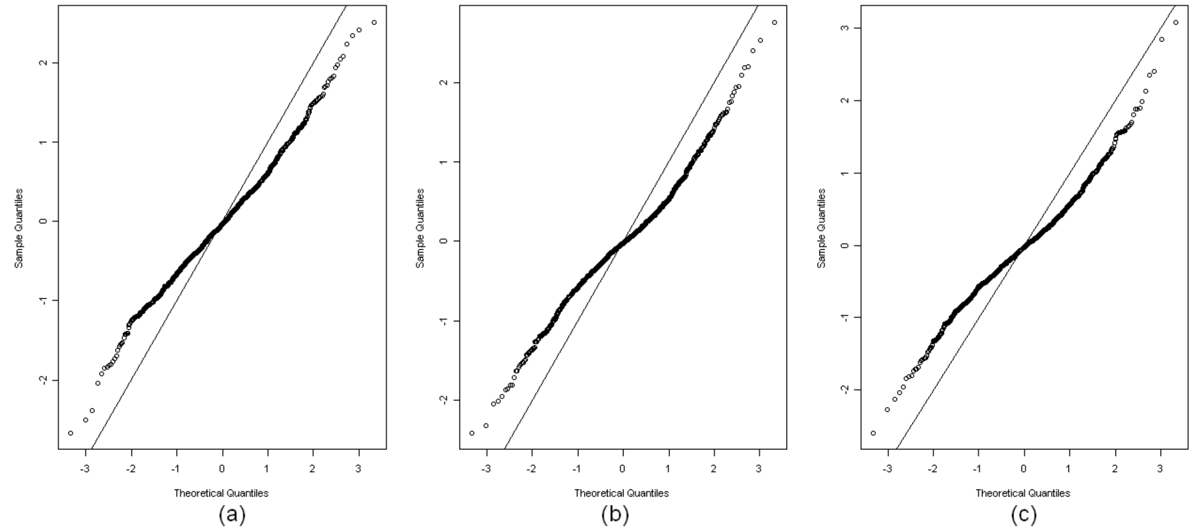


Figure 3-49 Normal QQ plots using residuals from three MCMC analyses on global data set 1. The labels (a), (b) and (c) refer to the labelled histories in Figure 3-47.

Another interesting comparison between the residuals from the analysis under the ND model (section 3.4) and analysis (a) from the current section is shown in Figure 3-50. It is difficult to draw any conclusions from Figure 3-50 since both plots highlight issues with constant variance. Sample Pearson correlation for (a) is 0.0994 and for (b) is 0.0600 which suggests that the new model produces a better fit to the data. The normal QQ plots in Figure 3-51 also suggest that the new model yields a superior fit since plot (a) indicates a lack of symmetry, which is a defining characteristic of the normal distribution. Plot (b), on the other hand, indicates that the distribution of the standardised residuals has rather light tails but is still symmetric about the mean.

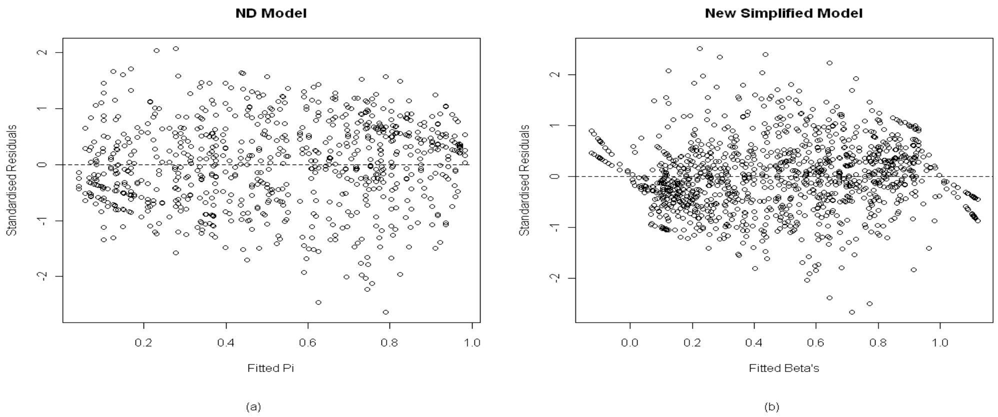


Figure 3-50 Standardised residuals vs fitted values plots for analysis of global data set 1 under (a) ND model and (b) new simplified model.

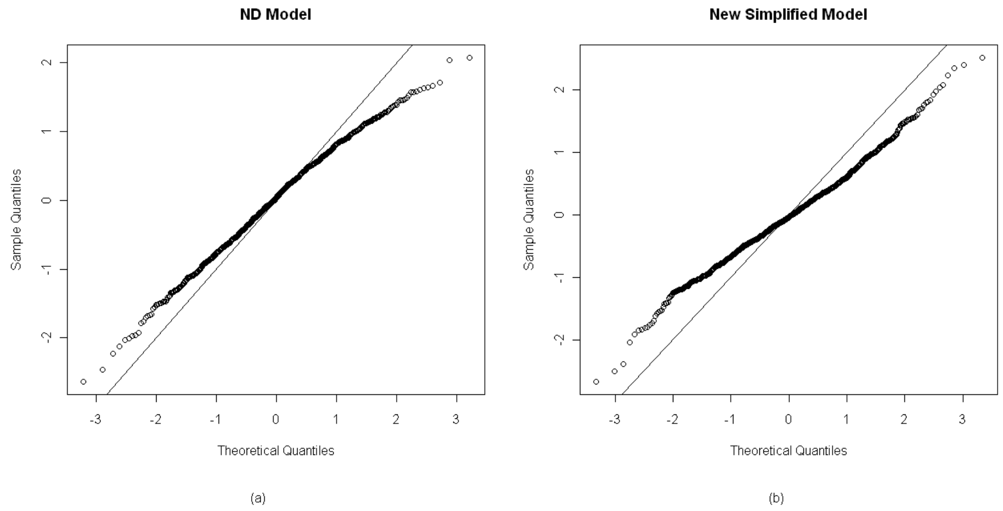


Figure 3-51 Normal QQ plots for analysis of global data set 1 under (a) ND model and (b) new simplified model.

Figures 3-52 and 3-53 show some graphical output regarding the c's from the MCMC analysis using labelled history (a). The trace plots show that the chain appears to mix well enough for all the c's and the estimated posterior distributions are all uni-modal and well-behaved. The estimates of the c's from the branches directly above the contemporary populations are interesting. The Pima population is still the most differentiated from its most recent common ancestor (with the Cambodians) but its value has decreased by a factor of approximately ten ($\hat{c}_4$ = 0.0723, Table 14), which still coincides with the explanation given previously since it is still the largest c of all the populations. The branch above the Mozabite population still has a small c ($\hat{c}_2$ = 0.0074, Table 14) probably reflecting its shared ancestry with Europeans. The c above the Cambodian population is also a lot smaller under the new model ($\hat{c}_3$ = 0.0195, Table 14) compared with the estimate using the ND model ($\hat{c}_3$ = 0.2420, Table 10), which seems more likely value for a south East Asian population (Nicholson et al., 2002). However, the estimate of c above the Biaka population does seem slightly small for Sub-Saharan Africa ($\hat{c}_1$ = 0.0385, Table 14).

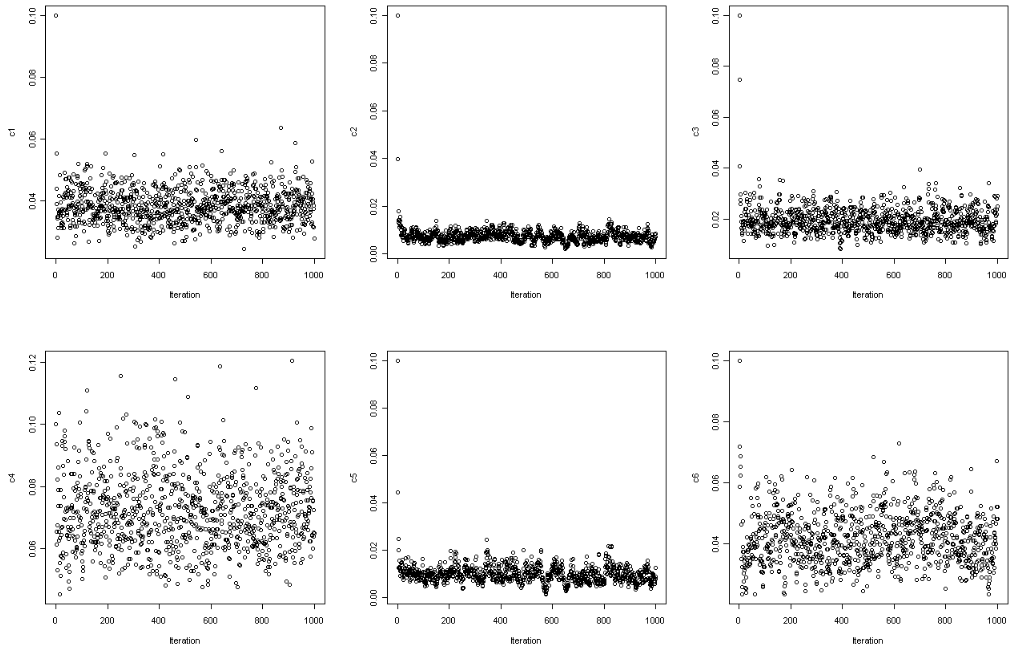


Figure 3-52 Chains for the *c*'s from the analysis under labelled history (a) Figure 3-47. Chains thinned by a factor of 25.

Table 14 Summaries of MCMC results for global data set 1.

Parameter	***Mean***	***Posterior Standard Deviation***	***90% Credible Region***
c_1	*0.0385*	*0.0056*	*(0.0303, 0.0478)*
c_2	*0.0074*	*0.0029*	*(0.0041,0.0109)*
c_3	*0.0195*	*0.0054*	*(0.0126, 0.0278)*
c_4	*0.0723*	*0.0119*	*(0.0547, 0.0940)*
c_5	*0.0099*	*0.0038*	*(0.0051, 0.0158)*
c_6	*0.0414*	*0.0084*	*(0.0289, 0.0561)*

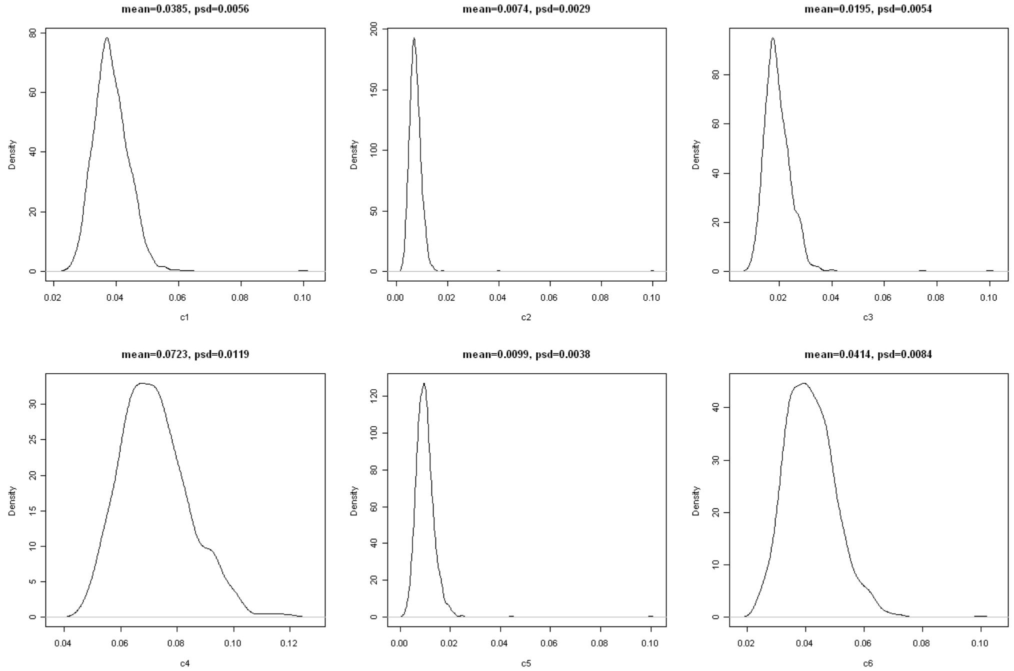


Figure 3-53 Estimated posterior density plots with means and p.s.d's for the *c*'s using labelled history (a) Figure 3-47.

In this section the residual diagnostics have been used to infer the most likely labelled history for global data set 1, given that the balanced topology is correct. The groupings suggested seem plausible since the African populations reside on one side of the tree and the Cambodians and Pima on the other. Using the new model also yields very different magnitudes for the values of the *c*'s, although the ordering of magnitude remains unchanged. When attempting to interpret the values of *c* under the simplified model one must remain sceptical since the simplifications made in order to be able to fit the model are not well justified by any population genetics theory. Nevertheless the model should still represent a close approximation to the more accurate but non-identifiable model, and the improvement it provides over the ND model is encouraging.

3.10 Global Data Set 2 – Analysis under Simplified Model

Here the second global data set is re-analysed under the new simplified model in an attempt to infer the most likely labelled history for these populations. Global data set 2 included a sample from an African Mandenka population, an Mbuti pygmy population, a French population and a Tuscan population. Figure 3-54 shows the three possible labelled histories using the balanced topology. The most obvious grouping would be the two African populations together and the two Europeans together (Figure 3-54 (a)), and these groupings are supported in the pairwise allele frequency plots in Figure 3-25.

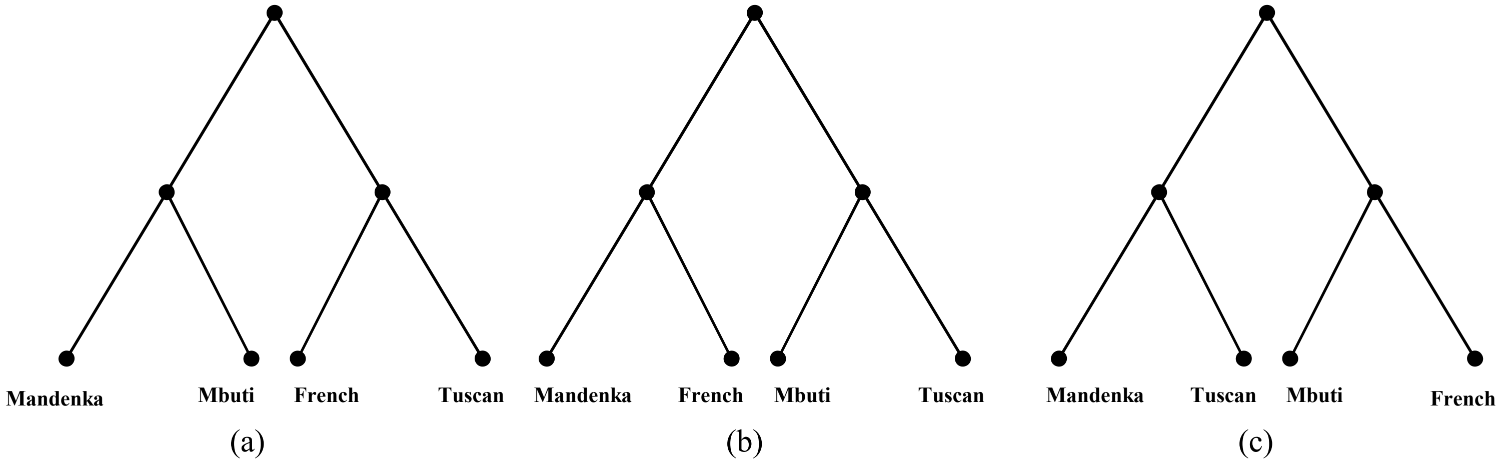


Figure 3-54 Three labelled histories used under the new simplified model fitted to global data set 2.

Each model in Figure 3-54 was fitted to global data set 2. The same prior distributions and starting values as in section 3.9 were used, except that $F_{ST} = 0.0425$ for these data.

The residuals plots in Figure 3-55 are all fairly consistent with the modelling assumption of constant variance, particularly plots (b) and (c). What is clear though is that the residuals are not informative about the most appropriate labelled history for these data since plots (b) and (c) are very similar. In fact the most obvious grouping mentioned before, from analysis (a), appears to yield the worst fit of the three analyses, which is very surprising. Even if one were to reject labelled history (a), which does not seem sensible, the best fit is still not clear. The sample Pearson correlations for analyses (a), (b) and (c) between the standardised residuals and the fitted values are 0.0967, 0.0498 and 0.0483 respectively. Again notice the similarity between (b) and (c). The QQ plots in Figure 3-56 suggest that labelled history (a) fits the data the best.

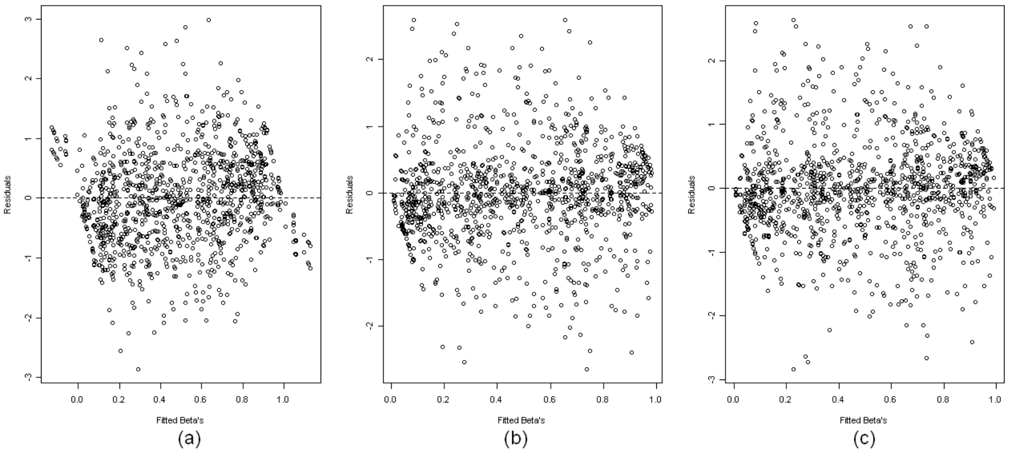


Figure 3-55 Residual plots from three MCMC analyses on global data set 2. The labels (a), (b) and (c) refer to the labelled histories in Figure 3-54.

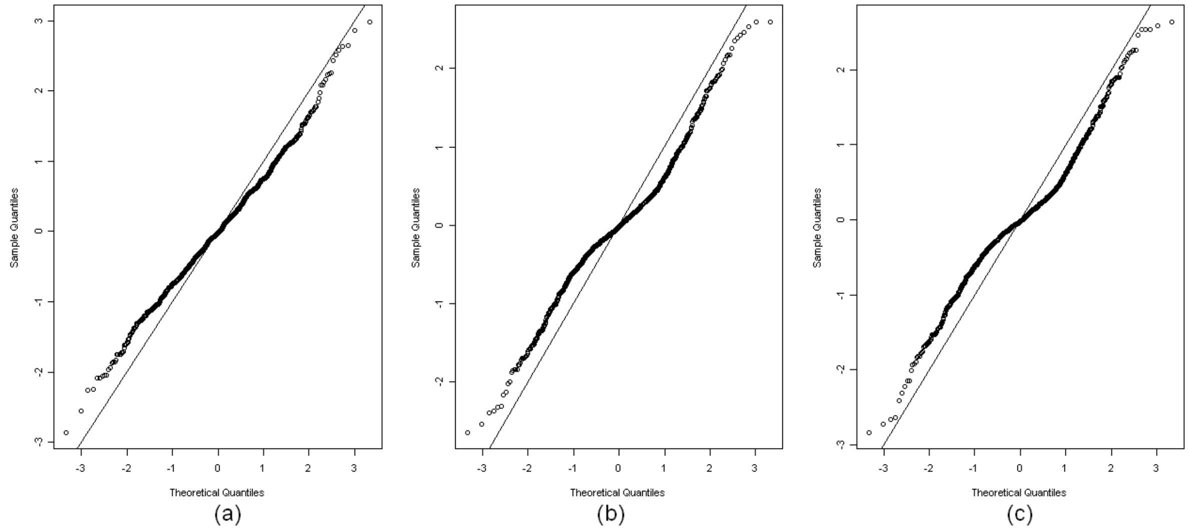


Figure 3-56 Normal QQ plots using residuals from three MCMC analyses on global data set 2. The labels (a), (b) and (c) refer to the labelled histories in Figure 3-54.

The results of this analysis are rather disappointing since the residuals fail to establish the labelled history that provides the best fit and also the most plausible grouping of populations is rejected. That the residuals were not able to distinguish between competing models is probably a reflection of the informal nature of this method of inference. It is likely the case that a more rigorous approach to model selection is needed for these data in particular, but also in general. This point will be discussed in more detail in the concluding chapter of this thesis. It is also worth noting that the fit provided by the new model under labelled histories (b) and (c) do clearly provide a better fit than the ND model (see Figure 3-28 (a)).

The analyses carried out in this section and the previous section show that residual diagnostics can be used to determine the labelled history of a group of populations for real data but also that such a method of inference has limitations. Another important point is the improvement in model fit when using the new simplified model as opposed to the ND model for both the real data sets. As previously argued, a statistical model that fits the data well can be a useful tool and the simplified model does seem to provide a better fit for both global data sets. However the improvement is not entirely clear and since the ND model is theoretically justified and is simpler to implement, it is still appealing as a statistical model.

Chapter 4

Conclusions and Discussion

This chapter includes a summary of the conclusions of this thesis, using direct reference to the aims set out in section 1.2. It also includes a discussion of the limitations, possible improvements and implications of the new model and the potential for future research.

4.1 Conclusions

The Bayesian hierarchical model proposed by Nicholson et al. (2002) provides a way of investigating population differentiation using SNP data. The initial aim was to develop an MCMC algorithm to sample from posterior distributions of parameters in the ND model using simulated and real data sets. This was accomplished using the R programming software (R Development Core Team, 2008) to implement the Metropolis-Hastings algorithm (see section 2.1.2.1). To increase the efficiency of the algorithm simplifications were made when calculating r for each group of parameters (π, α, c) and the variance of the proposal distributions were adjusted for each group to ensure that the chains moved through the parameter space in an adequate manner. Two re-parameterisations were used to simplify the implementation. The β's were introduced, whose parameter space spans the real line, to allow Normal proposal distributions without rejecting values out with [0, 1]. The truncation function $t(x)$ was also introduced to transform the β's back to the α's remembering that $t(\beta) = \alpha$. The c's were also transformed onto the log scale, again to allow the use of Normal proposal distributions, since the c's are strictly non-negative due to their relationship with the variance of the α's. A Un(0, 1) prior distribution on the π's was used throughout the analyses which represents an uninformative prior. The other potential prior was a Beta(2, 2) which reflects the tendency of SNP discovery process to find more polymorphic loci. It was decided to use the most conservative prior distribution, the Un(0, 1). The natural choice of prior distribution for the re-parameterised c's was the log-normal prior since the c's were transformed onto the log-scale.

The second aim was to assess the fit of the ND model for both simulated and real data sets. Model fit was assessed using residual diagnostic plots, and also by removing a population from the data, re-fitting the model and checking the stability of the estimates of the c's. Stability was assessed by calculating the difference between the draws from the two analyses at every step in the chain for corresponding c parameters, excluding burn-in, and computing 90% credible regions for the differences. If a credible region did not contain zero, then a significant difference was declared for that particular parameter.

As a preliminary, 100 independent data sets each including four populations were simulated and analysed under the ND model, in order to check that instability reported by Nicholson et al. (2002) for some data sets was not an unfortunate feature of the model. For each data set there were three credible regions, since removing a single population leaves three remaining populations and hence three c parameters to compare, giving 300 credible regions in total for one population removal. Two arbitrary populations were removed in turn yielding 600 credible regions, of which only one did not contain zero. This is a surprising result since one might expect some more non-zero intervals by chance alone, but it is no doubt evidence that estimates are not inherently unstable. The residuals also indicated a good fit as would be expected.

Then a set of four European populations were analysed from the HGDP-CEPH database under the ND model. The residuals from this analysis indicated a good fit, the estimates were stable under population removal and the estimates of the c's were consistent with the consensus that Europeans are the most genetically homogeneous continent. This result provides evidence that variation in SNP allele frequencies for European populations is well represented by the ND model.

It was then necessary to formulate a simulation procedure that allowed one to simulate SNP data under alternative bifurcating topologies and labelled histories, since this gave the opportunity to investigate in a simulation setting, the fit of the ND model when using data reflecting an alternative topology. The essence of the simulation procedure was the extension to the ND model used in later analyses.

100 independent data sets were simulated from four populations under a bifurcating topology (see Figure 3-13) and for each data set two separate groups of three comparisons were made: each group corresponding to a different population being removed. Again there were 600 intervals in total, but it is advantageous to summarise them in two groups of 300. When the first population was removed, 15.5% of the 300 credible regions did not contain zero; a sizeable proportion. When the second population was removed, 69% of the credible regions did not contain zero, in this case the majority. This result was particularly interesting since instability was reported by Nicholson et al. (2002) for some data sets whose evolutionary topology was likely to be different from the simple topology under the ND model. It has therefore been shown that the same instability is present when the topology is definitely incorrect and so it may be the case that the lack of robustness found in real data sets is due the incorrect topology. This result motivated the extension to the ND model defined in section 2.4, which allows alternative topologies to be specified.

Two more real data sets taken from the HGDP-CEPH database, each including populations dispersed throughout the globe, were analysed under the ND model. The first data set included Biaka pygmies, North African Mozabites, Cambodians and Native American Pima from Mexico. The estimates of the *c*'s from this analysis were rather surprising for some populations and were discussed in detail in section 3.4. The crucial result from this analysis was that the residuals suggested a lack of fit, since both the constant variance and normal distribution assumptions seem to be violated for these data. However the estimates were all stable under population removal. The second data set included two sub-Saharan African populations, Mbuti pygmies and the Mandenka, and two European populations: French and Tuscan. The estimates of the *c*'s were as would be expected for these populations, with the Europeans having small values relative to the Africans. Importantly, the residuals again showed a lack of fit but also instability in parameter estimates

The simulation analyses show that given the data are the result of a more complex topology, the diagnostics are able to highlight the discrepancy. The three analyses performed on real data highlight instances where the ND model appears to fit the data well and also when lack of fit is present. It may be the case that the data sets not well represented by the ND model may be better represented by a model with an alternative topology.

The third aim of this thesis was to develop an extension to the ND model which allows flexibility in the topology and labelled history of sampled populations. This was motivated by the poor performance of the ND model in certain situations, particularly when it was likely that the simple topology under the ND model was inadequate.

It was decided that only bifurcating topologies would be considered to limit the number of potential trees. The first model to be developed had the same probabilistic assumptions as the ND model. A set of indicator vectors were used to specify the ancestor and two offspring populations of each node, which completely defines the labelled history. This method relies on the particular labelling of the tree discussed in section 2.2. The implementation required an extra level in the hierarchy for internal nodes, corresponding to ancestral populations other than the MRCAP. The internal nodes were allowed to be fixed for a single allele and so it follows that since mutation or migration are not assumed to be present; all descendant populations at SNPs that are fixed for a single allele must also be fixed. A set of conditions were devised to ensure these properties were adhered to. When fitting this model it appeared that parameters were non-identifiable since the chains were very unstable and extremely large values were accepted. The reason for the problem is unclear and there is likely not a single cause. Some suggestions were offered in section 3.6, but it was decided that a simplification to the model was necessary. One potential simplification was to remove the dependence of the variance on the mean for the distribution of allele frequencies since this was suggested as a possible contributor to the non-identifiability. Although less well motivated by population-genetic theory, this step ameliorated the problem of identifiability and provided a model which could be fitted to data and its fit to the data assessed.

The final aim was to assess the fit of the newly developed model under different labelled histories for real and simulated data sets in an attempt to infer the most appropriate labelled history for a set of populations. The population removal diagnostic is not useful when using complex topologies since more often than not the role of a particular branch changes when a population is removed so instability would be expected, even under the correct model. The final analyses can be split up into two sections relating to the type of data used: simulated and real.

The two simulated data sets that were analysed produced differing results. For each analysis data were simulated under a particular labelled history and *c* configuration and analysed under the correct and two incorrect labelled histories. The first analysis used a *c* configuration where all c's equalled 0.1. Residuals were then calculated and compared between the three analyses. Unfortunately, the residuals were not able to distinguish the correct labelled history in this situation. Then, a more realistic data set was simulated where c = (0.05, 0.05, 0.3, 0.3, 0.1, 0.1). In the second scenario the residuals were informative about the labelled history and did in fact suggest the correct labelled history.

The global data sets analysed under the ND model were also analysed under the new model using the three potential labelled histories, given the specific topology being used. The residuals from the three analyses using global data set 1 suggested that the two African populations (Biaka and Mozabites) be grouped together, leaving the Cambodians and the Pima as the remaining group. This labelled history does seem the most obvious but the signal in the residuals was far from convincing. Another interesting comparison was between the residuals from the earlier analysis using the ND model and the analysis using the new model. Again the differences were very slight but there did seem to be an improvement in model fit when using the new model, particularly in the normal QQ plots. When assessing global data set 2 the residuals were unable to provide any information about the most likely labelled history since the residuals from two analyses were very similar. Rather disappointingly the labelled history that made the most intuitive sense (African and European groupings) produced the worst fit to the data.

4.2 Discussion

An important aspect of this thesis was to review an existing statistical model which describes variation in SNP allele frequencies. This review consisted of numerous analyses using simulated and real data to assess the applicability of the ND model in various scenarios. The recent explosion of publicly available human SNP data sets motivates a more rigorous investigation of the capability of the ND model by utilising the large volumes of genetic data now available, since an exhaustive review would have provided a much clearer perspective than is presented here.

Another important feature of the data now available is the coverage across the genome. The particular database used in this thesis was the HGDP-CEPH databank, which includes 1050 individuals typed at 650,000 SNPs distributed across the genome. Throughout the analyses on real data sets in this thesis, only 194 SNPs were used from a single chromosome; a small fraction of the available SNPs. The issue with using many SNPs is the computational efficiency needed to produce manageable running times for the MCMC simulations. The programming language used does not pertain to the type of methods involved in an MCMC analysis and so although the algorithm was efficient within the context of R, an alternative programming language such as C or Fortran would be needed if one were interested in analysing individuals at many more SNPs.

The decision not to model the ascertainment process when using SNP data was taken with the aims set out in section 1.2 in mind. However, this process is potentially important and it would be sensible to include it in any future investigations, since greater accuracy is clearly desirable.

The model proposed in section 2.4 attempted to comply with the probabilistic reasoning of the ND model but also to account for additional uncertainty in topology by using indicator vectors to define the ancestral relationships of the sampled populations. The causes of the identifiability issues encountered when fitting the model are unclear, however, removing the dependence of the variance of allele frequencies on the mean provided a model whose parameters were identifiable. This simplification was implemented with practical reasons in mind, since it allowed various labelled histories to be fitted to the data and the fit of each model assessed using residual diagnostics. Whether or not the simplified model is accurate for SNP data is debatable and it would be advantageous to seek and rectify the problems with the potentially more accurate model as a future task.

The method used to infer the most likely labelled history lacked a quantitative element present in most model selection procedures, such as Bayes factors, or likelihood ratio tests in a frequentist setting, and the sometimes ambiguous results reflected this. In any future enquiry a formal approach to model selection should be sought, which in addition to the residual approach, may provide more precise inference. A more optimistic approach would be to consider the labelled history a discrete parameter in the model and formulate an algorithm to automatically update the current tree by randomly selecting and relocating a branch at each step in the chain. The ratio of the joint conditional posterior densities for the current tree and the new tree could then be calculated and the move accepted or rejected using the standard criteria. Initial investigations into this method found that randomly choosing and relocating a branch caused problems with mixing since most moves are so unlikely relative to the current tree that they are repeatedly rejected. A branch removal algorithm which only provides moves that are not too unlikely may resolve this problem, since the posterior probability of any labelled history would be available, and could be pursued in the future.

An interesting comparison was also made between the fit of the ND model and the new simplified model using real data, again using residuals. These comparisons were not conclusive since it was not entirely clear whether the new model did provide a better fit. It must be taken into account that the distributional assumptions of the simplified model are not entirely justified by population genetics theory and although in one particular instance there was a suggestion that there was an improvement in fit when using simplified model, the ND model still performed relatively well. Again one must consider the rather subjective method used to compare the models, reflecting that an improved model selection procedure would provide a better comparison.

In conclusion, based on these analyses, it is not clear whether the new model does provide an improvement upon the ND model. But since only two real data sets were compared, using a small portion of the available SNPs, and inference was based on the informal assessment of residuals; future analyses, with the improvements that have been suggested, will undoubtedly provide more insight.

Bibliography

Atkinson, Q. D., Gray, R. D., and Drummond, A. J. (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory *Molecular Biology and Evolution* **25**, 468-474.

Balding, D., and Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and patenity. *Genetica* **96**, 3-12.

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* **53**, 370-418.

Beaumont, M. A., and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Genetics* **5**, 251-260.

Cavalli-Sforza, L. L. (1993). *The history and geography of human genes*: Princeton University Press.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., Nielson, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* **15**, 1496-1502.

Excoffier, L. (2007). Analysis of population subdivision. In *Handbook of statistical genetics*, D. J. Balding, M. Bishop, and C. Cannings (eds): Wiley.

Falconer, D. S. (1989). Changes of gene frequency. In *Introduction to quantitative genetics*, 24-50: Longman Scientific and Technical.

Felsenstein, J. (1993). Phylogeny inference package (http://evolution.genetics.washington.edu/phylip.html).

Fisher, R. A. (1930). *The genetical theory of natural selection*: Clarendon.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004a). Posterior simulation. In *Bayesian data analysis*, 289-292: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004b). Single parameter model. In *Bayesian data analysis*: Chapman and Hall.

Gillespie, J. H. (2004). *Population genetics: a concise guide*: John Hopkins University Press.

Hartl, D., L. , and Clark, A., G. (2007). Population subdivision. In *Principles of population genetics*: Sinauer Associates, Inc.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their appliations. *Biometrika*, **57**, 97-109.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* **437**, 1299-1320.

Jun, L. Z., Absher, D. M., Tang, H., *et al.* (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104.

Kimura, M. (1983). *The neutral theory of molecular evolution* Cambridge.

Lewin, B. (2004). *Genes VIII*, 8th edition: Pearson Prentice Hall.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36**, 512-517.

McColl, J. H. (2004). Multivariate probability: Arnold.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machine. *Journal of Chemical Physics*, **21**, 1087-91.

Nicholson, G., Smith, A. V., Jonsson, F., Gustafsson, O., Stefansson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society* **64**, 695-715.

Nielson, R. (2004). Population genetic analysis of ascertained SNP data. *Human Genomics* **1**, 218-224.

Olson, S. (2002). *Mapping human history*: Bloomsbury Publishing.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Roberts, G., and Rosenthal, J. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* **18**, 349-367.

Sharif, M. (2007). Statistical models of SNP variation applied to admixture analysis. MSc thesis, University of Glasgow.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **6**, 111-178.

Ye, S., Dhillon, S., Ke, X., Collins, A. R., and Day, I. N. M. (2001). An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Research* **29**, E88-8.

Appendix A

The residual formula in section 2.1.3, [13], includes a correction factor to account for the additional uncertainty in the contemporary allele frequencies, $\alpha_{ij}'s$. See below for a derivation of the variance correction factor.

Consider the two distributional assumptions of the ND model:

1. $x_{ij} \sim \mathrm{Bi}(n_{ij}, \alpha_{ij}).$

2. $\alpha_{ij} \sim \mathrm{Normal}(\pi_j, c_i\pi_j(1-\pi_j))$

Then,

$$\begin{aligned}
\frac{x_{ij}}{n_{ij}} - \hat{\pi}_j \approx \frac{x_{ij}}{n_{ij}} - \pi_j \Rightarrow E\left[\frac{x_{ij}}{n_{ij}} - \hat{\pi}_j\right] &\approx E\left[\frac{x_{ij}}{n_{ij}} - \pi_j\right] \\
&= E[\alpha_{ij} - \pi_j] \\
&= 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
Var\left[\frac{x_{ij}}{n_{ij}} - \hat{\pi}_j\right] &\approx Var\left[\frac{x_{ij}}{n_{ij}} - \pi_j\right] \\
&= E\left[Var\left(\frac{x_{ij}}{n_{ij}} \middle| \alpha_{ij}\right)\right] + Var\left[E\left(\frac{x_{ij}}{n_{ij}} \middle| \alpha_{ij}\right)\right] \quad \text{(The Law of Iterated Expectation, McColl, 2004)} \\
&= E\left[\frac{\alpha_{ij}(1-\alpha_{ij})}{n_{ij}}\right] + Var(\alpha_{ij}) \quad \text{(from (2))} \\
&= -\frac{1}{n_{ij}}\left[Var(\alpha_{ij}) + [E(\alpha_{ij})]^2 - E(\alpha_{ij})\right] + Var(\alpha_{ij}) \quad \text{(since } Var(x) = E(x^2) - [E(x)]^2\text{)} \\
&= -\frac{1}{n_{ij}}\left[c_i\pi_j(1-\pi_j) + \pi_j^2 - \pi_j\right] + c_i\pi_j(1-\pi_j) \quad \text{(from (1))} \\
&= \pi_j(1-\pi_j)\left[c_i + \frac{1}{n_{ij}}(1-c_i)\right].
\end{aligned}$$