The Use of Bayes Factors in Fine-Scale Genetic Association Studies

Robin Young

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

Department of Statistics

February 2011

Abstract

The aim of this thesis is to explore and compare methods that can be used for the purposes of finding possible genetic effects in the context of fine-scale genotype-phenotype association studies. Fine-scale genetic association studies present unique challenges for attempts at finding genetic effects, due to the strong linkage that can exist between different variants and issues that exist as a result of multiple testing. However, unlike Genome-Wide Association Studies (GWAS), there is potential to use the information from haplotypes arising from areas of low genetic recombination.

In order to test the effectiveness of approaches involved in fine-scale studies, the PheGe-Sim (Phenotype Genotype Simulation) application has been developed in order to simulate fine-scale phenotype-genotype data sets under a variety of scenarios. The simulations are based upon the coalescent model with extensions of population expansion, recombination, and finite sites mutations, that allow for real data sets to be more closely mirrored. The simulated data sets are subsequently used to assess the effectiveness of each of the methods that are used in this thesis, in attempting to find the known simulated causal variants.

One of the methods suitable for use in fine-scale genetic association studies for testing associations is Treescan (Templeton et al., 2005). Treescan is a method that attempts to use relationships between closely related haplotypes in an attempt to increase the power of finding genetic determinants of a phenotype. A haplotype tree is constructed, and each branch can be sequentially tested for any evidence of association from the resultant groups. To provide comparisons with the Treescan method, similar methods to the Treescan approach using each SNP (single nucleotide polymorphism) and haplotype have been implemented.

As a result of the issues of multiple testing in the context of GWAS, Balding (2006) advocated the use of Bayes factors as an alternative to the standard use of p-values for categorical data sets. In this thesis Bayes factors have been formulated that are suitable for continuous phenotype data, and for the context of fine-scale association studies. Bayes factors are used in a method that utilizes the Treescan approach of assessing various groupings from a haplotype tree, with the method being adapted to take advantage of the flexibility offered by Bayes factors. Single SNP and haplotype approaches have also been programmed using the same implementation of Bayes factors.

The PheGe-Find (Phenotype Genotype-Find) application has been developed that implements the association methods when supplied with the appropriate genotype and phenotype input files. In addition to testing the methods on simulated data, the approaches are also tested on two real data sets. The first of these concerns genotypes and phenotypes of the *Drosophila Melanogaster* fruit fly, that has previously been assessed using the original Treescan approach of Templeton et al. (2005). This allows for comparisons to be made between the different approaches upon a data set where there is strong evidence of a causal link between the genotype and phenotypes concerned. A second data set of genetic variants surrounding the human ADRA1A gene is also assessed for any potential causative genetic effects on blood pressure and heart rate phenotype measurements.

Acknowledgements

Firstly I would like to thank my supervisor, Dr. Vincent Macaulay. His suggestions and comments have proved invaluable to the work involved in this thesis and, although often unsaid, has been greatly appreciated.

I would also like to thank Professor Alan Templeton, for providing the original *Drosophila* data set, and for his efforts at obtaining permission to use other data sets which he had been involved with. I would also like to acknowledge the assistance of Dr Sandosh Padmanabhan for providing the ADRA1A data from the PAMELA study. Both Paul Johnson and Mark Bailey have also provided useful assistance at various stages throughout my studies.

I would like to acknowledge the funding support of EPSRC which allowed me to undertake this research. In addition, I would like to thank the Department of Statistics at the University of Glasgow for support in attending conferences. I would also like to thank all the academic and administration staff in the Department of Statistics, for assistance in numerous ways over the last four years.

Finally, I would like to thank all my friends, family, and associates for allowing me to make the most of my time in the last few years. Whether it be playing hockey, badminton or squash: going for lunch, tea, a drink, quiet or otherwise, or just generally being there to listen to me complain about one thing or another. It has been, and always will be, truly appreciated.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Fine-scale genetic association studies are an important component in finding genetic variants that increase an individual's risk of common complex diseases, such as type 2 diabetes and Alzheimer's disease. Once a genome-wide association study (GWAS) has identified an area of the genome that is likely to contain a causative genetic component, fine-scale mapping studies are required in order to locate the specific mutation that is responsible for the increased risk, or the cause of a change in some measurable component such as blood pressure or heart rate. If a genetic variant is found that is associated with a condition, then this could provide valuable information as to what biological features are involved in its aetiology. This information could then be used in order to direct efforts at attempting to develop treatments that could target the biological and genetic component of common complex diseases (WTCCC, 2007).

1.1 Mendelian and Complex Diseases

The association of inherited genetic variants with an observable phenotypic trait has been a key component in understanding the processes involved in genetics since the late 19th century. As the understanding behind the processes involved in inheritable diseases improved, the genetic variants that caused diseases were

discovered. Initial experiments on fruit flies by Thomas Hunt Morgan led to the discovery of the primary patterns of inheritance; namely dominant, recessive and additive conditions of such traits as eye colour and wing length (Morgan et al., 1915). Mendelian traits, those that are entirely controlled by the action of a single gene, are now generally well characterized in many species through use of methods such as linkage analysis within families of disease carriers. In humans the genetic variants responsible for heritable Mendelian diseases have been determined for several severe conditions, such as cystic fibrosis (Rommens et al., 1989; Kerem et al., 1989) and Huntington's Chorea (Huntington's Disease Collaborative Research Group, 1993).

The successful discovery of disease-causing genes in Mendelian disorders has led to efforts to determine genetic factors involved in more complex diseases, which are a result of both multiple genetic and environmental components and hence have a more complicated inheritance pattern than Mendelian diseases. The common disease-common variant (CDCV) hypothesis (Lander, 1996; Pritchard and Cox, 2002) proposes that there will be multiple common genetic variants that each confer a small increase in risk for common diseases, such as type 2 diabetes or Alzheimer's disease. Strong influences of environmental factors, and the late onset of such conditions resulting in limited selection pressures, can make efforts at finding the relatively small effects of contributing genetic factors difficult. This can be illustrated in the case of type 2 diabetes where there are well-characterized genetic effects (Sladek et al., 2007), however, there are also strong associations between an individual's diet and their risk of contracting the condition. There are further difficulties in detecting genetic associations with complex diseases due to the possibility of multiple variants in the genome interacting with each other to contribute to a change in the risk of contracting the disease.

1.2 Genome-wide Association Studies

Genome-wide association studies for complex diseases have become increasingly feasible, and as a result more popular, due to the accumulation of evidence about

genetic variation that has been documented in the HapMap Database (International HapMap Consortium, 2005) and the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010). A further practical consideration that has facilitated more widespread use of GWAS is the substantial advancements that have been made in genotyping technology. In particular, the development of DNA microarrays that can accurately type single nucleotide polymorphisms (SNPs) at previously discovered locations in the genome, has enabled the genotyping of hundreds of thousands of genetic variants quickly, cheaply and with relative ease compared to previously used technology.

The potential usage of one such microarray, the Affymetrix 500k, was illustrated in the influential genome-wide association study paper from the Wellcome Trust Case Control Consortium (WTCCC, 2007). In this study genotypes of individuals with one of seven complex diseases were compared to a common control group, to detect any associations between the genotypes and the disease under consideration. The study reports successful discovery of highly significant associations with six of the conditions, with significance for genome-wide association being defined as a p-value of less than 5×10^{-7}. As noted by the WTCCC, the associations that have been found require validation in further studies before there can be confidence in their association with the diseases. GWAS by Barrett et al. (2009) for associations with type 1 diabetes and by Trégouët,D.A and others (2009) for coronary artery disease, represent just two of the many such studies that have attempted to follow up or replicate the results obtained by the WTCCC.

For genome-wide association studies, it is unlikely that a true causative mutation will be typed in a study, due to there being many more SNPs present in the genome that will not have been typed. It is though hoped that a SNP will be found that is highly correlated to a SNP that is indeed causative, due to linkage disequilibrium between closely-spaced nucleotides. Fine-scale association studies are then used in an attempt to validate signals of association detected in a GWAS, which can be achieved by typing additional SNPs in identified regions of interest. Selected recent examples of the many such validation studies are Rung et al. (2009) and Lowe et al. (2007), assessing variants for association with type 1 and type 2 diabetes respectively.

Methods involved in fine-scale studies can be different to those used in GWAS, for both biological and practical reasons. Unlike GWAS, in fine-scale studies there is information that can be used advantageously about the haplotypes and the relationships in their inheritance over short genetic distances. In practical terms, simply due to there being fewer SNPs being considered, it is more feasible to consider gene-gene and gene-environment interactions in a fine-scale study. There is, however, the additional consideration of strong linkage existing between densely typed SNPs, that will generally not be the case for SNPs in a GWAS. This information can be useful in determining potential causative associations, although this can also lead to difficulties in identifying which of the linked variants are in fact causative for a phenotype. A further practical consideration with fine-scale studies, is that it is time consuming and expensive to obtain customized genotyping at a fine-scale level, whereas a GWAS will tend to use a DNA microarray with pre-determined SNPs to be typed. Irrespective of the approach used, if a SNP is found to be causatively associated with a condition, then this can potentially be used in efforts to develop more effective treatments for the condition.

The methods explored in this thesis are aimed at short gene segments in strong linkage disequilibrium, which have been covered with a dense panel of SNPs so as to capture a high percentage of the potential variants within a region being considered. This data can arise through resequencing studies of candidate genes identified in a GWAS, and will be referred to as fine-mapping approach. As commented previously, a GWAS will often have been designed so as to genotype at SNPs that can be used to capture the data represented at the variant itself, and nearby SNPs that are correlated with it. Imputation of the untyped SNPs using HapMap or 1000 Genomes data sets can then be used to obtain the likely alleles at the untyped variants, using programs such as IMPUTE (Howie et al., 2009), MACH (Li et al., 2010) or BIMBAM (Servin and Stephens, 2007). Although these approaches can be accurate at imputing untyped variants, there will be less accuracy at rare variants and SNPs that are separated by recombination from those that have been typed. As the genotyping technology improves, it is plausible that almost all genetic variation can be captured using genotyping chips and the need for imputation will diminish. However, in such situations of dense SNP coverage, multiple linked variants can all be apparently associated with a

phenotype being considered. Although the variant with the strongest signal is the most plausible candidate for true association, it is relevant to refine this further into determining if other nearby low frequency variants are in fact worthwhile candidates, and if there is the potential effect of one or more distinct signals that can be masked by considering each SNP independently. In such situations, there may be advantages of approaches that can test multiple variants together, or methods that can take advantage of the specific linkage patterns within a region of low recombination that can potentially be modelled using the ancestral relationships between individuals.

The methods that can be used for fine-scale studies are discussed further in section 1.5. However, first the theory of the coalescent is introduced, as this can be used to simulate genetic data upon which the fine-scale methods can be compared.

1.3 The Coalescent

In order to describe the genetic ancestry of individuals sampled from the present day the coalescent theory has been developed. Originally suggested in a series of papers by Kingman (1982a,b), the theory is a development of ideas about neutral evolution in population genetics attributable to Wright (1931) and Fisher (1930) in the Wright-Fisher model. A brief summary of the features of the coalescent model follows. However, further details of the Wright-Fisher and coalescent models can be found in Hein et al. (2005) and Wakeley (2008).

The coalescent aims to describe relationships between subjects' haplotypes that exist at the present time, by hypothesizing that at some time in history there is an ancestor common for a pair of individuals' haplotypes. This concept is extended for all individuals' haplotypes, and subsequently for groups of individuals, until a single most recent common ancestor to the entire sample of haplotypes is obtained. Formulation of the coalescent theory provides statistical distributions and properties of genetic data, that can subsequently be used for simulating realistic data sets.

Coalescent theory provides a mathematical framework for describing the genetic ancestry of a set of individuals and relies upon a set of fundamental assumptions (Hein et al., 2005). The first of these assumptions is that individuals are haploid in nature. Although not literally true for humans, the pairs of N individuals can be considered as $2N$ chromosomes or $2N$ haploid individuals to a very good approximation.

A second assumption made in the coalescent process is that of discrete and non-overlapping generations. This assumption is highly unrealistic, as for example it is never going to be the case that all individuals in a human generation are born and die at exactly the same time with a lifespan of approximately 25 years per generation. The coalescent is however a method of generalizing to a large population over a considerable period of time, and it turns out that this assumption is of minor relevance when these two factors are taken into consideration.

Arguably the most important of the assumptions of the coalescent process, and one that can provoke much discussion, is that mutations that occur are selectively neutral. This assumption is a key feature in the construction of methods to simulate the coalescent, as it allows the coalescent structure and the mutation pattern to effectively be treated as two independent processes. A justification for this neutrality assumption is that most inheritable complex diseases, such as Type 2 diabetes, are late onset and as such causative mutations are highly unlikely to have strong selection pressures acting upon them.

The initial formulation of coalescent theory required the assumption of a constant and finite population size, however subsequent developments in theory particularly by Donnelly and Tavaré (1995) have relaxed this assumption. Complex arrangements of population expansion and sudden bottlenecks in population size can be formulated, depending on the population ancestry that is required to be modelled. The PheGe-Sim program that has been developed in chapter 2 incorporates the possibilities of the simplest scenarios of a constant population size, or of a population that is exponentially increasing in size forward in time. More complex models of population change over time can be accomplished using other coalescent simulators. However, exploring the specifics of demographic history is not of primary importance in this thesis.

It is also assumed in the coalescent that no geographical or social structures

exist in the population, i.e the population is randomly mating. As with the assumption of discrete and non-overlapping generations, this is a highly unrealistic assumption to make for humans. There are multiple ways in which this assumption can be violated in practice, based upon common-sense arguments such as in human populations there are inherent social and geographical restrictions that have shaped the development and isolation of different populations and ethnic groups over history. There are, however, methods of attempting to reconstruct populations (demes) with migration activity between them, such as the finite islands model. Features of migration models have not been incorporated into the PheGe-Sim program, as methods of association will have difficulty in adequately detecting and modelling these features. As such, even large scale association studies tend to be based upon a single ethnic group or population, in attempts to reduce the risk of inflated type-I error rates that can arise when population structure is not taken into consideration.

A further assumption of coalescent theory, that has subsequently been relaxed, is that the region of DNA under consideration is not affected by any recombination events. Hudson (1983) developed methods to include recombination in the coalescent history, resulting in Ancestral Recombination Graphs (ARG) (Griffiths and Marjoram, 1996). Methods such as this have been implemented in ms (Hudson, 2002) and msHot (Hellenthal and Stephens, 2007), that simulate genetic samples with various scenarios of recombination and recombination hostspots. The PheGe-Sim program of Chapter 2 also allows for the inclusion of recombination events, however, it does not include extensions to variable recombination rates and recombination hotspots. Although accounting for recombination would be potentially useful, it will be seen in Chapter 7 that realistic fine-scale genotype data sets can be simulated without this feature.

1.4 Coalescent Simulation

Numerous simulators have been programmed that make use of the coalescent theory in order to create samples of DNA sequences from a population, such as: ms (Hudson, 2002), msHot (Hellenthal and Stephens, 2007) , CoaSim (Mailund et al., 2005), SimCoal (Excoffier et al., 2000), Genome (Liang et al., 2007) and

Table 1.1: Features of selected coalescent simulators; Rec = Recombination; FS = Finite Sites; PE = Population Expansion ; CDH = Complex Demographic Histories; CPS = Continuous Phenotype Simulation.

Method	Short Description	Rec	FS	PE	CDH	CPS	Comments & Criticisms
ms	ms is one of the first, and most commonly used, methods for simulating genetic data using a coalescent model.	✓	✗	✓	✓	✗	Modification would be required to allow for finite sites and to incorporate phenotype simulation.
msHot	msHot is an extension of ms that incorporates recombination hotspots, to separate blocks of low recombination.	✓	✗	✓	✓	✗	As with ms, modification would be required to allow for finite sites and to incorporate phenotype simulation.
Genome	Genome is a program that intends to efficiently simulate large sections of DNA.	✓	✗	✓	✓	✗	Primarily intended for genome-wide simulation, it can miss some of the details required for finer scale simulation.
CoaSim	CoaSim is a less well-known program, but provides a wide range of options for simulations.	✓	✓	✓	✓	✗	Simulates case-control and not continuous data, also relies on a penetrance approach as opposed to one based upon the linkage of SNPs in assigning phenotypes.
SimCoal	SimCoal is a coalescent simulator typically used for exploring features of complex demographic history.	✗	✓	✓	✓	✗	Does not allow for recombination or phenotype simulation, and is based on a discrete as opposed to continuous coalescent model.
PheGe-Sim	PheGe-Sim intends to simulate data that can be directly used for modelling fine-scale phenotype genotype association studies with continuous data.	✓	✓	✓	✗	✓	PheGe-Sim is slower than some of the other methods, and does not include options for migration or specified demographic events.

PheGe-Sim (Chapter 2). Table 1.1 details some of the features of these widely used programs for generating haplotype samples.

Alternative approaches of simulating genetic data are also possible, other than the backwards-in-time coalescent methods presented in table 1.1. Forward-in-time simulators such as Fregene (Hoggart et al., 2007; Chadeau-Hyam et al., 2008) are potentially more flexible but are substantially more computationally intensive for large sample sizes. Approaches of simulating data based upon permuting HapMap data are also used in certain situations (Spencer et al., 2009). The simulation of data in PheGe-Sim uses a backwards-in-time coalescent approach, although is not completely standard in its approach. This is because of resampling haplotypes at an intermediate stage, a procedure that is more computationally efficient. PheGe-Sim also integrates the mechanism of causative mutations and the history of the sample at the same time, allowing for specification of causative effects that are not independent of the history of the sample. This can subsequently be used to inform an effect size based upon multiple mutations and interactions that takes into account the history for finite-sites models of mutations.

The programs of table 1.1 could in theory be adapted to be suitable for simulations of fine-scale genotypes and continuous phenotypes, however, they would also have to be adapted to be suitable for integration with the methods used for association detailed in Chapter 4. The Phege-Sim application of Chapter 2 has therefore been coded in R (R Development Core Team, 2006), specifically to simulate fine-scale genotype-phenotype data and to be integrated with the other applications used for genotype-phenotype association studies outlined in this thesis.

In order to be suitable for generating fine-scale genetic data, the resulting data set must display certain distinct patterns of linkage between the SNPs. In GWAS, using a DNA microarray will provide a map of SNPs covering the whole genome. However, this will not usually provide the dense coverage as required for fine-scale analysis. Figure 1.1 illustrates the linkage pattern displayed within a real data set (Chapter 6), and this can be seen to have a different range and spread of colouring than can be achieved using any of the simulators that are primarily intended for genome-wide simulation (appendix C.1). The specific details of the

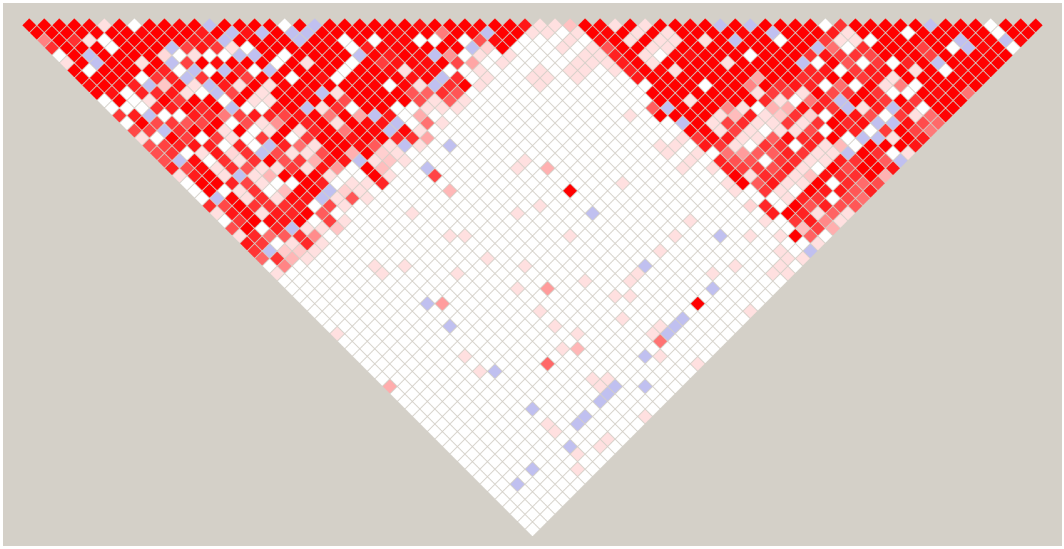
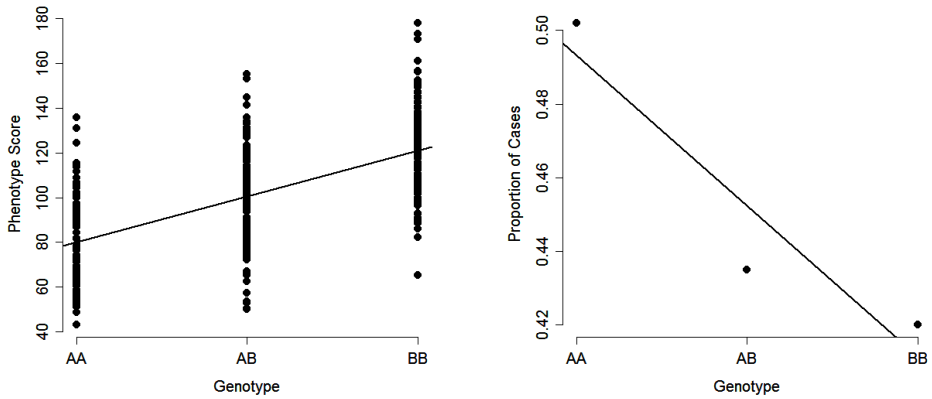


Figure 1.1: Example linkage plot from the data in Chapter 6.Each SNP involved is represented by a separate row of the plot, and therefore each square represents the linkage (correlation) between two SNPs. Increasing linkage is indicated by darker reds, and areas of low linkage or low sample size are indicated by white or blue squares respectively.

necessary parameters and features of PheGe-Sim that are required for the fine-scale genotype simulation are discussed further in Chapters 2 and 7.

1.5 Fine-scale Association Study Methods

In both frequentist and Bayesian approaches to association testing, the appropriate choice of method to use depends upon whether the outcome of interest is binary or continuous in nature. If the trait of interest is continuous, the simplest test of association in an attempt to detect if there are differences in the phenotype between the genotype groupings is to use a straight-forward analysis of variance (ANOVA) model. This approach is comparable to the proposed flexible general model of the Bayesian analysis (section 3.3), whereby each genotype class can have a different mean, but there is a common variance within each genotype class. If the data is categorical, as opposed to continuous, logistic regression or a standard χ^2 test can be used to test for differences in proportions of cases and controls, between the three genotype classes. The method suggested by Balding



(a) Additive Regression (b) Armitage Trend test

Figure 1.2: Illustrations of additive models for both continuous data (a), and the binary method of the Armitage Trend Test (b).

(2006) provides a Bayesian option for categorical analysis, and a modified version of this is detailed in section 3.1.

If it is assumed that the effects of alleles at a SNP are to act in an additive manner on a continuous outcome of interest, then a standard regression can be used as shown in equation 1.1 and figure 1.2(a):

$$y = \alpha + \beta x + \varepsilon, \tag{1.1}$$

where y is the phenotype response, α is the intercept of the regression equation, β is the slope parameter of the regression equation, x is the genotype class (0 for AA, 1 for AB and 2 for BB) and ε is the residual error that is Normally distributed with mean zero and variance σ^2.

The Armitage trend test (Armitage, 1955) is a comparable method for detecting additivity in binary data, where the outcome of interest is the proportion of cases in each of the three genotype classes. The test can be implemented through use of the R function *prop.trend.test*. An illustration of the approach is given in figure 1.2(b), and the corresponding χ^2 test statistic is calculated by equation 1.2:

$$\chi^2_G = \frac{N\left[N(r_1 + 2r_2) - R(n_1 + 2n_2)\right]^2}{(N-R)R\left[N(n_1 + 4n_2) - (n_1 + 2n_2)^2\right]} \sim \chi^2_1, \tag{1.2}$$

where, N = Total sum of cases and controls, R = number of cases, r_1 = count of the case heterozygote alleles, r_2 = count of the case homozygote alleles, n_1 = number of AB cases and controls, n_2 = number of BB cases and controls; with $\chi^2_G \sim \chi^2_1$ if there is no trend.

The additive model can be unsatisfactory in that, for a single SNP test, there will only be three possible groupings of genotype and therefore only three possible points from which the coefficients of a regression line are to be calculated. A further possible issue with this approach is that, used in isolation, tests for additivity can have low power for detecting non-additive differences between the genotypes such as may occur for dominant models, or models with three independent genotype means. In an extreme (albeit unlikely) situation, a regression approach for detecting additive effects would have no power to detect a true association if the heterozygote genotype grouping (AB) had a substantially different mean phenotype measurement compared to two similar homozygote phenotype means. Both general and additive tests should also be carried out to detect any differences between the genotypes, although care must be taken to ensure that issues of multiple testing involving non-independent tests are appropriately considered. Similarly, consideration must be given to the multiple testing issues in the calculation of p-values for dominant or recessive models, which are essentially the same as the calculation of p-values for the general model but constrained to only two genotype classes. Additive (section 3.4) and dominant/recessive (section 3.5) models can be fitted in the Bayesian setting for continuous data, along with Bayes factors for combinations of two or more SNPs (sections 3.6, 3.7 and 3.8). All possible one and two SNP models can be compared in the Bayesian framework, and although still to some extent subjective, there is explicit acknowledgement of the role of prior beliefs about the proposed models that will often nonetheless be inherently present in the frequentist approach. In addition, if desired, a Bayes factor can also be calculated by producing a weighted average of different models, with the weights being specified according to prior beliefs about the different models.

1.5.1 Haplotypes

In candidate-gene studies, or fine scale studies within several linked genes, there is potential to use haplotypes as opposed to SNPs in efforts to find causative mutations, as haplotypes are essentially the units of inheritance passed from one generation to the next. There are various advantages and disadvantages to a haplotype-based approach compared to assessing the affect of each SNP individually, an area which is explored in detail by both Clark (2004) and Schaid (2004). An immediate problem is that haplotypes are generally not observed, and have to be inferred to the correct alignment (phase) using the available genotype data with a program such as PHASE (Stephens et al., 2001) or fastPHASE (Scheet and Stephens, 2006) which implement a Bayesian approach to phase estimation. Although these programs are accurate in areas of low recombination rate and high SNP density (Scheet and Stephens, 2006; Marchini et al., 2006), phasing will inevitably introduce an additional source of error that would not be present in single SNP analysis. In strongly recombinant areas of the genome, there will be errors in phasing haplotypes due to uncertainty in determining the correct ordering of sequences, although in these situations it can be easier to identify potential individual causative variants. Figure 1.1 shows an example where there is a clear recombination hotspot that separates the SNPs into two haplotype blocks, and if this structure is ignored, then this can lead to a far greater number of haplotypes than if the phase is inferred for each block separately. Further details about this particular data set are given in Chapter 6.

It is, however, thought that there may be some advantages in using haplotypes as opposed to assessing each SNP individually. This is due to the SNPs in a block of DNA being correlated with each other, and so there may be interactions between SNPs that are best captured when considered as one single haplotype. The linkage between SNPs will also invalidate some of the correction methods for multiple testing in single SNP analysis, as the assumption of independent tests will not be valid. Methods that test the strength of association between each of the observable haplotypes and a phenotype of interest have been developed, to act in an analogous way to the single SNP methods that have been used. Such methods will however lack any ability in determining which of the SNPs on a

haplotype is crucial in determining a change in phenotypic score. The use of haplotype-based methods is therefore potentially only useful when there are multiple causative SNPs upon a single haplotype, each of which could be individually of small effect and thus missed in a single SNP-based analysis. Haplotype-based analysis can subsequently be modified to take into account a proposed tree structure relating the haplotypes, and this method could be better placed to identify groups of related haplotypes that are all associated with a change in phenotype measurement. This method will also have the added advantage, in that the SNPs that differentiate associated groups of haplotypes can be ascertained.

1.5.2 Phylogenetic methods and Treescanning

The structure of linkage present in small regions of chromosomes can be used to an advantage in detecting causative associations through the use of phylogenetic models. There can also be advantages in the interpretation of any found associations due to haplotypes forming a basic unit of inheritance (*Drosophila* 12 Genomes Consortium , 2007), although obtaining haplotypes for human populations can present difficulties (Durrant and Morris, 2005). Single SNP tests that ignore the correlations in alleles at different SNPs caused by shared ancestry would be expected to be less powerful than methods that explicitly model those correlations, for example by inferring aspects of the ancestral tree or graph that relates the haplotypes. Such approaches potentially allow signals coming from multiple correlated SNPs, occurring close to each other in the ancestry, to bolster one another.

Taking account of the ancestral relationship of haplotypes can increase the power of association studies (Eskin, 2008; Roeder et al., 2006), and numerous methods have been proposed that aim to take advantage of the linkage. One such method is that proposed by Durrant et al. (2004) in which a sliding window approach is taken to define haplotype blocks that are then related via a cladogram, which is subsequently used to create tests for association with a phenotype. Minichiello and Durbin (2006) take a related approach that estimates Ancestral Recombination Graphs (ARGs) as opposed to cladograms to relate haplotypes. Zöllner and Pritchard (2005) proposed a method that relies on an underlying

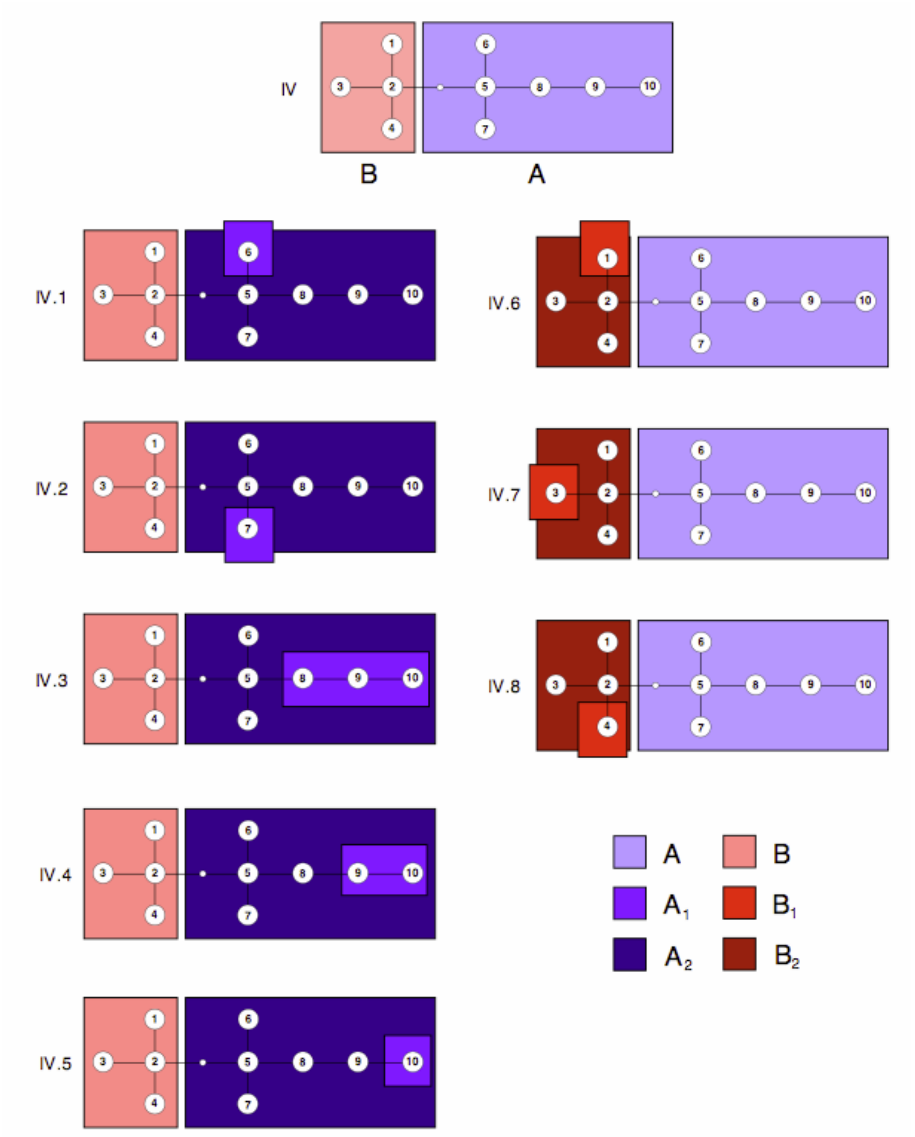


Figure 1.3: Example of the use of Tree-Scanning; reproduced from the Treescan Documentation (Templeton et al., 2005). Part IV shows two groups of haplotypes, represented by 'A' and 'B', that have been determined to have significantly different phenotype means. In a second round of tests, illustrated by parts IV.1 to IV.4, further groups of haplotypes are considered in an effort to determine if there are additional significantly different phenotype groups, conditional on retaining the differences found between groups 'A' and 'B'.

coalescent approach to relate individuals, in an attempt to increase the ability to find causal variants by observing non-random clustering of haplotypes. This approach can though be computationally expensive, a feature that Kimmel et al. (2008) aim to overcome using their method that relies on no recombination events and a finite sites model across each region being considered.

One of the phylogenetic methods that is assessed in this thesis through the use of simulated data, is the Treescan method (Templeton et al., 2005). This method is a development of the nested clade analysis method, also developed by Templeton and colleagues in a series of articles (Templeton et al., 1987, 1988, 1992). The nested clade analysis method involves deriving a network of related haplotypes, and then sequentially assessing for various groupings of haplotypes as to whether there is any evidence of a difference in phenotype scores between the clades (groups) using F-statistics. Treescan uses a similar method to the nested clade analysis, however a key difference is that Treescan allows for the use of diploid as opposed to haploid populations and is thus suitable for use on human populations.

In situations of no recombination and infinite sites mutation, the Treescan method will produce identical results to the single SNP analysis. There will however be an advantage of the Treescan approach in circumstances where a causative mutation may only result in a change in phenotype score on a particular lineage, and thus only on one haplotype carrying the causative form of the mutation. In this situation grouping together individuals with the same DNA base as in a single SNP analysis would lead to a reduction of power, as there would be no separation of causative and non-causative mutations that can be achieved in the Treescan analysis.

The first stage of a Treescan analysis is the construction of a haplotype tree, by pruning a network of haplotypes if there is ambiguity in the branching structure, the specific details of which are described in section 4.2. The tree is then partitioned sequentially at every branch, and an ANOVA carried out to compare the phenotypes of the three resulting groups; namely AA, BB or AB from the top image in figure 1.3. The p-values that result from the ANOVA must be corrected in some way, to compensate for multiple testing and for the correlation between

the different tests. Treescan obtains corrected p-values using a permutation approach, whereby the phenotype scores of the sample are randomly allocated to each of the three groups, whilst preserving the initial number of samples in each. A p-value is then obtained by taking the proportion of the F-statistics from the ANOVA's for each of the permuted data sets, that exceeds the original F-statistic relating to the grouping of clades. Monotonicity is then enforced through successive maximization of the p-values, to ensure that a higher F-statistic results in a lower permuted p-value.

Due to the correlation that exists between closely spaced SNPs on a haplotype, the Treescan method may detect branches with a low p-value due to a 'spill-over' effect from a branch that is truly associated with the phenotype. To compensate for this effect, Treescan performs a second stage of splitting the tree, conditional on a branch that is found significant in the first round of tests. Figure 1.3 illustrates the second round of partitions that are assessed, after initially finding the groupings A and B as significantly different, as adjudged by having a corrected p-value of less than 0.05. All splits that satisfy this criteria from the first round are then fixed, and a second stage of splitting conditional to this is implemented. A decision can then be formed as to whether a haplotype is strongly associated with the phenotype, by assessing the tables of first and second stage results (see section 4.4).

The methods discussed so far are all in the frequentist setting and rely on the use of p-values, which are corrected for instances of multiple testing. Although there are valid criticisms that can be made regarding the frequentist approach and its use of multiple-testing corrections, it is still useful for the context of association studies. A SNP that has been found to be strongly associated with a phenotype of interest by the use of p-values will also have a strong chance of being found through the use of a properly calibrated Bayesian analysis. The use of Bayes factors is though particularly appropriate for data in a genetic association context, for the following key reasons:

1. In fine-scale association studies, there is a large number of hypothesis to be tested as there are at least as many tests to be performed as there are SNPs that have been typed in a study. This leads to issues in appropriately

correcting for multiple testing, a situation that is even more crucial for GWAS, where the number of tests can run into the millions.

2. It is possible to end up with rare groupings of genotypes that have large p-values which are in fact a result of chance, due to there being low power to detect such differences. Appropriate reporting and analysis of such SNPs should take this into account. However, this situation is automatically accounted for in the Bayes factors through the use of the priors, but can often be neglected when assessing p-values. There are no known methods to correcting reliably for this issue in the frequentist setting, and any sensible approach will necessarily be introducing some form of subjective opinion as to what sample size and power are suitable criteria to determine significance.

3. Multiple different models, such as additivity, dominance or recessive effects, can be applied to the same SNP or combinations of SNPs. The weight of evidence for all the possible models should be calculated and compared, to ensure full exploration of a data set, as is the situation in the Bayes factors context. In a frequentist approach, consideration should be given to multiple testing issues for the non-independent tests of the different models.

4. In fine-scale studies in particular, SNPs may be highly correlated, and there are no suitable methods of correcting p-value cut-offs for such data. The linkage that occurs in data will also be a feature of the Bayesian analysis, however as noted in the previous comment, the Bayesian approach can fully explore all combinations of models and SNPs that are deemed appropriate without fear of multiple testing issues, which are replaced by a priori models of effect sizes and the number of causative SNPs.

5. There are no steadfast rules about what can be considered as 'significance' in reporting of a found association, and a criticism of the Bayes factor method in that there is subjective use of priors can equally be applied to the interpretation and reporting of results from p-values. The strength of an association of a SNP with a trait will depend on many factors, and the weight of evidence that is found for each possible model should be reported. This is possible in a Bayesian approach, where the evidence can be presented

in such a way that can convey the uncertainty in whether true associations are present.

The following sections detail the approaches that could be applicable in the calculation of Bayes factors, in addition to introducing the possibility of the exactly computed Bayes factors that are used in the PheGe-Find application of Chapter 4.

1.6 Bayesian Methods

The theory of Bayes factors was first developed by Jeffreys (1939), as a way of quantifying the evidence in favour of competing models or hypotheses. The methodology behind Bayes factors was re-introduced with the review paper of Kass and Raftery (1995).

The Bayes factor for two competing models H_0 and H_1, are part of the following relationship:

$$Posterior\ Odds = Prior\ Odds \times Bayes\ Factor. \tag{1.3}$$

That is,

$$\frac{p(H_1|y)}{p(H_0|y)} = \frac{p(H_1)}{p(H_0)} \times \frac{p(y|H_1)}{p(y|H_0)}, \tag{1.4}$$

where y symbolizes the data, and the calculation of the marginal likelihoods $p(y|H_1)$ and $p(y|H_0)$ that form the Bayes factor, requires calculation of the following integrals:

$$p(y|H_0) = \int p(\theta_0|H_0)p(y|\theta_0, H_0)d\theta_0, \tag{1.5}$$

and

$$p(y|H_1) = \int p(\theta_1|H_1)p(y|\theta_1, H_1)d\theta_1, \tag{1.6}$$

where θ_i are the parameters of model H_i, $i = 0, 1$.

The primary consideration in the use of Bayes factors is in evaluating the integrals of equations 1.5 and 1.6. The posterior distributions for a Bayesian analysis should be proper, i.e. integrate to 1, thus also ensuring that appropriate comparisons between competing hypothesis can be made. To satisfy this criteria

careful consideration must be made when deciding on the prior distributions that are chosen, particularly if using non-conjugate priors.

The use of Bayes factors as a method of ranking associations has been increasing in popularity in the context of genetic studies. However, the relationship between p-values and Bayes factors is not straightforward as the interpretation depends strongly on the sample size and minor allele frequency of each test being considered (Wakefield, 2009). Methods providing a compromise on the frequentist and Bayesian approaches have also been proposed such as treating Bayes factors as a test statistic and obtaining p-values by permutation, or by calculation of a posterior probability of association (Stephens and Balding, 2009; Servin and Stephens, 2007). The Wellcome Trust Case Control Consortium (2007) used Bayes factors for *case-control* studies with conjugate priors developed in the SNPtest program (Marchini et al., 2007), and the Bayes factors used have been useful in dealing with the uncertainty caused by imputed genotypes. As GWAS studies of *continuous* phenotypes are increasingly being tested, the Bayes factors have to be adapted to be suitable for this context, resulting in increased difficulty in the choice and specification of suitable priors.

In some instances the integral can be calculated analytically, if suitable prior distributions are chosen. The advantages of this approach is accuracy in the answer, and in reduced computation time in comparison to numerical methods of evaluating the integral. In order to use the analytic approach, conjugate priors provide the most straightforward approach, to ensure that the posterior is of the same form as the prior. In the examples in Chapter 3, a conjugate Normal Inverse Chi-squared distribution (Gelman et al., 2004) is used for the prior for the mean and variance.

If the integrals cannot be performed analytically, there are various alternative methods that can be used. Laplace' method can be employed in order to estimate Bayes factors. An un-normalized posterior density is obtained, and then this distribution is approximated with a Normal distribution by matching of the mean and variance. The normalizing constant of the posterior distribution, that is required for the calculation of the Bayes factors, is then approximated by the normalizing constant of the matched Normal distribution. The Laplace approach is also potentially useful for calculation of the Bayes factors for multiple tests

of association, as the normalizing constant can be found relatively quickly and easily.

Depending on the chosen prior and likelihood distributions the calculation of the integrals for the competing models can be intractable, and the results have to be approximated through methods such as Markov-Chain Monte Carlo and importance sampling. Although a Monte Carlo approach can be useful in many situations where the evaluation of the integrals is impossible analytically, for the high dimensional SNP data the problems in high run time and in automating the assessment of convergence make the use of these approaches less appealing in this context.

As with all Bayesian analysis suitable parameters for the prior distribution must be chosen, the choice of which is both a positive and a negative feature of Bayesian analysis. In choosing a prior, information that is previously known about a data set can be included in the analysis which can increase the power of a study. However, if unsuitable priors have been chosen then incorrect or misleading conclusions can be inferred. The various issues with selection of prior parameters are documented further by Gelman et al. (2004). Further discussions relating to the appropriate choice of prior parameters for the exactly computed Bayes factors used in this thesis, are given separately for each of the data specific Chapters (5, 6 and 7). BimBam (Servin and Stephens, 2007) is an application that also uses exactly computed Bayes factors, but uses a different approach for the prior distributions (section 4.4.1).

The exact analytic approach that has been chosen to calculate the integrals in equations 1.5 and 1.6, requires the use of conjugate priors, so that the posterior distribution of model parameters can be obtained easily, and integrated in a relatively straightforward manner. Using exact evaluation of the integrals will result in the most accurate answer if the assumptions underlying the chosen distributions are valid, and will also be the quickest method to implement. This approach is though restricted to the exponential family of models, as they have suitable conjugate prior distributions. For the purposes of most continuous phenotype traits they will approximately follow, or can be transformed to represent, a Normal distribution and so the exact analytic computed Bayes factors can be used.

Table 1.2: Interpreting the strength of associations using Bayes factors, represented by B.

$2\log_e B$	B	Evidence against H_0
0 to 2	1 to 3	Minimal
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very Strong

The relationship between p-values and Bayes factors is not straightforward, and as much care must be taken in their interpretation as with their formulation. In frequentist studies a significance level of 0.05 or 0.01 is used, and this is corrected in some way to take into account multiple testing. In a Bayes factor approach, there is a less rigid specification of a value that is required for 'significance', however, table 1.2 presents a reproduction of approximate guidelines for interpretation as specified by Kass and Raftery (1995).

The posterior odds that are obtained from the Bayes factors will be heavily dependent on the choice of the prior odds of association for the test under consideration, and this choice will depend on the situation being considered. For example, in a GWAS it is unlikely that any one particular SNP will have an effect on a trait of interest, however, for a candidate gene study it would be reasonable to expect that there already exists some justification that there is an association within the region being considered. A crucial element in the use of Bayes factors is that this choice of prior odds is made by some careful reasoning, and is not tailored to simply accommodate strong results being artificially found. The upside of choosing prior odds is that, unlike in the use of p-values, a Bayes factor may be influenced less by small changes in the observable data.

1.7 Example Data Sets

Simulations of the coalescent will provide some impressions about the likely effectiveness of the methods that are assessed. However, they are also tested on real data sets as there will inevitably be features of genetic data that are not

replicated in the simulated data. Two data sets were available upon which to test the methods used for association. The first of these data sets in Chapter 5 is that which was initially analyzed using the Nested Clade Analysis method (Templeton et al., 1988), and subsequently in the Treescan (Templeton et al., 2005) approach to provide comparisons between the two procedures. This data set is relatively small and involves homozygous inbred *Drosophila melanogaster* fruit flies, as opposed to human data which would involve heterozygous genotype groupings. The analysis of this data set will however provide useful initial indications as to the similarities and differences of the methods assessed in this thesis.

Data has also been made available from the PAMELA (Padmanabhan et al., 2010) study relating to phenotype and genotype measurements of an Italian population that has not previously been analyzed using Treescan or Bayesian approaches. The available data set involves 70 SNPs relating to a gene that is targeted by anti-hypertensive treatments, and so is a plausible candidate gene for association with the measurements of heart rates and blood pressures that have been recorded.

Both of the data sets have been analyzed using the interactive PheGe-Find application, developed in Chapter 4, which has been created to allow for phenotype-genotype data sets to be simply and quickly analyzed.

1.8 Overview of the Thesis

- Chapter 2 introduces the coalescent data simulation features that PheGe-Sim uses in order to simulate data for fair comparison of the association methods presented elsewhere in the thesis.

- Chapter 3 provides further details behind the methods used in the PheGe-Find application, in particular detailing the formulations of the Bayes factors used throughout the thesis.

- Chapter 4 contains information on the features and methods that are present in the PheGe-Find and PheGe-Sim applications.

- Chapter 5 uses the methods of Chapter 3 on the *Drosophila melanogaster* data set.
- Chapter 6 uses the methods of Chapter 3 on a data set involving heart rate and blood pressure measurements.
- Chapter 7 uses the methods to simulate data as presented in Chapter 2, in order to test the effectiveness of the various association study approaches given in Chapters 5 and 6.
- Chapter 8 provides a summary, some conclusions, and potential further work.

1.9 Aims of the Thesis

- To simulate data that can reasonably represent data sets arising in a fine-scale genetic association study.
- To investigate the advantages and disadvantages of various approaches that can be used in fine-scale genetic association studies, through the use of simulated data sets. In particular, to compare the effectiveness of methods based upon Treescan to those that test each SNP individually; and to assess the potential benefits of using a Bayes factor approach compared to standard frequentist p-values.
- To provide applications that can be used to implement the association study methods for real data sets.
- To use the applications on real data sets to detect any associations present, and to illustrate any deficiencies of the methods that are not apparent with simulated data.

Chapter 2

PheGe-Sim

The PheGe-Sim application (figure 2.1) has been written in the R programming language (version 2.4.1)[1] in order to simulate data from the coalescent process, and if specified, to apply the various association study methods to detect causative loci. PheGe-Sim has been written as a function contained within the Rpanel (Bowman et al., 2007) environment, which creates a windows-based application that can be used to easily specify the many possible input variables. For the variables that can be numerical, or take on a range of categorical values, the simulator will check that the input is of the correct type and that it is consistent with the other selected variable options. If an inappropriate input is entered an error message such as that in figure 2.2 will be produced, suggesting what inputs should be changed to allow the program to run. The following section details the possible input variables, and the options that are permissible to be entered for each.

[1]PheGe-Sim was initially programmed in R 2.4.1 and Rpanel (1.0-4) , but the code will also run in the newest version of both R (2.10.1) and Rpanel (1.0-5). There are however minor differences with the versions, as R 2.4.1 will run with the textentry boxes as they appear, whereas R 2.10.1 requires a carriage return after changing any of the values for the change to be recognized.

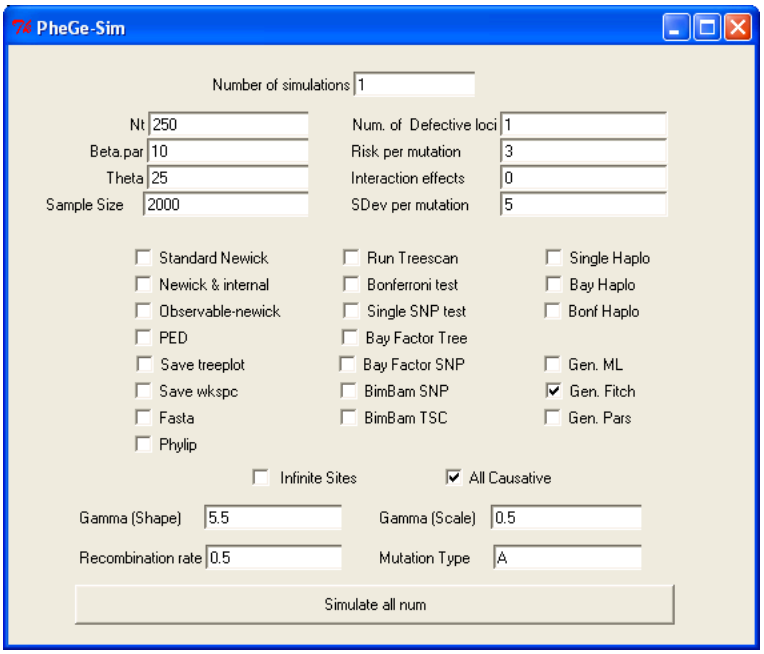


Figure 2.1: PheGe-Sim simulator screen shot.

2.1 Input Options

- ***Number of simulations : {1,2,...}*** = Run and summarize results (section A.4) for the chosen number of simulations.
- N_t ***: {5,6,...}*** = Number of terminal nodes of Ancestral Recombination Graph (ARG) used to simulate genotypes (section 2.2).
- ***Beta.par :*** $[0, \infty)$ = The population size expansion parameter β, where $\beta = 2Nb$ is the scaled expansion rate, N is the effective population size, and b is the expansion rate per generation.
- ***Theta :*** $[1, \infty)$ = Mutation rate per $4N$ generations.
- ***Sample size : {50, 51, 52, ...}*** = The number of individuals with genotypes and phenotypes that are to be sampled from the population.
- ***Newick*** options***: {T,F}*** = Save different variations of Newick format files (Felsenstein et al., 2010) of coalescent trees resulting from the simulated ARG.

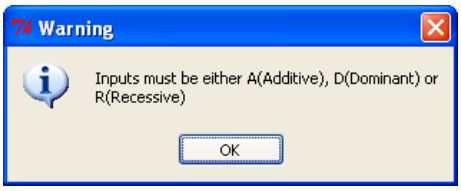


Figure 2.2: Example of error message from PheGe-Sim.

- ***Number of Defective loci : {1,2,...}*** = Specify how many mutations are selected to be causative in the simulations.
- ***Risk per mutation : $(-\infty, \infty)$*** = Specify the average increase in phenotypic measurement for each copy of a causative mutation. This is for the heterozygote mean of an additive model, the value of dominant or recessive models is therefore twice the specified value (section 2.6).
- ***Interaction Effects : $(-\infty, \infty)$*** = Specify any two-way interactions between causative mutations, if it has been chosen to have more than one defective locus. Alternatively, if it is chosen to have only one defective loci, the single value entered for an interaction effect will represent an 'excess additive' form of mutation (section 2.6).
- ***SDev per mutation : (0, ∞)*** = Specify a common standard deviation for each genotype grouping.
- ***Fasta : {T,F}*** = Save the genotypes in the Fasta format (section A.1).
- ***Phylip : {T,F}*** = Save the genotypes in the Phylip format (section A.2).
- ***PED : {T,F}*** = Save the genotypes in a PED file format (section A.3). Will also run the Haploview (Barrett et al., 2005) application to provide a linkage plot of the simulated data set.
- ***Save Treeplot : {T,F}*** = Save the true coalescent plots and ARG used to simulate data (section A.11).
- ***Infinite Sites : {T,F}*** = Use an infinite sites model for the mutations, or a finite-sites model if unselected (section 2.4.1).

- ***All Causative : {T,F}*** = Select whether or not a causative mutation is causative in all positions that it occurs on the ARG (2.4.1).
- ***Gen*** options : ***{T,F}*** = Use Maximum Likelihood, Fitch algorithm or the Parsimony method in the reconstruction of a haplotype tree (sections 4.2 and 4.3).
- ***Run Treescan : {T,F}*** = Runs the Treescan application and interprets the results from the output (sections 1.5.2 and B.2).
- ***Bonferroni Test : {T,F}*** = Implements a test of association with the Bonferroni correction on each SNP and interprets the results (section 4.4).
- ***Single SNP test : {T,F}*** = Runs a single SNP method that corrects p-values in an analogous way to Treescan (section 4.4).
- ***Bay Factor Tree : {T,F}*** = Use the Bayes factors method on haplotype groupings given by the Treescan method (section 1.5.2).
- ***Bay Factor SNP : {T,F}*** = Runs a single SNP method using Bayes factors to test for associations (Chapter 3 and section 4.4).
- ***Haplo*** options ***: {T,F}*** = Runs the appropriate method using haplotypes instead of SNPs to test for associations (section 4.4).
- ***Gamma*** options ***: (0, ∞)*** = The shape and scale parameter of the Gamma distribution of the finite sites model (section 2.4.1).
- ***Recombination rate : [0,2]*** = Specify the recombination rate for the ARG per $4N$ generations.
- ***BimBam SNP : {T,F}*** = Runs the single SNP BimBam (Servin and Stephens, 2007) method (section 4.4.1).
- ***BimBam TSC : {T,F}*** = Runs the Treescan method using the Bayes factors of BimBam (sections 1.5.2 and 4.4.1).
- ***Mutation Type : {A,D,R}*** = Specifies whether the simulated causative variants act in an Additive, Dominant, or Recessive manner. A non-additive

model for one causative site can be obtained using an additive model with an interaction (section 2.6).

- ***Simulate all num*** **:** = Upon selection will start execution of the simulations, or will notify of any errors in the specified entries.

2.2 Simulate ARG matrix

Input:

Recombination rate - ρ

Number of (terminal) nodes - n

Population expansion parameter - β

The first stage in simulating the coalescent process consists of generating a vector representing the sequence of coalescent and recombination events that occur, going back in time from the present. A time is simulated until either a coalescent or recombination event occurs, based upon the number of nodes that are currently remaining (k) and the recombination parameter (ρ). The waiting time to a coalescent event is modeled with an exponential distribution with parameter $\frac{1}{2}k(k-1)$ and the time till a recombination event is modelled with an exponential distribution with parameter $\frac{1}{2}\rho k$; therefore the waiting time till an event of either type is modelled as an exponential with the sum of these rates:

$$\text{Exponential}\left(\frac{k(k-1)}{2}+\frac{\rho}{2}k\right). \tag{2.1}$$

The type of event occurring is then chosen according a Bernoulli distribution with probabilities given according to the values from each of the two independent coalescent and recombination distributions. If the simulated event is of a coalescence between two nodes, then the two nodes that coalesce are randomly chosen from the available nodes at that stage in the simulation, as illustrated in figure 2.3 where nodes 1 and 2 are chosen resulting in the creation of node 6. The number of nodes present is reduced by one, i.e., $k \to k-1$, and so in figure 2.3 after the first coalescent event only the set of nodes {4,3,6,5} remains.

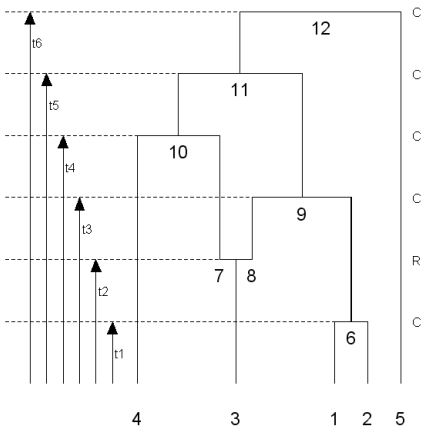


Figure 2.3: Sequence of coalescent(C) and recombination(R) events and times. The set of {t1,t2,t3,t4,t5,t6} represents the sum of the waiting times at each event.

The next time till an event occurs is then generated, and if this event happens to be a recombination event then the result is to increase the number of nodes remaining by one, i.e., $k \rightarrow k + 1$. The lineage on which recombination occurs is chosen randomly from those that are present at that stage in the simulation, so in figure 2.3 node 3 has been chosen thus resulting in the creation of branches 7 and 8. The position of the recombination event is chosen according to a Uniform distribution covering the section of DNA. The process of sampling coalescent and recombination events is then continued until all nodes have coalesced to a common ancestor, so that $k = 1$, which in this example occurs at node 12, the most recent common ancestor of this sample. Careful selection of (ρ) must be made to ensure that the most recent common ancestor is reached within a 'reasonable' amount of time, and is therefore restricted to an upper limit of two $4N$ recombination units.

As the sequence of coalescent and recombination events is occurring, the evolutionary time between events is stored. This results in the branch lengths of a particular lineage being simply calculated from the differences in event times on that lineage. In order to model the potential effect of a non-constant population size (Hein et al., 2005), the branch lengths can be adjusted for exponentially

increasing population size using:

$$t_k = \frac{1}{\beta}\log(1 + \beta t_k^*), \tag{2.2}$$

where t_k represents the adjusted branch lengths, β is the scaled growth rate, and t_k^* are the branch lengths from a coalescent process with no population growth.

Output of:

ARG branch lengths - *branch.length.ARG*

ARG node labels - *ARG.matrix*

Sequence of coalescent and recombination events - *ARG.matrix*

2.3 Plotting the ARG

Inputs:

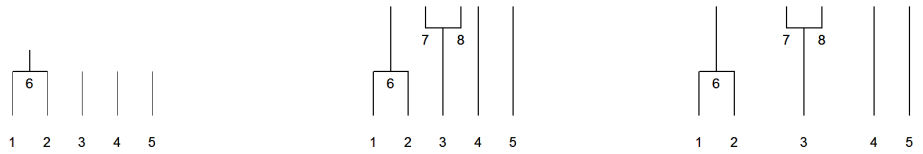
ARG branch lengths - *branch.length.ARG*

ARG node labels - *ARG.matrix*

Sequence of coalescent and recombination events - *ARG.matrix*

The ARG that has been generated in the previous section can be plotted, with the aim being to produce the clearest plot possible by reducing the number of lines that cross over as a result of recombination events. Although not of primary concern to any of the results of the simulations, the methods involved in determining the relationship between different lineages for the plot are closely related to the format of the methods used to allocate the list of branches for each terminal node. This information is then used once mutations have been added to the branches in section 2.4 to determine the sequence of bases of each of the sample haplotypes.

Figure 2.4 and table 2.1 represent the first three stages involved in plotting the ARG that was illustrated in figure 2.3, and demonstrates the algorithm that can be generalized for plotting any ARG. At the initial stage of the plotting algorithm, none of the terminal nodes are fixed in position on the x-axis. A coalescent event then occurs which fixes nodes 1 and 2 in position adjacent to



(a) (b) (c)

Figure 2.4: First three stages involved in plotting the example ARG of figure 2.3.

each other, as illustrated in figure 2.4(a).

The second event that occurs, as shown in figure 2.4(b), is a recombination on the lineage of node 3. A recombination event does not alter the ordering of terminal nodes, but the spacing of the terminal nodes is adjusted so as to maintain a clear appearance of the plot (2.4(c)). Information regarding the ancestry of the terminal node 3 is retained, with it being either connected to lineage 7 for sites to the left of the recombination breakpoint, or connected to lineage 8 for sites to the right of the recombination position.

The next event that occurs is a coalescent event involving two nodes that have been involved in events from earlier on in the plotting process. In this case, the nodes that are joining together are the node labelled 6 resulting from the first coalescent event, and the node labelled 8 that is the result of the first

Table 2.1: First three stages involved in plotting the example ARG of figure 2.3. The 'Fixed' row refers to terminal nodes that have been fixed in a sequence, and the 'Current Nodes' row illustrates the nodes that are remaining at each stage of the plotting.

Stage in figure 2.4	-	a	b	c
Fixed		1 2	1 2	1 2
Current Nodes	1 2 3 4 5	3 4 5 6	4 5 6 7 8	3 4 5 6

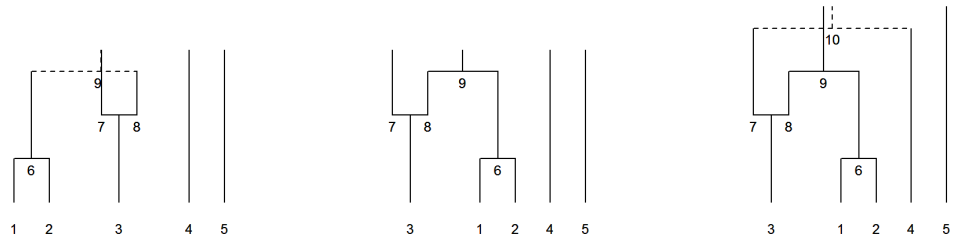


Figure 2.5: Second stages involved in plotting the ARG of figure 2.3.

recombination event. The plotting algorithm attempts to minimize the number of instances of lines crossing over to maintain the clearest plot, and so the sequence of terminal nodes in figure 2.5(b) is chosen in preference to that of figure 2.5(a).

The fourth event in this example is that of a coalescence between nodes 4 and 7. The algorithm again checks for crossing over of branches, and subsequently chooses the sequence of terminal nodes that minimizes the number of instances of this happening. For this coalescent event, that results in the terminal node sequence being that as shown in 2.6(a) as opposed to that of figure 2.5(c).

The penultimate event occurring in figure 2.6(b) is of a coalescence between nodes 9 and 10, resulting in the creation of node 11. As the sequence of terminal nodes have already been fixed for both nodes 9 and 10, the nodes are joined together irrespective of whether there are any branches crossing over. This can lead to the base of a plot being well structured but with the top of a plot being cluttered because of many branches crossing over: as a greedy algorithm is being used that decides upon the best configuration based only on the current level

Table 2.2: Second stages involved in plotting the ARG of figure 2.3.

Stage in figure 2.5	a	b	c
Fixed	1 2 3	1 2 3	3 1 2 4
Current Nodes	4 5 7 9	4 5 7 9	9 10 5

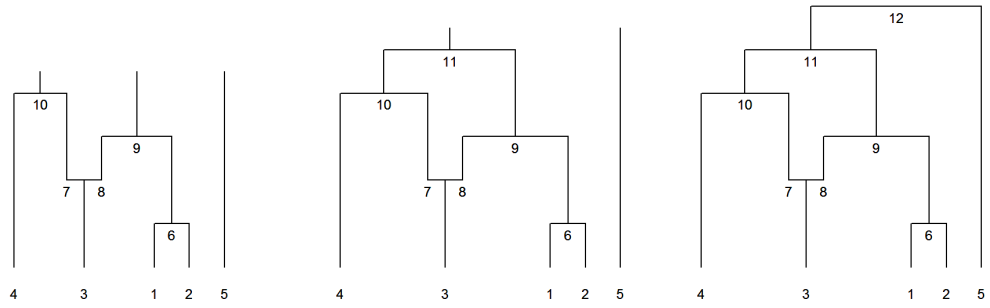


Figure 2.6: Final three stages involved in plotting the ARG of figure 2.3.

in the process. This situation is however unlikely to occur in practice unless a relatively large recombination rate is chosen in comparison to the rate of coalescence. Possible improvements could be made to the algorithm to improve the appearance of plots in situations such as this, although the algorithm works successfully in its current form for its primary function of determining the sequence of branches for each terminal node.

The final event for this example, as illustrated in figure 2.6(c), is of the coalescence of nodes 11 and 5. In this instance there is no preference for an ordering of {4,3,1,2,5} or {5,4,3,1,2}, as neither alternative will result in branches crossing over each other on the plot. The sequence of terminal nodes is therefore chosen randomly from the two possibilities.

The resultant plot can then be saved to an appropriate file if the *Save Treeplot* option is selected on the Rpanel simulator. The information regarding the sequences of internal nodes for each side of the recombination breakpoint are saved

Table 2.3: Final three stages involved in plotting the ARG of figure 2.3.

Stage in figure 2.6	a	b	c
Fixed	1 2 3 4	4 3 1 2	4 3 1 2 5
Current Nodes	9 10 5	11 5	12

$$\begin{array}{c} \\ TN_1 \\ TN_2 \\ TN_3 \\ TN_4 \\ TN_5 \end{array}\begin{pmatrix} 1 & 6 & 9 & 11 \\ 2 & 6 & 9 & 11 \\ 3 & 7 & 10 & 11 \\ 4 & 10 & 11 & 0 \\ 5 & 0 & 0 & 0 \end{pmatrix} \qquad \begin{array}{c} \\ TN_1 \\ TN_2 \\ TN_3 \\ TN_4 \\ TN_5 \end{array}\begin{pmatrix} 1 & 6 & 9 & 11 \\ 2 & 6 & 9 & 11 \\ 3 & 8 & 9 & 11 \\ 4 & 10 & 11 & 0 \\ 5 & 0 & 0 & 0 \end{pmatrix}$$

(a) (b)

Figure 2.7: Matrices of terminal nodes (TN) for the left (a) and right (b) of the recombination breakpoint.

in a series of matrices for each of the Terminal Nodes (TN). Continuing the example ARG used previously, the saved matrices would take the form of that in figure 2.7 for the left and right hand sides of the recombination event.

Outputs of:

Sequence of internal nodes - *lineage.nodes.matrix*

ARG plot - Saved to file

2.4 Assign Mutations onto the ARG

Inputs:

Branch Lengths - *branch.length.ARG*

Mutation Rate - θ

Number of defective Loci - *defect.loci*

Sequence of internal nodes - *lineage.nodes.matrix*

The ARG that has been created can now have the mutations that affect the observed genotypes superimposed on it. This is possible due to the coalescent theory allowing the evolutionary and mutation history to be separable from each other, if the effects of mutations are assumed to be selectively neutral. Each branch is treated independently, and the number of mutations that occur on a branch is chosen according to a $Poi(\theta t)$ distribution, where t represents the branch length and θ represents the mutation rate. Continuing the example of the

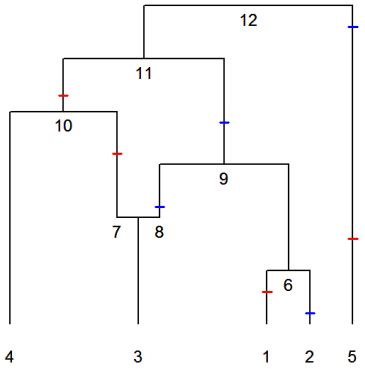
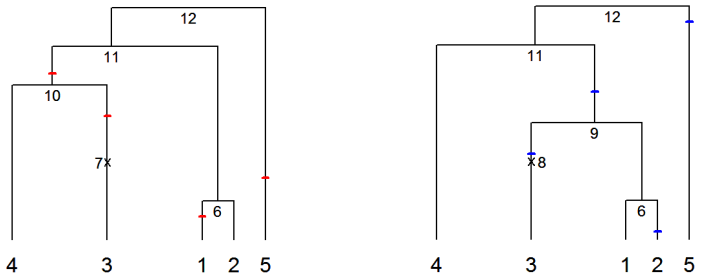


Figure 2.8: ARG with overlaid mutations, with red lines indicating mutations occurring on sites to the left of the recombination breakpoint, and blue mutations representing mutations that occur to the right of the site at which recombination occurs.

ARG used in sections 2.2 and 2.3, the plot of an ARG overlaid with mutations is shown in figure 2.8.

In order to allocate the positions of sites that mutate at each SNP on the ARG, the randomly selected locations at which recombination events occur must also be taken into consideration. For each mutation that occurs on the ARG, the position of the site that the mutation represents is chosen according to the regions that are possible on the branch under consideration. For example, in figure 2.9(a) only sites to the left of the recombination breakpoint are possible on branch 7, whereas the complete set of sites can mutate on branch 11 as this occurs in the coalescent trees of both regions. This procedure ensures that all mutations correspond to a change in the base at a selected site, except in cases of trapped ancestral material, where branches of the ARG are not possible in any of the coalescent trees and so do not contribute to the sample.

Each recombination event will result in a separate coalescent tree describing a different region of a sequence of DNA bases. To ensure that a model of infinite sites is possible, the number of sites in a region must be at least as large as the number of mutations that are sampled to occur for that region. An initial sequence length of 500 is used as this is assumed to be at a level far above that of the number of SNPs in a fine-scale study. If a larger number of mutations are



(a) Left hand side of recombination event

(b) Right hand side of recombination event

Figure 2.9: Mutations allocated to coalescent trees.

sampled than there are possible sites in a region, the region is simply extended in size to accommodate the number of mutations. The recombination breakpoints are then adjusted according to the new size of the regions. This procedure is possible as the distances between recombination breakpoints and SNPs are assumed to be irrelevant in practice. Dealing with the recombination breakpoints in this manner also prevents large numbers of simulations from being discarded for not allowing the possibility of the specified number of mutations to occur within a region.

The information regarding the terminal node sequences are determined for each coalescent region separately, and the terminal node sequence for the entire range of the ARG is obtained by simply joining these regions together. Figure 2.10 illustrates matrices that could result from the coalescent trees, with mutations as in figures 2.9(a) and 2.9(b). Information from the stored matrices can also be shown graphically (figures 2.10(d), 2.10(e) and 2.10(f)), using 'lineplots' that illustrate the relative locations of mutations for each terminal node.

2.4.1 Finite Sites

Additional Input:

Infinite sites option - *infinite.sites*

$$\begin{matrix}1\\2\\3\\4\\5\end{matrix}\begin{pmatrix}A & G & A & A\\ A & C & A & A\\ A & C & G & G\\ A & C & G & A\\ G & C & A & A\end{pmatrix}$$

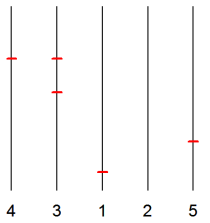
(a) LHS

$$\begin{matrix}1\\2\\3\\4\\5\end{matrix}\begin{pmatrix}A & A & T & T\\ A & T & T & T\\ A & A & T & C\\ A & A & G & T\\ C & A & G & T\end{pmatrix}$$

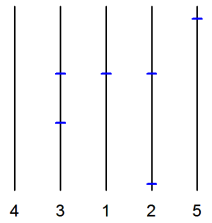
(b) RHS

$$\begin{matrix}1\\2\\3\\4\\5\end{matrix}\begin{pmatrix}A & G & A & A & A & A & T & T\\ A & C & A & A & A & T & T & T\\ A & C & G & G & A & A & T & C\\ A & C & G & A & A & A & G & T\\ G & C & A & A & C & A & G & T\end{pmatrix}$$

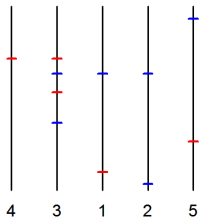
(c) ALL



(d) LHS



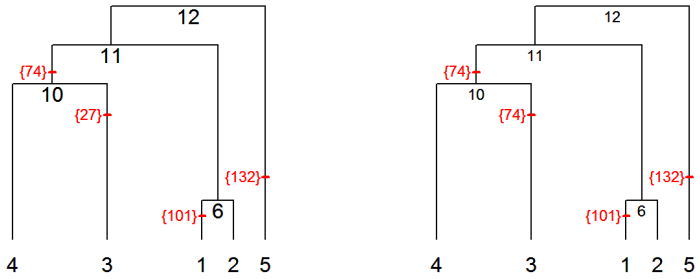
(e) RHS



(f) ALL

Figure 2.10: *phy.matrix* for both left (a) and right (b) hand sides of the recombination event. This results in the matrix *phy.matrix.ARG* (c) which covers the full ARG. The recombination event in the region can be determined by the observation that there exists 4 haplotype configurations (AT, GT, GG, AG) between the third snp of (a) and the third snp of (b). Figures (d) to (f) illustrate the 'line plots' corresponding to (a) to (c).

The simulator allows the possibility of either a finite or infinite-sites model of mutation to be implemented, since recurrent mutation may have a strong effect. Figure 2.11(a) illustrates the model of infinite sites whereby each mutation is constrained to occur only once on the ARG. A finite-sites model corresponding to the same number of mutations on the tree, but where site number 74 has mutated in two positions is illustrated in figure 2.11(b). In this case the effect of the mutation on branch 10 is reversed by the mutation on branch 3, but is still observable on branch 4. The program allows for the possibility of a mutation to be completely removed by reverse mutations, however, a site that is allocated to cause a change in phenotype measurement in section 2.6 must occur in two forms. Although unlikely, the program also guards against all the sites affected by mutations existing in only one form by discarding simulations where this occurs.



(a) Infinite Sites (b) Finite Sites

Figure 2.11: Illustration of infinite and finite-sites models.

If the program consistently fails to complete due to the set of initial parameter choices, the program will terminate and inform the user that a different set of parameter values should be chosen.

It is ensured that there are only two bases present at each SNP, irrespective of how many mutations at that position occur on the ARG, with the causative form of the mutation always chosen to be the base that a mutation changes to away from the ancestral base.

2.4.2 Gamma Model of Finite Sites

Additional Inputs:

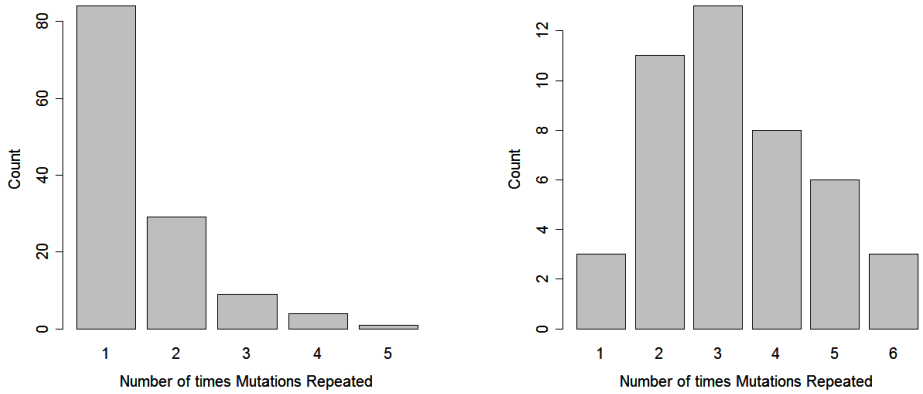
Shape parameter of Gamma distribution - Γ_α

Scale parameter of Gamma distribution - Γ_β

Ancestral Recombination Graph - *ARG.matrix*

Recombination breakpoints - *break.points*

The first stage of the finite sites model is to replicate the generation of mutations according to the procedure of infinite sites as previously described in section 2.4.1. For each SNP that is sampled, the number of times that it occurs on the ARG is randomly chosen based on a Gamma distribution with shape parameter Γ_α and scale parameter Γ_β. A Gamma distribution has been chosen as it is plausible that there are a few sites that mutate many times, whereas at the majority



(a) Gamma distributed rates, with shape parameter of 1 and scale parameter of 1

(b) Gamma distributed rates, with shape parameter of 6 and scale parameter of 0.5

Figure 2.12: An example from the discrete Gamma distribution for finite sites.

of locations there will only be a small number of mutations. To ensure that only positive integers are selected, the sampled value from the Gamma distribution is rounded up to the nearest integer value which results in discrete groups of the sampled mutations, as illustrated in figure 2.12. Although the y axis scales are different for both 2.12(a) and 2.12(b), it should be noted that in both cases the total number of mutations involved is the same. It can also be the case that if certain values of Γ_α and Γ_β are chosen, the data can begin to resemble the infinite sites model.

As this is a procedure involving random sampling it can, and indeed will, be the case that the total number of mutations sampled to occur according to the finite sites method exceeds the number of mutations that actually occur under the infinite sites model in section 2.4. In order to deal with this situation, various corrections are applied to the sampling of the number of mutations, so that the correct mutation rate is retained and in order to obtain as close to the desired finite sites distribution as possible. The steps taken are shown in the flow chart in supplementary section A.10. In practice in the majority of situations, the steps taken have minimal impact on the specified finite sites distribution.

The sites involved of each chosen mutation are then randomly 'thrown' onto

the coalescent plot of the region involved, and subsequently the ARG, as shown in figure 2.11(b).

Outputs of:

Ancestral Recombination graph with site positions - *ARG.matrix*

2.5 Collapsing the ARG

Inputs:

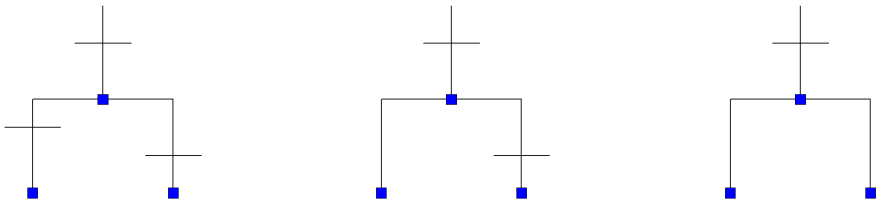
ARG branch lengths - *branch.length.ARG*

ARG node labels - *ARG.matrix*

The ARG, and therefore the matrix *phy.matrix.ARG*, that have been created have to now be collapsed to a data structure containing only branches on which mutations occur. The reasoning for this is that in an association study, using the available data, an analyst would not be able to reconstruct ancestral events that have no discernible effects on an individuals genotype. The only features that would be observable are the unique haplotypes of a sample of individual's and their corresponding phenotypes. It would therefore be inappropriate to retain any information from the construction of the data other than the unique haplotypes and phenotypes of a simulated population, before analysis using the methods detailed in the following chapter.

There are four situations that can occur in determining whether a branch should be collapsed, applied separately to each of the coalescent regions of the ARG. Three are given in figure 2.13, where each plot represents a section of the ARG with potential configurations of mutations. The fourth condition is that of a branch where no mutations occur at all, and as such the entire section would be removed until there occurred further back in the tree a mutation structure similar to one of the other three configurations.

The removal of branches from figure 2.13 results in the sections of trees being reduced to that shown in figure 2.14. This procedure is applied to all branches involved in each non-recombining region and, with the example data set used throughout this chapter, would result in trees relating haplotypes given in figure



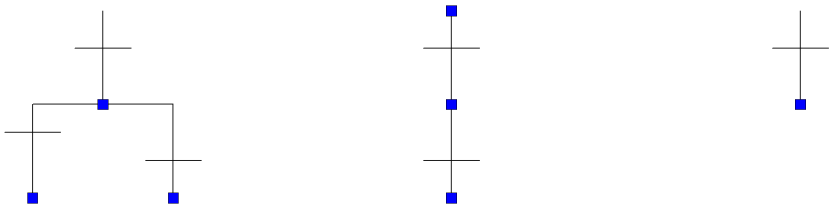
(a) (b) (c)

Figure 2.13: Examples of the initial configuration of branches from the ARG that can subsequently be collapsed due to some branches having no occurrences of observable mutations, with the horizontal lines representing such mutations.

2.15.

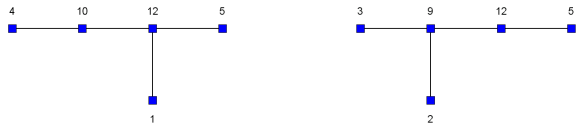
The collapsed ARG is then used in order to generate the potential haplotypes that can reasonably be sampled within the simulation, as any unobserved internal nodes will be discounted. The information regarding the number of terminal nodes on the un-collapsed tree carrying each of the collapsed trees nodes is also used in determining the relative probabilities of selecting an individual haplotype for the sample.

To create a sample 'individual' from a population, two haplotypes are randomly selected, based on the assumptions discussed in section 1.3. Information regarding the sample from a population is retained in the matrix *c.p.details*, which is subsequently used in assigning phenotype scores for each sampled individual.



(a) (b) (c)

Figure 2.14: Examples of the new structures that result after collapsing branches of the forms in figure 2.13.



(a) Left of split collapsed (b) Right of split collapsed

Figure 2.15: Collapsed coalescent haplotype trees for the example data, for both the left and right hand side of the recombination breakpoint.

Outputs:

Sampled individuals' haplotypes - *c.p.details*

Collapsed matrix of haplotypes - *phy.matrix.ARG*

2.6 Assigning phenotype scores

Inputs:

Collapsed matrix of haplotypes - *phy.matrix.ARG*

Number of causative mutations - *defect.loci*

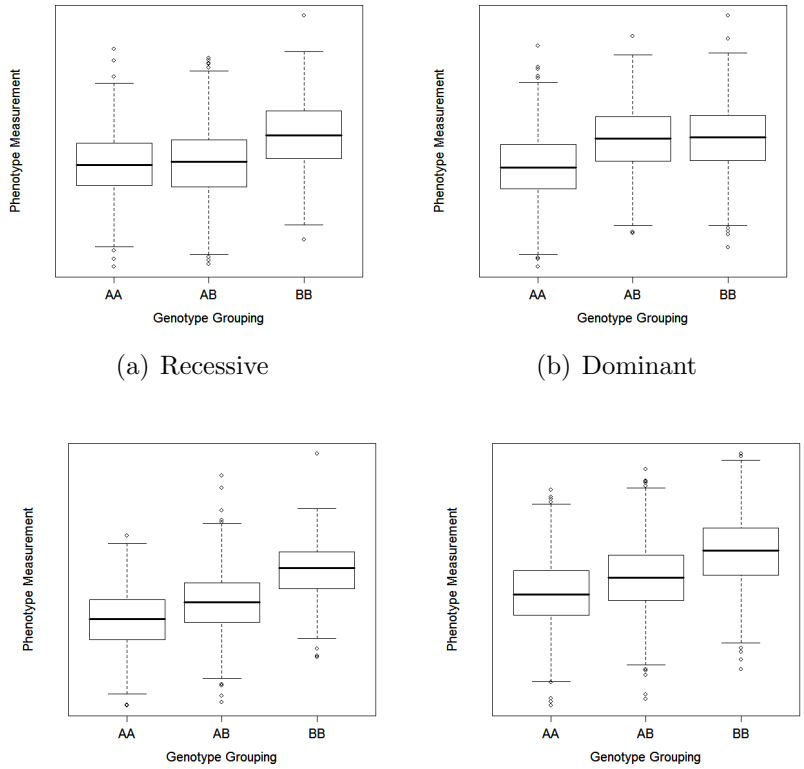
Causative effect per mutation - *mut.par.rsk*

Standard deviation of phenotype scores - *sids.par.rsk*

Types of mutation(s) - *mut.type*

Interaction effects between mutations - *interact.effect*

In PheGe-Sim, phenotype scores are assigned to an individual based upon a range of parameters. The first of these is the choice of the number of causative mutations (*defect.loci*) that will increase the mean phenotype of affected individuals by effect sizes given by the parameter *mut.par.rsk*. If more than one mutation is chosen as causative, there is the potential to specify that the mean effect size of each mutation is different to obtain different patterns of phenotype measurements. The phenotype of each sampled individual is chosen according to a normal distribution, with the mean parameter fixed according to the causative mutations that are present in that individual, whether any interactions are present, and the specified mutation types. The standard deviation parameter is chosen to be the same value for each distribution, and is specified by the *sids.par.rsk* parameter.



(a) Recessive (b) Dominant

(c) Additive (d) Excess Additive

Figure 2.16: Models of phenotype measurements.

For each mutation that is chosen to be causative, the haplotypes that are affected by the mutation are determined by following the structure of the coalescent tree away from the root. If a model of finite sites is chosen, it may be that a causative mutation is cancelled out in some of the haplotypes that would otherwise have been affected. It is however ensured that at least one haplotype will carry the causative form of the mutation. When using a finite-sites model a choice can also be made with the *All Causative* parameter as to whether a mutation is causative in all positions at which it occurs on a tree, or that it is only causative on one specific branch of the ARG.

Another influencing factor on the phenotype scores is the choice of whether a mutation is to act in an additive, dominant or recessive manner. In the simulations, a recessive model requires the causative mutation to be present in both

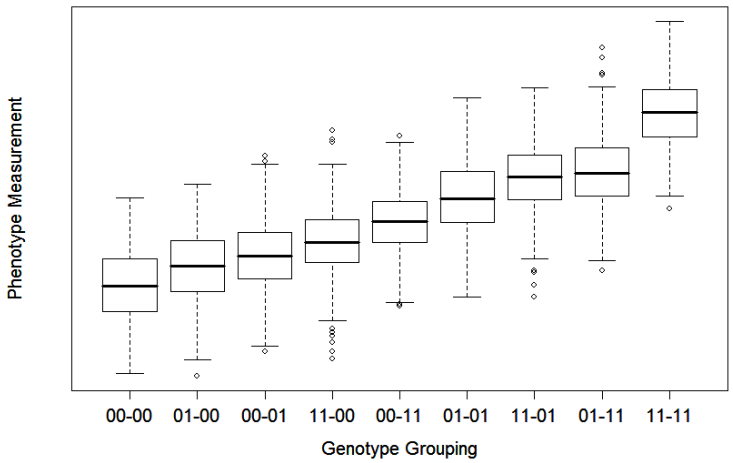


Figure 2.17: Example of the possible groups resulting from two causative SNPs and their interaction. The two copies present for each individual are shown, with '1' being the causative form for both of the two SNPs.

haplotypes of an individual to cause an increase in phenotype, whereas a dominant model requires at least one copy to cause the relevant increase in score. These definitions are chosen as the simulations work on the arbitrary assumption that mutations cause an increase in scores as opposed to a decrease in scores. The program will however operate under either condition.

If only one causative SNP is selected the *interaction effects* option can be used to specify a model of the mutant homozygous group, containing the causative form of the mutation, having an extra effect compared to the heterozygous grouping. For recessive models this will have no relevance and will only act as an addition of phenotype measurement to the already specified mutation effect size. However, in the case of a specified additive mutation the interaction will allow for specification of models with non-perfect additivity, the type of model being dependent on whether the interaction has a positive or negative value. A positive interaction effect can specify an 'excess additive' model as shown in figure 2.16(d), whereby the effect of having two copies of the causative mutation is stronger than an additive presence of only one causative mutation (figure 2.16(c)). If on the other hand the interaction is chosen to have a negative effect, the additive model

will begin to resemble the phenotype pattern of the dominant model. If the interaction is chosen to be larger than the original effect size, this can result in the situation whereby the mean of the heterozygous grouping is higher than either of the two homozygous groups.

If two or more mutations are chosen to be causative, then the program allows for the effect of two-way interactions (i.e., epistasis) to exist if the causative forms of both mutations are present. Interaction effects of up to the number of two-way combinations involving the number of causative mutations can be specified, with the allocation of effects to each being randomly chosen amongst the possible combinations.

Outputs:

Sampled phenotypes for each individual - *c.p.details*

2.7 Output Files

PheGe-Sim creates various files and folders as determined by the choice of input parameters, features previously described in section 2.1 and throughout this chapter. In particular, for each of the association methods and tree construction approaches, a folder will be created to contain the appropriate summary files of that method.

In addition to the output files of the programs that are collected into relevant folders, two output files are also created that collate together the details and results of all the simulations. The *sim_results* file summarizes information about the effectiveness of each method that is being considered in terms of finding the true causative mutations. An example of *sim_results* is given in section A.4, where it can be seen that for each simulation and method comparisons are made between the true causative mutations and any mutations that have been 'found' by the method. The left hand side of table 2.4 illustrates the criteria used for assessing the effectiveness of each method in finding causative SNPs for the simulations when the true causative mutations are known.

In addition to assessing whether causative SNPs have been found correctly using the Treescan method, it is also tested as to whether Treescan correctly

Table 2.4: Determining the criteria of found SNPs or branches.

Criteria	SNP	Branch
Found	As determined by the Frequentist (figure 4.6) or Bayesian (figure 4.7) approach that is taken	
Correct Find	A SNP that has been found corresponds to the true causative SNP.	A found haplotype/branch contains the causative form of a true causative SNP.
False + (Type I error)	A found SNP is not any of the true causative SNPs.	A found haplotype contains no SNPs in their causative form.
False − (Type II error)	A true causative SNP has not been found.	NA
Linked +	A SNP has been found that is in perfect linkage with a true causative SNP.	NA

finds branches containing the true causative SNP(s). The rationale for this is that the tree construction method can suggest multiple site changes on the same branch, and will therefore have to declare all of these SNPs as being 'found' if that branch is significantly associated with the phenotype. This can result in a high False Discovery Rate that is largely due to errors in the tree construction method, and not necessarily errors in the Treescan method itself. The 'branch' correction method can therefore be used as a potentially fairer comparison in terms of the False Discovery Rates. However, the downside of this approach is that there is insufficient knowledge to ascertain specifically which of the SNPs upon a haplotype are indeed causative. This approach is similar to the methods employed when using the simple haplotype methods of association.

At the end of the *sim_results* file, a summary of the performance of each association method in finding true causative mutations over all the simulations is also displayed. The False Discovery Rate (FDR), that can be used as an indicator of what proportion of the mutations that are declared as significantly associated

Table 2.5: Summary of output files and folders of PheGe-Sim. Items preceded with † are common to PheGe-Sim and PheGe-Find, and are described in Chapter 4.

Location	Description
Sim_Out	Directory containing output folders
Folders	
PHYLIP	PHYLIP format files of the unique genotypes
FASTA	FASTA format files of the unique genotypes
†Bayes_Factor_TSC	Treescan version of Bayes factors
†Bayes_Factor_SNP	Single SNP (and haplotype) Bayes factor files
†Fitch_Files	Files involved in the Fitch tree construction from PHYLIP
†Max_L_Files	Files involved in the Maximum Likelihood construction from PHYLIP
†Pars_Files	Files involved in the Maximum Parsimony Tree construction from PHYLIP
Treeplot_Files	Plots of the ARG, coalescent and line plots.
†Treescan_Input_Files	Input Treescan files for all the tree construction approaches
†Treescan_Out_Files	Treescan Output files for all the tree construction approaches
StdNewick_Out_Files	Newick output files for the true coalescent regions
Int.Newick_Files	Newick output files, with labeled internal nodes, for the true coalescent regions
Observable_Trees	Collapsed haplotype trees of each coalescent region
†SingSNP_Out_Files	Results files of the Single SNP (and haplotype) analysis based on Treescan
†PED_Files	PED files and Haploview Linkage plots
Wkspc_Folder	R workspace of the key variables that have been saved in the simulation
†Bonf_Files	Results files of the Bonferroni SNP (and haplotype) analysis
†BIMBAM	Output Files from BimBam relating to single SNP analysis
†BIMBAM.TSC	Output Files from BimBam relating to Treescan based analysis
Files	
Sim_Results.txt	Text file summarizing results of the association approaches
details.txt	Text file summarizing various features of the parameters involved in the simulations

are indeed truly associated with the phenotype, is also displayed in the *sim_results* file. The definition of the false discovery rate is:

$$\text{FDR} = \frac{\text{Sum of False Positives}}{\text{Sum of False Positives} + \text{Sum of correctly found SNPs}}. \tag{2.3}$$

The second file specific to PheGe-Sim is the *details* file, an example of which is given in section A.5. This file contains information about the causative sites for each simulation, and the haplotypes that are affected as a result of carrying the causative form of the mutation. A brief summary of the input parameters that have been chosen for that simulation is also given at the end of the *details* file. A brief summary of all the output files from the PheGe-Sim application is given in table 2.5.

Chapter 3

Association Methods

The exactly computed Bayes factors that are used in the PheGe-Sim application are discussed in the following sections, for the different models of mutation effects: Additive, Dominance, Recessive and a General model. Section 3.1 introduces the use of the exactly computed analytic Bayes factors in a similar manner to that suggested by Balding (2006), for use in case-control studies. Sections 3.2 - 3.5 subsequently introduce the Bayes factors that have been developed and implemented here for use with continuous phenotype measurements, in the case of assessing only one SNP (or branch) with the outcome. The concepts that have been applied for the single SNP/branch tests are then extended in sections 3.6 to 3.8, in order to accommodate the testing of multiple SNPs in combination.

3.1 Binary Data

In the supplementary information of 'Bayesian approaches to single-SNP association' (Balding, 2006), a Bayes factor approach was used as a method to find causative SNPs in a case-control setting. The Bayes factors are used to test for evidence of an association comparing the null model of no association (M_0), against the alternative (M_1) that there are differences in the proportions of cases and controls across the heterozygote and the two homozygote genotype classes. The likelihood for the data (D) is specified in terms of the probability, θ, of an

individual involved in the study being a case, in a binomial form:

$$P(D|\theta) = \binom{n_A + n_U}{n_A} \theta^{n_A} (1-\theta)^{n_U}, \tag{3.1}$$

where n_A is the number of cases and n_U is the number of controls. This applies independently in the three genotypes, with a common θ under M_0, and three different θs under M_1.

The approach taken by Balding is to use a Uniform prior for θ in (0,1), for illustrative purposes and to ensure a proper posterior distribution. Alternatively a Beta prior distribution can be used and, as it is conjugate to the likelihood, this will also ensure that a proper Beta posterior distribution is obtained. One reason for using the Beta in favour of the Uniform prior is that the resultant Bayes factor may be less sensitive to small sample sizes of cases or controls at any one SNP, depending on the choice of hyperparameters. Thus the prior is taken as:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{3.2}$$

for the single θ in M_0, and as a product of three such terms under M_1.

The Beta prior of equation 3.2 is a generalization of the Uniform prior, which can be obtained by setting both α and β equal to one. Initial impressions using the data in Chapter 6 suggest that setting α and β to approximately five will be effective in dampening down the stochastic effects when there is only a small sample size available for the test under consideration. However, this can be quite an informative prior for a low sample size and tests with lower values of α and β may also be appropriate depending on the specific context of the outcome being considered.

The posterior distribution using the Beta prior can be calculated analytically, as follows:

$$\begin{aligned} p(\theta|D) &\propto \theta^{n_A}(1-\theta)^{n_U}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{n_A+\alpha-1}(1-\theta)^{n_U+\beta-1} \\ \text{i.e. } \theta|D &\sim \text{Be}(\alpha+n_A, \beta+n_U). \end{aligned} \tag{3.3}$$

which applies to the global θ of M_0. Integrating out the parameter θ to obtain the marginal likelihood for the null model, M_0, will result in:

$$P(D|M_0) = \int_0^1 p(D|\theta)p(\theta)d\theta = cB(\alpha + n_A, \beta + n_U), \tag{3.4}$$

where c is a constant. In a similar manner, assuming the groups are independent, the marginal likelihood for an alternative model, M_1, allowing for different probabilities of being affected within each genotype class is:

$$\begin{aligned} P(D|M_1) = & cB(n_{A0} + \alpha, n_{U0} + \beta) \\ & \times B(n_{A1} + \alpha, n_{U1} + \beta) \\ & \times B(n_{A2} + \alpha, n_{U2} + \beta), \end{aligned} \tag{3.5}$$

where the subscripts 0, 1 and 2 refer to the three genotype groups, and c is the same constant as in (3.4). Therefore, the Bayes factor for the alternative model compared to the null model in the case of binary data reduces to (3.5) divided by (3.4). The posterior odds of the alternative model can then be obtained by multiplying the Bayes factor by a suitable choice of prior odds.

3.2 Null Model

This section describes a null model for the case of continuous phenotype data, where the null model involves a single parameter for describing the means of each of the three genotype classes in the first round of tests. Subjects are assumed to be independent, and the null model assumes a normal likelihood for the data, for the phenotype of the ith subject:

$$y_i \sim \mathrm{N}\left(\mu\,,\,\sigma^2\right). \tag{3.6}$$

Conjugate priors for the mean μ and the within group variance σ^2 are assumed:

$$\mu\,|\,\sigma^2 \sim \mathrm{N}\left(\mu_0\,,\,\frac{\sigma^2}{\kappa_0}\right), \tag{3.7}$$

$$\sigma^2 \sim \text{Inv-}\chi^2\left(\nu_0\,,\,\sigma_0^2\right), \tag{3.8}$$

where μ_0, κ_0, ν_0 and σ_0^2 are hyperparameters.

The full joint distribution is obtained by multiplying the likelihood for the n data points (the total number of subjects in all 3 genotype groups), by the prior densities for μ and σ:

$$\begin{aligned} p(y\,,\,\mu\,,\,\sigma^2) = c\,(\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{(n-1)s^2+n(\mu-\bar{y})^2}{2\sigma^2}\right) \\ \times \quad (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)}\exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \\ \times \quad \sigma^{-1}\exp\left(-\frac{\kappa_0(\mu-\mu_0)^2}{2\sigma^2}\right), \end{aligned} \tag{3.9}$$

where c is a constant that does not involve μ and σ, and takes the form:

$$c = \frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\,\frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\,(\sigma_0^2)^{\frac{\nu_0}{2}}\,(2\pi)^{-\frac{n}{2}}. \tag{3.10}$$

The resulting posterior distribution for μ and σ^2 is a N-Inv$\chi^2(\mu_n, \frac{\sigma_n^2}{\kappa_n}; \nu_n, \sigma_n^2)$ distribution (Gelman et al., 2004), due to the priors being conjugate, the form of which is:

$$p(\mu,\sigma^2|y) \propto \sigma^{-1}(\sigma^2)^{-\left(\frac{\nu_n}{2}+1\right)}\exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2+\kappa_n(\mu_n-\mu)^2]\right), \tag{3.11}$$

where

$$\begin{aligned} \mu_n &= \frac{\mu_0\kappa_0+n\bar{y}}{\kappa_0+n}, \\ \kappa_n &= \kappa_0+n, \\ \nu_n &= \nu_0+n, \\ \sigma_n^2 &= \frac{\nu_0\sigma_0^2+\sum_{i=1}^{n}(y_i-\bar{y})^2+\frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2}{\nu_n}. \end{aligned} \tag{3.12}$$

To obtain the marginal likelihood for the null model, the unknown parameters

μ and σ^2 can be integrated out of the full density (3.9). This yields:

$$\begin{aligned}
p(y) &= \int_0^\infty \int_{-\infty}^\infty p(y \,|\, \mu,\, \sigma^2)\, p(\mu, \sigma^2)\, d\mu\, d\sigma^2 \\
&= c \iint \sigma^{-1} (\sigma^2)^{-\left(\frac{\nu_n}{2}+1\right)} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_n \sigma_n^2 + \kappa_n (\mu - \mu_n)^2\right]\right) d\mu\, d\sigma^2 \\
&= c \int \sigma^{-1} (\sigma^2)^{-\left(\frac{\nu_n}{2}+1\right)} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_n \sigma_n^2\right]\right) \\
&\quad \times \left\{ \int \exp\left(-\frac{1}{2\sigma^2}\left[\kappa_n (\mu - \mu_n)^2\right]\right) d\mu \right\} d\sigma^2 \\
&= c \int (\sigma^2)^{-\left(\frac{\nu_n}{2}+1\right)} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_n \sigma_n^2\right]\right) \frac{\sqrt{2\pi}}{\sqrt{\kappa_n}}\, d\sigma^2. \qquad (3.13)
\end{aligned}$$

The resulting marginal likelihood for the data under the null model is therefore:

$$p(y|H_0) = \frac{\sqrt{\kappa_0}}{\sqrt{\kappa_n}} \frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} \left(\sigma_0^2\right)^{\frac{\nu_0}{2}} (2\pi)^{-\frac{n}{2}} \frac{\Gamma\left(\frac{\nu_n}{2}\right)}{\left(\frac{\nu_n}{2}\right)^{\frac{\nu_n}{2}} \left(\sigma_n^2\right)^{\frac{\nu_n}{2}}}. \qquad (3.14)$$

3.3 Alternative Model

The alternative model allows for each of the three groups to have separate means, but a common variance for each group is assumed. Similar to the null model scenario, the mean of each group is chosen to be normally distributed a priori,with the within-group variance distributed according to an Inv-χ^2 distribution. The likelihoods are again taken to be normal. That is,

$$\begin{aligned}
\mu_A \,|\, \sigma^2 &\sim \mathrm{N}\left(\mu_1, \frac{\sigma^2}{\kappa_0}\right), \quad y_{Ai} \sim \mathrm{N}(\mu_A, \sigma^2), \\
\mu_B \,|\, \sigma^2 &\sim \mathrm{N}\left(\mu_2, \frac{\sigma^2}{\kappa_0}\right), \quad y_{Bi} \sim \mathrm{N}(\mu_B, \sigma^2), \\
\mu_C \,|\, \sigma^2 &\sim \mathrm{N}\left(\mu_3, \frac{\sigma^2}{\kappa_0}\right), \quad y_{Ci} \sim \mathrm{N}(\mu_C, \sigma^2), \\
\sigma^2 &\sim \text{Inv-}\chi^2\left(\nu_0\,,\, \sigma_0^2\right), \qquad (3.15)
\end{aligned}$$

with all y's independent (conditional on their group membership), and μ_1, μ_2, μ_3, κ_0, ν_0 and σ_0^2 are hyperparameters.

The full joint distribution $p(\mu_A\,,\mu_B\,,\mu_C\,,\sigma^2\,,y)$ can be found by multiplying the priors for the means and variance, by the likelihood of the data:

$$\begin{aligned} p(\mu_A\,,\mu_B\,,\mu_C\,,\sigma^2\,,y) \;=\; & p(\mu_A\,|\,\sigma^2)p(y\,|\,\mu_A\,,\,\sigma^2)p(\mu_B\,|\,\sigma^2)p(y\,|\,\mu_B\,,\,\sigma^2) \\ & \times p(\mu_C\,|\,\sigma^2)p(y\,|\,\mu_C\,,\,\sigma^2)p(\sigma^2). \end{aligned} \tag{3.16}$$

Hence the full distribution is:

$$\begin{aligned} p(\mu_A,\,\mu_B,\,\mu_C,\,\sigma^2,\,y) = c_2 \exp & \left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2\right]\right) \\ & \times \exp\left(-\frac{1}{2\sigma^2}\left[\kappa_0(\mu_A-\mu_1)^2+\kappa_0(\mu_B-\mu_2)^2+\kappa_0(\mu_C-\mu_3)^2\right]\right) \\ & \times \exp\Big(-\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n_A}(y_i-\bar{y}_A)^2+n_A(\bar{y}_A-\mu_A)^2 \\ & \qquad +\sum_{i=1}^{n_B}(y_i-\bar{y}_B)^2+n_B(\bar{y}_B-\mu_B)^2 \\ & \qquad +\sum_{i=1}^{n_C}(y_i-\bar{y}_C)^2+n_C(\bar{y}_C-\mu_C)^2\Big]\Big) \\ & \times\sigma^{-1}\sigma^{-1}\sigma^{-1}(\sigma^2)^{-\frac{n_A}{2}}(\sigma^2)^{-\frac{n_B}{2}}(\sigma^2)^{-\frac{n_C}{2}}(\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)}, \end{aligned} \tag{3.17}$$

where c_2 is a constant that does not depend on μ_A, μ_B, μ_C or σ^2 and takes the form:

$$c_2 \;=\; \frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\,\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\,\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\,\frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\,(\sigma_0^2)^{\frac{\nu_0}{2}}\,(2\pi)^{-\frac{n_A}{2}}\,(2\pi)^{-\frac{n_B}{2}}\,(2\pi)^{-\frac{n_C}{2}}. \tag{3.18}$$

However, since $n_A+n_B+n_C=n$, equation 3.18 can be simplified as:

$$c_2 \;=\; \left(\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\right)^3\,\frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\,(\sigma_0^2)^{\frac{\nu_0}{2}}\,(2\pi)^{-\frac{n}{2}}. \tag{3.19}$$

Bringing together the arguments of the exponentials in 3.17, and multiplying out the squares, results in:

$$-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2+(n_A-1)s_1^2+\kappa_0\mu_A^2-2\kappa_0\mu_A\mu_1+\kappa_0\mu_1^2\right.$$

$$
\begin{aligned}
&+ n_A\bar{y}_A^2 - 2n_A\bar{y}_A\mu_A + n_A\mu_A^2 \\
&+ (n_B-1)s_2^2 + \kappa_0\mu_B^2 - 2\kappa_0\mu_B\mu_2 + \kappa_0\mu_2^2 \\
&+ n_B\bar{y}_B^2 - 2n_B\bar{y}_B\mu_B + n_B\mu_B^2 \\
&+ (n_C-1)s_3^2 + \kappa_0\mu_C^2 - 2\kappa_0\mu_C\mu_3 + \kappa_0\mu_3^2 \\
&+ n_C\bar{y}_C^2 - 2n_C\bar{y}_C\mu_C + n_C\mu_C^2\Big]. \qquad (3.20)
\end{aligned}
$$

Collecting together the terms in μ_A and completing the square yields:

$$
-\frac{1}{2\sigma^2}\left[(\kappa_0+n_A)\left(\mu_A - \frac{\mu_1\kappa_0 + n_A\bar{y}_A}{\kappa_0+n_A}\right)^2 - \frac{(\mu_1\kappa_0+n_A\bar{y}_A)^2}{\kappa_0+n_A}\right]. \qquad (3.21)
$$

Similar terms for μ_B and μ_C can be obtained. In a similar manner, the terms that involve the prior means μ_1, μ_2 and μ_3 can be treated separately from each other. For example, the terms involving μ_1 can be collected together with the term $n_A\bar{y}_A^2$ and a square completed, to yield:

$$
\frac{1}{2\sigma^2}\left[\frac{(\mu_1\kappa_0+n_A\bar{y}_A)^2}{\kappa_0+n_A} - \kappa_0\mu_0^2 - n_A\bar{y}_A^2\right] = -\frac{1}{2\sigma^2}\frac{n_A\kappa_0(\bar{y}_A-\mu_1)^2}{\kappa_0+n_A}. \qquad (3.22)
$$

Bringing all this together results in:

$$
\begin{aligned}
p(\mu_A,\mu_B,\mu_C,\sigma^2,y) = c_2\ &\sigma^{(-5-n_A-n_B-n_C-\nu_0)} \\
&\times \exp\Bigg(-\frac{1}{2\sigma^2}\Bigg[\nu_0\sigma_0^2 + (n_A-1)s_1^2 + (n_B-1)s_2^2 + (n_C-1)s_3^2 \\
&+ \frac{n_A\kappa_0(\bar{y}_A-\mu_1)^2}{\kappa_0+n_A} + \frac{n_B\kappa_0(\bar{y}_B-\mu_2)^2}{\kappa_0+n_B} + \frac{n_C\kappa_0(\bar{y}_C-\mu_3)^2}{\kappa_0+n_C} \\
&+ (\kappa_0+n_A)\left(\mu_A - \frac{\mu_1\kappa_0+n_A\bar{y}_A}{\kappa_0+n_A}\right)^2 \\
&+ (\kappa_0+n_B)\left(\mu_B - \frac{\mu_2\kappa_0+n_B\bar{y}_B}{\kappa_0+n_B}\right)^2 \\
&+ (\kappa_0+n_C)\left(\mu_C - \frac{\mu_3\kappa_0+n_C\bar{y}_C}{\kappa_0+n_C}\right)^2\Bigg]\Bigg). \qquad (3.23)
\end{aligned}
$$

If we define:

$$\begin{aligned}
\mu_{NA} &= \frac{\mu_1\kappa_0 + n_A\bar{y}_A}{n_A + \kappa_0}, & \kappa_{NA} &= \kappa_0 + n_A,\\
\mu_{NB} &= \frac{\mu_2\kappa_0 + n_B\bar{y}_B}{n_B + \kappa_0}, & \kappa_{NB} &= \kappa_0 + n_B,\\
\mu_{NC} &= \frac{\mu_3\kappa_0 + n_C\bar{y}_C}{n_C + \kappa_0}, & \kappa_{NC} &= \kappa_0 + n_C,\\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n_A-1)s_1^2 + (n_B-1)s_2^2 + (n_C-1)s_3^2\\
&\quad + \frac{n_A\kappa_0(\bar{y}_A-\mu_1)^2}{\kappa_0+n_A} + \frac{n_B\kappa_0(\bar{y}_B-\mu_2)^2}{\kappa_0+n_B} + \frac{n_C\kappa_0(\bar{y}_C-\mu_3)^2}{\kappa_0+n_C},
\end{aligned}$$

then the full joint distribution can be written as:

$$\begin{aligned}
p(\mu_A,\mu_B,\mu_C,\sigma^2,y) =& c_2\ \sigma^{(-5-n_A-n_B-n_C-\nu_0)}\\
&\times\exp\Big(-\frac{1}{2\sigma^2}\Big[\nu_n\sigma_n^2 + \kappa_{NA}(\mu_A-\mu_{NA})^2 + \kappa_{NB}(\mu_B-\mu_{NB})^2\\
&\quad + \kappa_{NC}(\mu_C-\mu_{NC})^2\Big]\Big). \qquad (3.24)
\end{aligned}$$

Exploiting the normality of this expression in μ_A, μ_B and μ_C, and the inverse chi-squared form in σ^2, integration of (3.24) with respect to μ_A, μ_B, μ_C and σ^2 gives the following marginal likelihood:

$$\begin{aligned}
p(y|H_{\text{alt}}) =& \left(\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\right)^3 \frac{(\frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}\,(\sigma_0^2)^{\frac{\nu_0}{2}}\,(2\pi)^{-\frac{n}{2}}\\
&\times\frac{\sqrt{2\pi}}{\sqrt{\kappa_{NA}}}\,\frac{\sqrt{2\pi}}{\sqrt{\kappa_{NB}}}\,\frac{\sqrt{2\pi}}{\sqrt{\kappa_{NC}}}\,\frac{\Gamma(\frac{\nu_n}{2})}{(\frac{\nu_n}{2})^{\frac{\nu_n}{2}}(\sigma_n^2)^{\frac{\nu_n}{2}}}. \qquad (3.25)
\end{aligned}$$

Therefore the Bayes factor representing the evidence in favour of the alternative model (3.25) compared to the null model (3.14) is:

$$\frac{\left(\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\right)^3 \frac{(\frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}\,(\sigma_0^2)^{\frac{\nu_0}{2}}\,(2\pi)^{-\frac{n}{2}}\,\frac{\sqrt{2\pi}}{\sqrt{\kappa_{NA}}}\,\frac{\sqrt{2\pi}}{\sqrt{\kappa_{NB}}}\,\frac{\sqrt{2\pi}}{\sqrt{\kappa_{NC}}}\,\frac{\Gamma(\frac{\nu_n}{2})}{(\frac{\nu_n}{2})^{\frac{\nu_n}{2}}\sigma_{\text{alt}}^{\nu_n}}}{\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\,\frac{(\frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}\,(\sigma_0^2)^{\frac{\nu_0}{2}}\,(2\pi)^{-\frac{n}{2}}\,\frac{\sqrt{2\pi}}{\sqrt{\kappa_N}}\,\frac{\Gamma(\frac{\nu_n}{2})}{(\frac{\nu_n}{2})^{\frac{\nu_n}{2}}\sigma_{\text{null}}^{\nu_n}}}, \qquad (3.26)$$

where σ_{alt} and σ_{null} are the σ_n for the alternative and null models respectively.

Hence,

$$\frac{p(y|H_{\text{alt}})}{p(y|H_{\text{null}})} = \left(\frac{\kappa_0}{\sqrt{\kappa_{NA}}\sqrt{\kappa_{NB}}\sqrt{\kappa_{NC}}} \frac{1}{\sigma_{\text{alt}}^{\nu_n}} \right) \Bigg/ \left(\frac{1}{\sqrt{\kappa_N}} \frac{1}{\sigma_{\text{null}}^{\nu_n}} \right). \tag{3.27}$$

3.4 Simple Additive Model

An additive model, for the one split or SNP being considered, can be considered in addition to the general model that allows flexibility in having separate means for each group. The following priors are assigned for the unknown means of the homozygous genotype classes:

$$\begin{aligned} \mu_A \,|\, \sigma^2 \; &\sim \mathrm{N}\left(\mu_1, \frac{\sigma^2}{\kappa_0} \right), \\ \mu_B \,|\, \sigma^2 \; &\sim \mathrm{N}\left(\mu_2, \frac{\sigma^2}{\kappa_0} \right), \end{aligned} \tag{3.28}$$

and the corresponding likelihoods for an individual from the three possible groups are:

$$\begin{aligned} y_A \; &\sim \mathrm{N}\left(\mu_A, \sigma^2 \right), \\ y_{AB} \; &\sim \mathrm{N}\left(\tfrac{1}{2}(\mu_A + \mu_B), \sigma^2 \right), \\ y_B \; &\sim \mathrm{N}\left(\mu_B, \sigma^2 \right), \end{aligned} \tag{3.29}$$

where the heterozygote mean is half-way between the homozygote means. Thus the full joint distribution is:

$$\begin{aligned} p(y,\, \mu_A,\, \mu_B,\, \sigma^2) = c_3 \exp &\left(-\frac{1}{2\sigma^2} \left[\nu_0 \sigma_0^2 \right] \right) \\ &\times \exp \left(-\frac{1}{2\sigma^2} \left[\kappa_0 (\mu_A - \mu_1)^2 + \kappa_0 (\mu_B - \mu_2)^2 \right] \right) \\ &\times \exp \left(-\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n_A} (y_i - \bar{y}_A)^2 + n_A (\bar{y}_A - \mu_A)^2 \right. \right. \\ &\qquad \left. \left. + \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})^2 + n_{AB} \left(\bar{y}_{AB} - \tfrac{1}{2}(\mu_A + \mu_B) \right)^2 \right. \right. \end{aligned}$$

$$+\sum_{i=1}^{n_B}(y_i-\bar{y}_B)^2+n_B(\bar{y}_B-\mu_B)^2\Big]\Bigg)$$
$$\times\,\sigma^{-1}\sigma^{-1}(\sigma^2)^{-\frac{n}{2}}(\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)}, \tag{3.30}$$

where n represents the total sample size and c_3 is a constant that does not depend on μ_A, μ_B, μ_C or σ^2, taking the form:

$$c_3 \;=\; \left(\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\right)^2 \frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\left(\sigma_0^2\right)^{\frac{\nu_0}{2}}(2\pi)^{-\frac{n}{2}}. \tag{3.31}$$

Multiplying out the arguments to the exponentials leads to:

$$\begin{aligned} l \equiv -2\sigma^2\ln p = {} & -2\sigma^2\ln c_3 - 2\sigma^2\ln\left(\sigma^{-\mathrm{n}-\nu_0-4}\right)+\nu_0\sigma_0^2 \\ & +\left(\kappa_0+n_A+\tfrac{1}{4}n_{AB}\right)\mu_A^2+2\left(\tfrac{1}{4}n_{AB}\right)\mu_A\mu_B+\left(\kappa_0+n_B+\tfrac{1}{4}n_{AB}\right)\mu_B^2 \\ & -2\left(\kappa_0\mu_1+n_A\bar{y}_A+\tfrac{1}{2}n_{AB}\bar{y}_{AB}\right)\mu_A-2\left(\kappa_0\mu_2+n_B\bar{y}_B+\tfrac{1}{2}n_{AB}\bar{y}_{AB}\right)\mu_B \\ & +\sum_{i=1}^{n_A}(y_i-\bar{y}_A)^2+\kappa_0\mu_1^2+n_A\bar{y}_A^2+\sum_{i=1}^{n_{AB}}(y_i-\bar{y}_{AB})^2+n_{AB}\bar{y}_{AB}^2 \\ & +\sum_{i=1}^{n_B}(y_i-\bar{y}_B)^2+\kappa_0\mu_2^2+n_B\bar{y}_B^2. \end{aligned} \tag{3.32}$$

This is a quadratic form in μ_A and μ_B, which implies that the density is bivariate normal in these parameters, the mean of which is identical to the mode. The mean can therefore be found by identifying the minimum of $-2\sigma^2\ln p$. First the partial derivative, $\frac{\partial l}{\partial \mu_A}$, is obtained:

$$\frac{\partial l}{\partial \mu_A} = 2\mu_A\left(\kappa_0+n_A+\tfrac{1}{4}n_{AB}\right)+2\mu_B\left(\tfrac{1}{4}n_{AB}\right)-2\kappa_0\mu_1-2n_A\bar{y}_A-n_{AB}\bar{y}_{AB}. \tag{3.33}$$

Equating this to zero at $(\hat{\mu}_A,\hat{\mu}_B)$ gives :

$$2\hat{\mu}_A\left(\kappa_0+n_A+\tfrac{1}{4}n_{AB}\right)=2\kappa_0\mu_1+2n_A\bar{y}_A+n_{AB}\bar{y}_{AB}-2\hat{\mu}_B\left(\tfrac{1}{4}n_{AB}\right). \tag{3.34}$$

Solving for $\hat{\mu}_A$:

$$\hat{\mu}_A = \frac{\kappa_0\mu_1 + n_A\bar{y}_A + \frac{1}{2}n_{AB}\bar{y}_{AB} - \frac{1}{4}\hat{\mu}_B n_{AB}}{\kappa_0 + n_A + \frac{1}{4}n_{AB}}. \tag{3.35}$$

Similarly, the partial derivative ,$\frac{\partial l}{\partial \mu_B}$, set to zero results in the following estimate for μ_B:

$$\hat{\mu}_B = \frac{\kappa_0\mu_2 + n_B\bar{y}_B + \frac{1}{2}n_{AB}\bar{y}_{AB} - \frac{1}{4}\hat{\mu}_A n_{AB}}{\kappa_0 + n_B + \frac{1}{4}n_{AB}}. \tag{3.36}$$

Solving these two simultaneous equations for $\hat{\mu}_A$ and $\hat{\mu}_B$ gives the estimator $\hat{\mu}_A$ of μ_A:

$$\hat{\mu}_A = \frac{(\kappa_0 + n_B + \frac{1}{4}n_{AB})(\kappa_0\mu_1 + n_A\bar{y}_A + \frac{1}{2}n_{AB}\bar{y}_{AB}) - \frac{1}{4}n_{AB}(\kappa_0\mu_2 + n_B\bar{y}_B + \frac{1}{2}n_{AB}\bar{y}_{AB})}{(\kappa_0 + n_A + \frac{1}{4}n_{AB})(\kappa_0 + n_B + \frac{1}{4}n_{AB}) - \frac{1}{16}n_{AB}^2}. \tag{3.37}$$

Similarly the estimator $\hat{\mu}_B$ of μ_B is:

$$\hat{\mu}_B = \frac{(\kappa_0 + n_A + \frac{1}{4}n_{AB})(\kappa_0\mu_2 + n_B\bar{y}_B + \frac{1}{2}n_{AB}\bar{y}_{AB}) - \frac{1}{4}n_{AB}(\kappa_0\mu_1 + n_A\bar{y}_A + \frac{1}{2}n_{AB}\bar{y}_{AB})}{(\kappa_0 + n_A + \frac{1}{4}n_{AB})(\kappa_0 + n_B + \frac{1}{4}n_{AB}) - \frac{1}{16}n_{AB}^2}. \tag{3.38}$$

The standard form of a bivariate normal distribution is:

$$f(\mu_A, \mu_B) = \frac{1}{2\pi\sqrt{\det(\mathbf{Q})}}\exp\left\{-\frac{1}{2}(\mu_A - \hat{\mu}_A, \mu_B - \hat{\mu}_B)\mathbf{Q}^{-1}(\mu_A - \hat{\mu}_A, \mu_B - \hat{\mu}_B)^T\right\}, \tag{3.39}$$

where $\mathbf{Q}$ is the positive semi-definite and symmetric covariance matrix . Therefore, comparing (3.32) to (3.39), the elements of $\mathbf{Q}^{-1}$ can be identified:

$$\mathbf{Q}^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} \kappa_0 + n_A + \frac{1}{4}n_{AB} & \frac{1}{4}n_{AB} \\ \frac{1}{4}n_{AB} & \kappa_0 + n_B + \frac{1}{4}n_{AB} \end{pmatrix}. \tag{3.40}$$

The remaining terms in (3.32) are $-2\sigma^2\ln c_3$, $-2\sigma^2\ln\left(\sigma^{-n-\nu_0-4}\right)$ and:

$$\nu_n\sigma_n^2 \equiv \nu_0\sigma_0^2 + \sum_{i=1}^{n_A}(y_i - \bar{y}_A)^2 + \sum_{i=1}^{n_{AB}}(y_i - \bar{y}_{AB})^2 + \sum_{i=1}^{n_B}(y_i - \bar{y}_B)^2$$

$$
\begin{aligned}
&+ k_0\mu_1^2 + n_A\bar{y}_A^2 + k_0\mu_2^2 + n_B\bar{y}_B^2 + n_{AB}\bar{y}_{AB}^2 \\
&- (\kappa_0 + n_A + \tfrac{1}{4}n_{AB})\hat{\mu}_A^2 - (\kappa_0 + n_B + \tfrac{1}{4}n_{AB})\hat{\mu}_B^2 \\
&- \tfrac{1}{2}n_{AB}\hat{\mu}_A\hat{\mu}_B, \qquad (3.41)
\end{aligned}
$$

where ν_n is given by

$$
\begin{aligned}
\nu_n &= \nu_0 + n_A + n_{AB} + n_B \\
&= \nu_0 + n. \qquad (3.42)
\end{aligned}
$$

Integration of (3.32) with respect to μ_A, μ_B and σ^2, exploiting the bivariate normal form in μ_A and μ_B, yields:

$$
\begin{aligned}
p(y \mid \mathrm{H_{add}}) = \frac{\kappa_0}{2\pi}\frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\left(\sigma_0^2\right)^{\frac{\nu_0}{2}}(2\pi)^{-\frac{n}{2}} \\
\times \frac{2\pi}{\sqrt{\left(\kappa_0 + n_A + \frac{1}{4}n_{AB}\right)\left(\kappa_0 + n_B + \frac{1}{4}n_{AB}\right) - \frac{1}{16}n_{AB}^2}} \\
\times \frac{\Gamma\left(\frac{\nu_n}{2}\right)}{\left(\frac{\nu_n}{2}\right)^{\frac{\nu_n}{2}}\sigma_{\mathrm{add}}^{\nu_n}}. \qquad (3.43)
\end{aligned}
$$

Therefore the Bayes factor for comparing the additive model to the null model is:

$$
\begin{aligned}
\frac{p(y|H_{\mathrm{add}})}{p(y|H_{\mathrm{null}})} = \left(\frac{\sqrt{\kappa_0}}{\sqrt{\left(\kappa_0 + n_A + \frac{1}{4}n_{AB}\right)\left(\kappa_0 + n_B + \frac{1}{4}n_{AB}\right) - \frac{1}{16}n_{AB}^2}}\frac{1}{(\sigma_{\mathrm{add}})^{\nu_n}}\right) \\
\div \left(\frac{1}{\sqrt{\kappa_n}}\frac{1}{(\sigma_{\mathrm{null}})^{\nu_n}}\right). \qquad (3.44)
\end{aligned}
$$

3.5 Dominant and Recessive Models

In the first round of association testing, dominant and recessive models can be tested using the Bayes factors. The specification of both forms of models is essentially the same as that for a general model, but where the phenotypes of the heterozygote grouping (AB) are allocated to either of the homozygote groupings

(AA or BB). Table 3.1 illustrates the definitions of dominant and recessive mutations that are used, where AA.AB represents the combined phenotypes of groups AA and AB, and similarly BB.AB represents the combined phenotypes of groups BB and AB. As a dominant model for one allele is the same as a recessive model for the other allele, it has been chosen to refer to a model as dominant or recessive according to the group containing a higher phenotype mean.

Table 3.1: Dominant and recessive allocation.

Criterion 1	Criterion 2	Conclusion
$\mathrm{BF(AA.AB)} > \mathrm{BF(BB.AB)}$	$\overline{\mathrm{BB.AB}} > \overline{\mathrm{AA.AB}}$	Recessive in B
$\mathrm{BF(AA.AB)} \leq \mathrm{BF(BB.AB)}$	$\overline{\mathrm{BB.AB}} > \overline{\mathrm{AA.AB}}$	Dominant in B
$\mathrm{BF(AA.AB)} > \mathrm{BF(BB.AB)}$	$\overline{\mathrm{BB.AB}} \leq \overline{\mathrm{AA.AB}}$	Dominant in A
$\mathrm{BF(AA.AB)} \leq \mathrm{BF(BB.AB)}$	$\overline{\mathrm{BB.AB}} \leq \overline{\mathrm{AA.AB}}$	Recessive in A

If group A is being considered, the marginal likelihood resulting from the recessive model is simply as in equation 3.45. This marginal likelihood is equivalent to that for the dominant model of group B. That is,

$$p(y \,|\, \mathrm{H}_{\mathrm{rec_A}}) = \left(\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\right)^2 \frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} \left(\sigma_0^2\right)^{\frac{\nu_0}{2}} (2\pi)^{-\frac{n}{2}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{\mathrm{NAA}}}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{\mathrm{NAB.BB}}}} \frac{\Gamma\left(\frac{\nu_\mathrm{n}}{2}\right)}{\left(\frac{\nu_\mathrm{n}}{2}\right)^{\frac{\nu_\mathrm{n}}{2}} \sigma_{\mathrm{rec_A}}^{\nu_\mathrm{n}}}. \tag{3.45}$$

This is derived by a very similar argument to that which led to expression (3.25), with the difference being that two of the genotypes can be grouped together as they are phenotypically indistinguishable in a dominant or a recessive model. In an analogous manner, the marginal likelihood resulting for the dominant model in A (or the recessive model for group B), can be found:

$$p(y \,|\, \mathrm{H}_{\mathrm{dom_A}}) = \left(\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\right)^2 \frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} \left(\sigma_0^2\right)^{\frac{\nu_0}{2}} (2\pi)^{-\frac{n}{2}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{\mathrm{NAA.AB}}}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{\mathrm{NBB}}}} \frac{\Gamma\left(\frac{\nu_\mathrm{n}}{2}\right)}{\left(\frac{\nu_\mathrm{n}}{2}\right)^{\frac{\nu_\mathrm{n}}{2}} \sigma_{\mathrm{dom_A}}^{\nu_\mathrm{n}}}. \tag{3.46}$$

Information regarding the dominant or additive nature of the first-round split is

stored, and used if required in the more complex second-round splits of section 3.8.

3.6 Alternative Model II

As the Bayes factors for different branches are correlated, some branches will be declared as significant as a result of being close to a branch that carries a true causative mutation. The same is also true if the SNPs are to be assessed individually, ignoring the tree structure in the analysis. In order to compensate for this, a further model can be proposed that can assess whether a branch is declared significant, conditional on the split in data determined from another branch.

The most basic model that could represent this conditional stage would be to have a model with six groups representing the six combinations that could arise. A Bayes factor can then be determined to assess the evidence in favour of the six-group model, compared to that of the previous alternative of three groups. The marginal likelihood of this model can be calculated in an analogous way to that of the three group model, and this is implemented for all non-empty groups that are present at a stage of splitting:

$$\begin{aligned} p(y \,|\, \mathrm{H_{AB1B2}}) = & \left(\frac{\sqrt{\kappa_0}}{\sqrt{2\pi}}\right)^6 \frac{\left(\frac{\nu_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})} \left(\sigma_0^2\right)^{\frac{\nu_0}{2}} (2\pi)^{-\frac{n}{2}} \\ & \times \frac{\sqrt{2\pi}}{\sqrt{\kappa_{NAA}}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{NAB1}}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{NAB2}}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{NB1B1}}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{NB2B2}}} \frac{\sqrt{2\pi}}{\sqrt{\kappa_{NB1B2}}} \\ & \times \frac{\Gamma(\frac{\nu_n}{2})}{\left(\frac{\nu_n}{2}\right)^{\frac{\nu_n}{2}} \sigma_{AB1B2}^{\nu_n}}, \end{aligned} \tag{3.47}$$

where class B has been split into B1 and B2 at the second round (see figure 1.3). This will result in a Bayes factor comparing the six-group model to that of the

three-group model, which simplifies down to:

$$\frac{p(y\,|\,\mathrm{H_{AB1B2}})}{p(y\,|\,\mathrm{H_{gen}})} = \left(\frac{\kappa_0^{\frac{3}{2}}}{\sqrt{\kappa_{NAA}}\sqrt{\kappa_{NAB1}}\sqrt{\kappa_{NAB2}}\sqrt{\kappa_{NB1B1}}\sqrt{\kappa_{NB2B2}}\sqrt{\kappa_{NB2B2}}} \times \frac{1}{(\sigma_{alt})^{\nu_n}} \right) \div \left(\frac{1}{\sqrt{\kappa_{nA}}\sqrt{\kappa_{nB}}\sqrt{\kappa_{nC}}} \times \frac{1}{(\sigma_{gen})^{\nu_n}} \right) \tag{3.48}$$

In a similar manner, the comparison can be made between the additive or the dominant/recessive model, according to which model performed best in the first round of tests.

3.7 Complex Additive Model

A model for second or further stage splits can be calculated, if all the splits involved are determined from the first round as having acted in an additive manner. There is an option of either introducing a new mean for the cross-term at the second-stage split, or assuming that the effect of the cross-term is an additive effect of both groups. The second interpretation will be discussed here, as the approach of adding a new mean for the cross-term will be dealt with in section 3.8. In a similar approach to that used for the simple additive model of section 3.4, each homozygous grouping can be given a Normal prior for the unknown mean. The likelihood for the heterozygous groupings consisting of the cross terms of those groups involved. Solving the system of equations for a second-stage split will result in the following estimators of the means for the three possible groupings involved.

$$\hat{\mu}_A = \frac{\kappa_0\mu_1 + n_A\bar{y}_A + \frac{1}{2}n_{AB}\bar{y}_{AB} + \frac{1}{2}n_{AC}\bar{y}_{AC} - \frac{1}{4}n_{AB}\hat{\mu}_B - \frac{1}{4}n_{AC}\hat{\mu}_C}{\kappa_0 + n_A + \frac{1}{4}n_{AB} + \frac{1}{4}n_{AC}},$$

$$\hat{\mu}_B = \frac{\kappa_0\mu_1 + n_B\bar{y}_B + \frac{1}{2}n_{AB}\bar{y}_{AB} + \frac{1}{2}n_{BC}\bar{y}_{BC} - \frac{1}{4}n_{AB}\hat{\mu}_A - \frac{1}{4}n_{BC}\hat{\mu}_C}{\kappa_0 + n_B + \frac{1}{4}n_{AB} + \frac{1}{4}n_{BC}},$$

$$\hat{\mu}_C = \frac{\kappa_0\mu_1 + n_C\bar{y}_C + \frac{1}{2}n_{AC}\bar{y}_{AC} + \frac{1}{2}n_{BC}\bar{y}_{BC} - \frac{1}{4}n_{AC}\hat{\mu}_B - \frac{1}{4}n_{BC}\hat{\mu}_A}{\kappa_0 + n_C + \frac{1}{4}n_{AC} + \frac{1}{4}n_{BC}}. \tag{3.49}$$

An example of a second round of testing has been given , however but the pattern can be easily extended for further levels of splits as required. These simultaneous equations can be solved for $\hat{\mu}_A$, $\hat{\mu}_B$ and $\hat{\mu}_C$ using matrix algebra, by finding the solution for $\mathbf{X}$ of the equation $\mathbf{AX} = \mathbf{b}$, where $\mathbf{X} = (\hat{\mu}_A, \hat{\mu}_B, \hat{\mu}_C)^{\mathbf{T}}$. This can be done through the use of the *solve* function in R, where the relevant matrices required for a two-split model are of the form:

$$\mathbf{b} = \begin{pmatrix} \frac{\kappa_0\mu_1 + n_A\bar{y}_A + \frac{1}{2}n_{AB}\bar{y}_{AB} + \frac{1}{2}n_{AC}\bar{y}_{AC}}{\kappa_0 + n_A + \frac{1}{4}n_{AB} + \frac{1}{4}n_{AC}} \\ \frac{\kappa_0\mu_2 + n_B\bar{y}_B + \frac{1}{2}n_{AB}\bar{y}_{AB} + \frac{1}{2}n_{BC}\bar{y}_{BC}}{\kappa_0 + n_B + \frac{1}{4}n_{AB} + \frac{1}{4}n_{BC}} \\ \frac{\kappa_0\mu_3 + n_C\bar{y}_C + \frac{1}{2}n_{AC}\bar{y}_{AC} + \frac{1}{2}n_{BC}\bar{y}_{BC}}{\kappa_0 + n_C + \frac{1}{4}n_{AC} + \frac{1}{4}n_{BC}} \end{pmatrix} \tag{3.50}$$

and

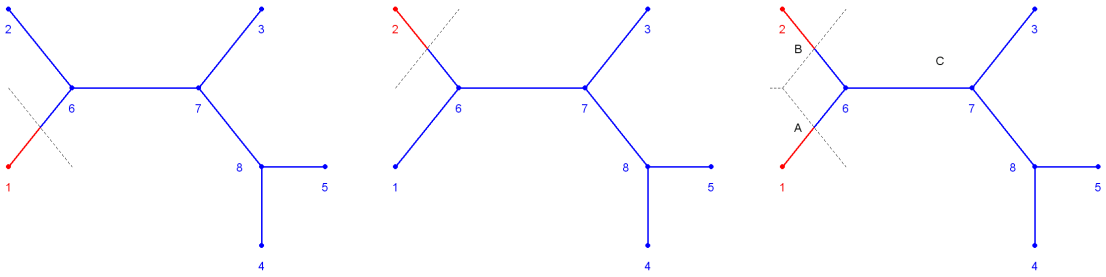
$$\mathbf{A} = \begin{pmatrix} 0 & -\frac{n_{AB}}{\kappa_0 + n_A + \frac{1}{4}n_{AB} + \frac{1}{4}n_{AC}} & -\frac{n_{AC}}{\kappa_0 + n_A + \frac{1}{4}n_{AB} + \frac{1}{4}n_{AC}} \\ -\frac{n_{AB}}{\kappa_0 + n_B + \frac{1}{4}n_{AB} + \frac{1}{4}n_{BC}} & 0 & -\frac{n_{BC}}{\kappa_0 + n_B + \frac{1}{4}n_{AB} + \frac{1}{4}n_{BC}} \\ -\frac{n_{AC}}{\kappa_0 + n_C + \frac{1}{4}n_{AC} + \frac{1}{4}n_{BC}} & -\frac{n_{BC}}{\kappa_0 + n_C + \frac{1}{4}n_{AC} + \frac{1}{4}n_{BC}} & 0 \end{pmatrix}. \tag{3.51}$$

To obtain a Bayes factor for the complex additive model, the marginal likelihood can be obtained by integrating out the parameters in an analogous way to the simple additive model, and compared to the highest marginal likelihood of the models involved in the previous round of comparisons.

3.8 Complex Mixture Model

This section details how at second or higher-order splits, an alternative likelihood is calculated for groupings that have been found to be mixtures of additive, dominant, and recessive in the first round of tests. This ensures that fair comparisons are made between each level of splits, by taking into account all the information present in the data.

In order to retain information about the mutation mechanisms involved from



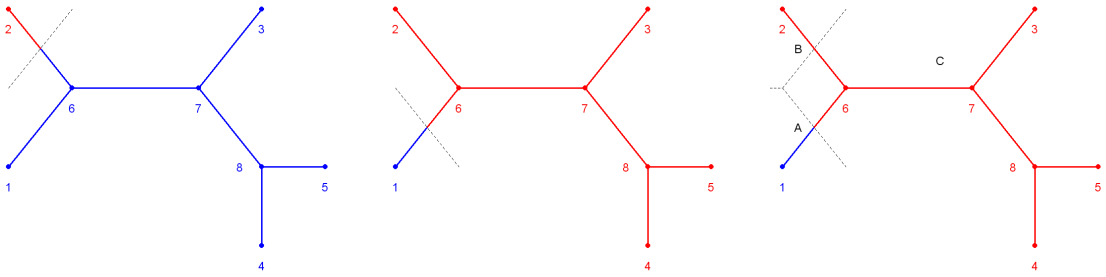
(a) First round split(1) (b) First round split (2) (c) Second round split

(d) Dominant splits (e) A additive with C; B dominant over C

Figure 3.1: Significant splits (red) that have been found in the first round of the procedure (a,b) and the resultant 'strong' (red) and 'weak' (blue) groupings to be defined in the second round of tests (c). Figures (d) and (e) represent the second-stage groupings of phenotypes, according to two possible combinations of first round mutation models.

the first round of splits, a method has been developed that relies upon a concept of 'strong' and 'weak' groupings of SNPs or haplotypes. In order for a grouping to be considered 'strong', there must be a significant effect for that group with regards to an increase in a phenotype measurement, as determined by a p-value of less than 0.05 or a Bayes Factor of greater than 150. Equivalently, a grouping is considered 'weak' if it is a group that contains the lowest phenotype mean from the initial split considered.

Figure 3.1(c) illustrates the first possible combination of strong and weak splits, as determined from a first round of tests (figures 3.1(a) and 3.1(b)). In this instance, both the split at branches *1-6* and *2-6* have been determined to be associated with the phenotype, with the 'strong' effect being apparent in both the A and B groups of haplotypes. This situation could arise if there is a baseline (wild-type) group where there is no apparent difference in phenotype scores, but on two separate branches a mutation occurs that results in an increased phenotype measurement. The illustrations in 3.1(d) and 3.1(e) shows the groupings of haplotypes that would be obtained for two possible combinations of additive and



(a) First round split (1) (b) First round split (2) (c) Second round split

BB AB | CC AC | AA | BC

(d) Dominant splits

AA | AC | CC | BB AB | BC

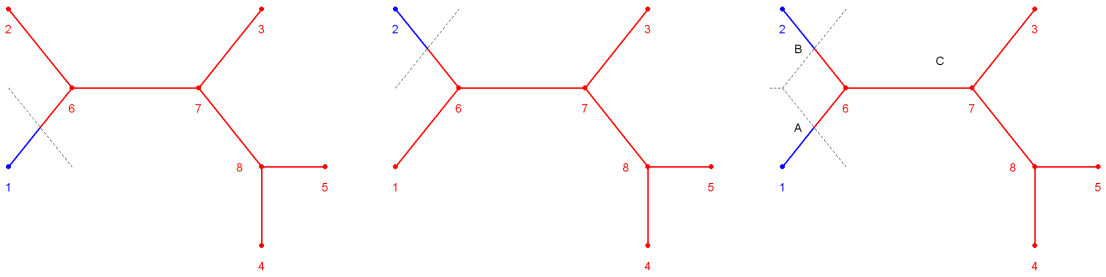
(e) C additive with A; B dominant over A

Figure 3.2: Significant splits (red) that have been found in the first round of the procedure (a,b) and the resultant 'strong' (red) and 'weak' (blue) groupings to be defined in the second round of tests (c). Figures (d) and (e) represent the second-stage groupings of phenotypes, according to two possible combinations of first round mutation models.

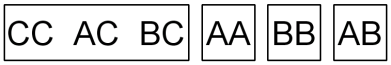
dominant mutations.

A further combination of strong and weak groupings is given in figure 3.2. This scenario could arise due to two causative mutations occurring on the same lineage, having been derived from a wild-type group. The concept of weak and strong groupings can once again be used to determine the clustering of groups. However, in this case the sub-tree involving nodes {3,4,5,6,7,8} has been declared as strong in one scenario, and weak in the other. The groupings that result from such conditions are given for two scenarios of additive and dominant mutations in figures 3.2(d) and 3.2(e).

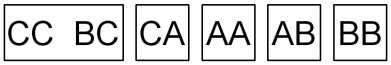
The final combination of strong and weak classes can be shown in figure 3.3. This situation is the reverse of the first scenario, in that here both haplotypes 1 and 2 are declared as strong in one of the first round splits, but weak in the other. This situation would be likely to appear if mutations result in a decrease, as opposed to an increase, of some phenotypic score. As with the previous examples, the groupings that could result for selected scenarios can be illustrated for second-round mutations following the pattern of figure 3.3(c).



(a) First round split (1) (b) First round split (2) (c) Second round split



(d) Dominant splits



(e) C additive with A; C dominant over B

Figure 3.3: Significant splits (red) that have been found in the first round of the procedure (a,b) and the resultant 'strong' (red) and 'weak' (blue) groupings to be defined in the second round of tests (c). Figures (d) and (e) represent the second-stage groupings of phenotypes, according to two possible combinations of first round mutation models.

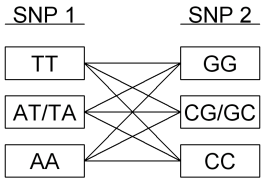


Figure 3.4: Second stage SNP groups, with homoplasy being present.

For single SNP methods, the number of groupings can be different in comparison to the Treescan approach at the second stage of splitting due to the presence of homoplasy. For example, if there were the choice of bases {A,T} at the first SNP, and at the second site there existed the sites {C,G}, then the nine possible combinations of SNPs are as given in figure 3.4. It should be noted that the method assumes that there are only two possible bases at each site, as is the case for all the simulated and real data available. The approach could however be extended to accommodate sites with three or four observable bases, although this situation is less likely to occur in real data sets.

However, if for the same two SNPs homoplasy was not present, then a situation analogous to that of the tree-based methods can be implemented whereby only six possible combinations of bases could occur, as shown in table 3.5.

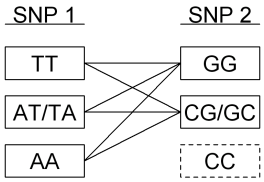


Figure 3.5: Second stage SNP groups, with homoplasy not being present.

3.9 Covariates

Although not included in the PheGe-Find application, covariates can also be straightforwardly added to the Bayes factor methods. This could then be useful for including non-genetic components into the Bayes factors, so that the genetic and environmental effects can be assessed together. The simplest case of adding one covariate to a single genetic group is briefly introduced.

The mean phenotype value is written as $\mu_A + \alpha_A x$, where μ_A is the intercept, x is the covariate and α_A is the slope parameter. Both μ_A and α_A are assigned independent normal priors:

$$\begin{aligned} \mu_A \,|\, \sigma^2 &\sim \mathrm{N}\left(\mu_1, \frac{\sigma^2}{\kappa_0}\right), \\ \alpha_A \,|\, \sigma^2 &\sim \mathrm{N}\left(\alpha_1, \frac{\sigma^2}{\kappa_0}\right). \end{aligned} \tag{3.52}$$

In this situation a common variance has been chosen for simplicity, with the prior again being chosen as an inverse-chi-squared distribution. However, models for unequal variances can also be introduced if required. It is also natural to set α_1 as zero in this context, as this would correspond to the prior belief that there is no covariate effect. The likelihood for a single observation is given by:

$$y_A \sim \mathrm{N}\left(\mu_A + (\alpha_A x), \sigma^2\right). \tag{3.53}$$

Assuming independence of the observations, the full joint distribution can be obtained in a similar manner as to the previous models that have been discussed. The marginal likelihood of the covariate model can be obtained, and a Bayes factor calculated for comparing this to an appropriate model not involving the covariate. Further details of the 'Bayesian regression' approach, and the accompanying issues of model selection, can be found in Gelman et al. (2004).

Chapter 4

PheGe-Find

PheGe-Find is a program that has been written in R version 2.4.1 with the aim being to implement various methods that can be applied for fine-scale association studies. As with PheGe-Sim the program is contained within the Rpanel (Bowman et al., 2007) environment to allow for easy specification of the various input options. Figure 4.1 illustrates a screen shot of the application with the various input options that are possible.

PheGe-Find allows for the input of files containing the genotypes and phenotypes required for the association testing. These will be checked against each other to ensure that they match, and the program will subsequently run using only individuals with both genotype and phenotype measurements.

The methods of association implemented in PheGe-Find are the same as those used upon simulated data in the PheGe-Sim application of Chapter 2. This is as a result of the requirement that the detection of causative SNPs must be treated independently from the simulation. As such none of the information relating to the simulations can be used, aside from the genotypes and corresponding phenotypes in the sample. The following section details the input options that are specific to the use of real data in PheGe-Find. The input options for the methods of association having previously been discussed in section 2.1.

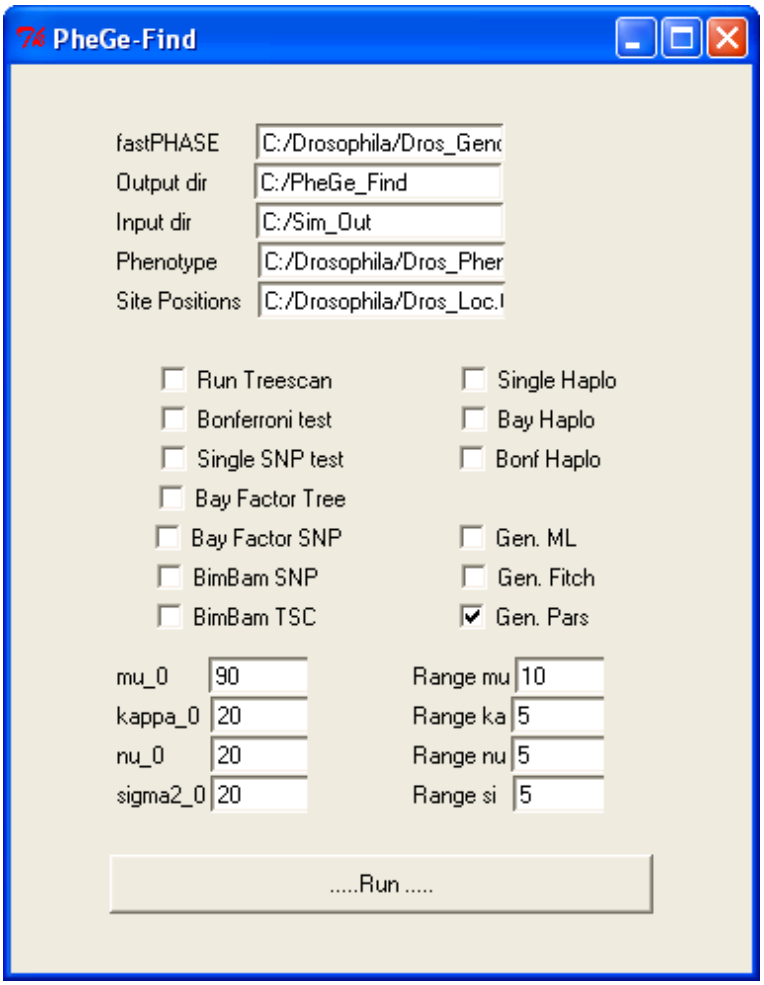


Figure 4.1: PheGe-Find screen shot.

4.1 Input Options

- ***fastphase: (String))*** = The location of the output file from fastPHASE (Scheet and Stephens, 2006) with the phased haplotypes of the genotype data. PED, or BIMBAM format, files can also be used.
- ***Output dir : (String)*** = Location of the directory where output files and plots are to be produced.
- ***Input dir : (String)*** = The location of the folder containing the external applications that are required, namely Treescan, PHYLIP, BimBam and

Haploview.

- ***Phenotype : (String)*** = Location of the file with Phenotype scores (section A.8).
- ***Site Positions : (String)*** = (Optional) Location of the file with SNP locations (section A.9).
- ***mu_0*: $(-\infty, \infty)$,** $[\mathbf{D}] = \mu_0$ hyperparameter for the Bayes factor approach (chapter 3).
- ***kappa_0* : $(\mathbf{0}, \infty)$** $= \kappa_0$ hyperparameter for the Bayes factor approach (chapter 3).
- ***nu_0* : $(\mathbf{4}, \infty)$** $= \nu_0$ hyperparameter for the Bayes factor approach (chapter 3).
- ***sigma2_0* : $(\mathbf{0}, \infty)$,** $[\mathbf{D}] = \sigma_0^2$ hyperparameter for the Bayes factor approach (chapter 3).
- ***Range options* : $(\mathbf{0}, \infty)$** = The range of values of the hyperparameters to be assessed in the sensitivity analysis plots (section 7.2.5).

4.2 Reconstruction of a Tree Based upon the Sequences

Inputs:

Sequences of SNPs in Phylip format

Although the methods involved in reconstructing phylogenetic trees are not of primary importance in this thesis as the trees are primarily considered nuisance parameters for the use of Treescan, the results of the Treescan-based methods could potentially be dependent on the method of tree construction used. There are numerous phylogenetic programs, applying a variety of methods, that could be used for reconstruction of a phylogeny: such as PAUP* (Swofford, 2003), TCS

(Clement et al., 2000) and PHYLIP (Felsenstein, 2005). The original Treescan paper (Templeton et al., 2005) advocates the use of TCS, however this program does not lend itself to the simulations since the output is a graphic, and is also restricted in that only the Parsimony method of tree construction is possible. In order to assess multiple methods of tree construction, and to be suitable for automating in simulations, it was decided to use the PHYLIP package of methods. This package allows for the phylogeny to be reconstructed using any of three commonly used methods: Maximum Parsimony, Maximum Likelihood and a distance-based method using the Fitch-Margoliash algorithm (Fitch and Margoliash, 1967).

Bayesian methods of reconstructing haplotype trees could also be implemented, such as those advocated by the application MrBayes (Huelsenbeck and Ronquist, 2001). However, there is a disadvantage with the Bayesian approach, in terms of the long run time required for the convergence and sampling of an MCMC chain used to approximate posterior densities of alternate trees, which makes it less suitable for the simulations of Chapter 7. The increased difficulties of interpreting MCMC output automatically, to assess whether convergence to the posterior distribution has been obtained, is a further issue that makes this approach less suitable for simulations and any form of automated use that could be integrated into PheGe-Sim. The implementation of a Bayesian MCMC approach is though a feature that could plausibly obtain small improvements in the results of the Treescan-based methods, although there would still be uncertainty caused by recombination and the same mutation occurring on different lineages (homoplasy), that would be difficult to resolve in any method of tree construction.

The resulting output files from the chosen PHYLIP application are then manipulated in R in order to get the information in a suitable form for the use in Treescan and the Bayes factor version of Treescan. The aim in manipulating the Newick (Felsenstein et al., 2010) format of trees returned by the PHYLIP programs is to process the bifurcating trees to multifurcating trees, where each branch involved corresponds to at least one mutation (the PHYLIP methods returning branches unsupported by any mutation).

Details of the tree construction methods used are given in the following sections, with the approaches being chosen able to run within a reasonable time

data sets of up to 2000 individuals. More in depth discussion of the methods involved can however be found in Felsenstein (1988) and Chapter 16 of Balding et al. (2001). In order to visualize the trees, the Newick code that is produced can be viewed using the Treeview program (Page, 1996).

4.2.1 Parsimony Tree Construction

Uses the following application from the Phylip Package:

dnapars

The application *dnapars* of the PHYLIP package is used to implement the method of maximum parsimony, whereby a tree is chosen that requires the lowest number of mutations to describe the observable haplotypes. The method does not however guarantee to find the globally shortest tree. In situations where there are multiple configurations of tree structure and mutations that are equally fit in describing the data, for simplicity it is chosen to only analyze a single best tree. Details of the specific choices of inputs to *dnapars* are shown in figure B.1(a).

The parsimony method is generally fast to implement for reasonably sized data sets and, if there is no recombination and an infinite sites model is to be assumed, will always reconstruct the true underlying tree. The accuracy of the method can however begin to deteriorate if the assumption of infinite sites is not valid, as the true tree may no longer be consistent with the tree containing the least number of mutations.

The output from *dnapars* details the estimated branch distances between nodes, and also reconstructs unknown ancestral sequences. For the ancestral sequences, it is however required to add arbitrary node labels to the tree as given by the *dnapars* application. The methods of section 4.3 can then be applied in order to reduce the stated output tree to a format that is compatible with Treescan, whereby branches that are estimated to be of zero length are removed.

Outputs:

Parsimony output file saved to folder

4.2.2 Maximum-Likelihood Tree Construction

Uses the following application from the Phylip Package:

dnaml

The maximum likelihood method of reconstructing an unknown phylogenetic tree is applied through the use of the *dnaml* application of the PHYLIP package. Further details of the methods used are detailed in Felsenstein and Churchill (1996). The key aim of the method is to maximize the likelihood with respect to three multi-dimensional parameters: the topology of the tree τ, the branch lengths $\boldsymbol{\nu}$ and a vector of evolutionary parameters θ. Assuming independent evolution at different positions in the nucleotide sequence, the likelihood can be written as follows:

$$f\left(\mathbf{X}|\tau,\boldsymbol{\nu},\theta\right)=\prod_{i=1}^{c}f\left(\mathbf{x}_i|\tau,\boldsymbol{\nu},\theta\right), \tag{4.1}$$

where $\mathbf{X}$ is the full sequence data, and $\mathbf{x}_i$ is the data at the i'th SNP ($i=1,\ldots,c$).

The *dnaml* program allows for the specification of various parameter choices, with the specific choices that have been made illustrated in figure B.1(d). As with the maximum parsimony method, the program can return reconstructed hypothetical sequences for internal nodes and these can subsequently be used to determine whether it is required to retain any unobserved internal nodes, which the R code will arbitrarily label. Branches that are estimated to have a non-significant length are removed according to the methods in section 4.3.

Outputs:

Maximum Likelihood output file saved to folder

4.2.3 Fitch-Margoliash Distance Tree Construction

Uses the following applications from the Phylip Package:

dnadist

fitch

A pairwise distance matrix between sampled haplotypes is computed using the

PHYLIP application *dnadist*. The distances are calculated under the assumption of a Jukes-Cantor (Jukes and Cantor, 1969) model of evolution, a Markov model which assumes that there is an equal probability of changing from the current base to any of the other three bases. More complicated models are though possible, which could take advantage of there being different rates of mutations between different types of bases. Equation 4.2 illustrates the formula used for calculating a pairwise distance, where d is the estimated distance, and p is the proportion of sites with different nucleotides:

$$d = -\tfrac{3}{4} \log_e \left(1 - \tfrac{4}{3} p\right). \tag{4.2}$$

The distance matrix is then used as an input for the fitch program, which constructs a tree using the weighted least squares method of the Fitch-Margoliash (Fitch and Margoliash, 1967) algorithm which aims to minimize the sum of squares:

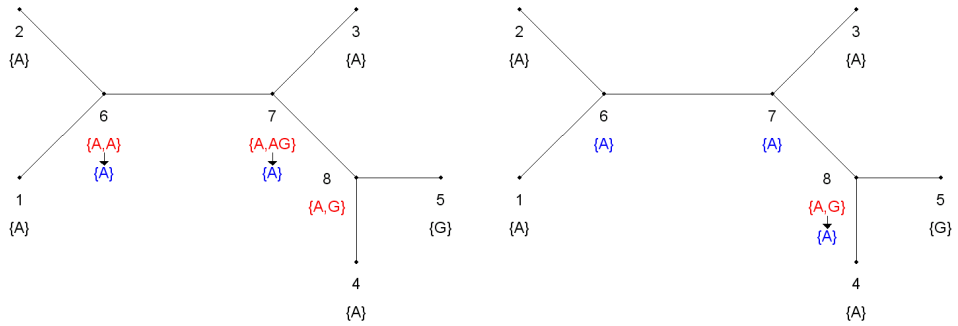
$$SS = \sum_i \sum_j \frac{(D_{ij} - d_{ij})^2}{D_{ij}^2} \tag{4.3}$$

where D is the observed distance between species i and j, and d is the expected distance.

The output of the program details where branches have been estimated to have zero length although, unlike the maximum likelihood and parsimony methods, the fitch application does not indicate the reconstruction of unobserved internal nodes. So, after the branches of estimated zero length have been collapsed, the Fitch algorithm (Fitch, 1971) is used to reconstruct the likely sequences of the non-terminal nodes. An example of a tree output from the Fitch program is shown in figure 4.2. This can be represented in Newick format, where an arbitrary root is taken and nodes are then grouped together inside brackets, until all the nodes present are collected together inside one expression. For this example, the corresponding Newick code with arbitrary branch lengths of one, is:

$$\text{(1:1,2:1,(3:1,(5:1,4:1):1):1);} \tag{4.4}$$

where no internal nodes labels have been assigned.



(a) First 'sweep' of the Fitch algorithm (b) Second 'sweep' of the Fitch algorithm

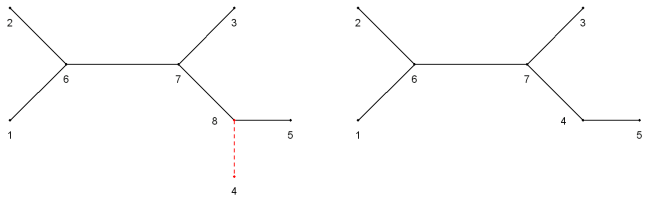
Figure 4.2: Fitch algorithm, where unresolved sites are in red, and subsequently resolved sites are coloured blue. Terminal nodes are fixed, as these are observed, and are coloured black.

All internal nodes are then added to the Newick code, and the bases at each site of these nodes are determined by application of the Fitch algorithm. Figure 4.2 represents the process for a single site, where the root has been chosen to be the node labelled 7. The bases present at each of the internal nodes, 6,7 and 8, are determined by the intersection of the sets of possible bases at nodes immediately below it in the tree and the resulting base is illustrated below in figure 4.2. Node 8 however cannot be resolved, and so the algorithm then proceeds in reverse, as in figure 4.2(b), which therefore assigns the state to be an 'A'. This process is repeated for all the SNPs of the terminal nodes, and ensures that the minimum number of base changes occurs for the constructed tree, while reconstructing the internal nodes. Branches are then removed according to the procedures of section 4.3, if the sequences at the nodes at both ends of the branch are identical. In the example this results in the final Newick form:

$$(1{:}1,2{:}1,6{:}0,(3{:}1,7{:}0,(5{:}1,4{:}1,8{:}0){:}1){:}1); \tag{4.5}$$

Outputs:

Fitch output file Saved to Fitch folder



(a) (b)

Figure 4.3: Initial (a) and reconstructed tree (b), when the branch in red is estimated as carrying no mutations.

4.3 Removal of Branches

Inputs:

Parsimony output file and estimated Parsimony tree

Maximum Likelihood output file and estimated Maximum Likelihood tree

Fitch output file and estimated Fitch tree

The following conditions are applied for each of the tree construction methods to evaluate whether a branch is required or not, as it may be that an interior node is predicted to have the same sequence as a terminal node. Each tree is refined so that there is at least one mutation on every branch on the tree, as otherwise multiple branches will result in the same test statistic value. It should also be noted that for the purposes of Treescan it is only relevant whether a branch length could be equal to zero or not as indicated by the presence or absence of a mutation, and the actual distances involved are of no importance.

There is no known format for giving a coded representation of an ARG, however the Newick format can be used for illustrating a haplotype tree with multifurcating branches. The Newick code returned from each of the PHYLIP programs can subsequently be altered by counting the number of internal and external nodes present within each set of brackets. It is ensured that within each set of brackets there is one node that is labelled as being internal, and that no adjacent nodes are equal to each other as indicated by having the same real or reconstructed sequence.

Figure 4.3(a) relates to the situation whereby the internal node 8 is the same as the terminal node 4. In this situation, the node labelled 8 is removed and node 4 is relabelled as an internal node. This is illustrated in figure 4.3(b),and the corresponding transformation of the Newick form is:

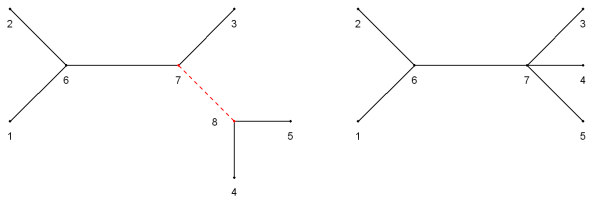
$$(1{:}1,2{:}1,6{:}0,(3{:}1,7{:}0,(5{:}1,4{:}1,8{:}0){:}1){:}1); \rightarrow (1{:}1,2{:}1,6{:}0,(3{:}1,7{:}0,(5{:}1,4{:}0){:}1){:}1);. \tag{4.6}$$

The second criterion to be assessed relating to whether to remove node labels is if the initial tree structure is as indicated in figure 4.4(a), where the internal node 8 is different from both of the terminal nodes 4 and 5, but is the same as the internal node 7. If this combination arises, the effect is to remove node 8 and this results in a multifurcation at the internal node 7, an illustrated in figure 4.4(b). The corresponding transformation of the Newick code is:

$$(1{:}1,2{:}1,6{:}0,(3{:}1,7{:}0,(5{:}1,4{:}1,8{:}0){:}1){:}1); \rightarrow (1{:}1,2{:}1,6{:}0,(3{:}1,7{:}0,5{:}1,4{:}1){:}1);. \tag{4.7}$$

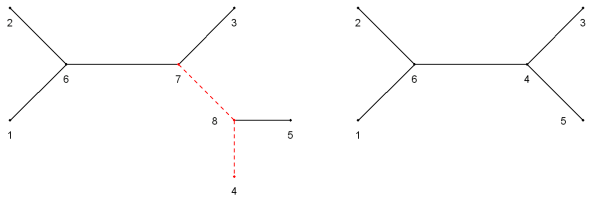
A further criterion to be considered is given in figure 4.5(a). This details a situation whereby the internal node 8 is equal to one of the terminal nodes, in this case 4, and also to the internal node 7. This results in node 8 being removed, and node 7 being replaced with node 4, leading to the tree in figure 4.5(b) and a Newick code as shown in equation 4.8.

$$(1{:}1,2{:}1,6{:}0,(3{:}1,7{:}0,(5{:}1,4{:}1,8{:}0){:}1){:}1); \rightarrow (1{:}1,2{:}1,6{:}0,(3{:}1,5{:}1,4{:}0){:}1);. \tag{4.8}$$



(a) (b)

Figure 4.4: Initial (a) and reconstructed tree (b), when the branch in red is estimated as carrying no mutations.



(a) (b)

Figure 4.5: Initial (a) and reconstructed tree (b), when the branches in red are estimated as carrying no mutations.

The final possible consideration is where the internal node 8 is not equal to any of the nodes that it is connected to. In this situation, the tree remains unchanged as none of the branches can be collapsed. Each section of the haplotype tree is assessed in turn using these criteria, working from the terminal nodes towards the chosen root, as determined by the tree construction method. The resultant haplotype tree that remains will be a multifurcating network of observable nodes and hypothesized ancestral sequences, as is required for the Treescan-based methods. Information relating to the haplotypes of the nodes remaining in the tree is retained, so that the SNPs that relate to each branch can be identified.
Outputs:

Collapsed Fitch tree - *pentax.pars.fitch*

Collapsed Parsimony tree - *pentax.pars.pars*

Collapsed Maximum Likelihood tree - *pentax.pars.ml*

4.4 Methods of Association

Inputs:

Observable haplotype sequences of SNPs

Number of permutations for p-value calculation - *num.perm*

There are numerous association methods that have been programmed for use in both the PheGe-Sim and PheGe-Find programs, and these are illustrated in table 4.1. The approach used for the standard Treescan method has previously

Table 4.1: Methods used in the association studies of PheGe-Find and PheGe-Sim.

Method	Frame	Extra Tests	Correction	Comment
Tscan	Standard Bayesian Bimbam	Pars Fitch Max Lik	Snp Branch	Treescan based methods are programmed to run under all combinations of frame, extra tests and correction (resulting in a total of 18 different interpretations).
Single SNP	Standard Bayesian Bimbam Bonferroni		Snp	Single SNP methods can only assess SNPs individually, or in combination with one or two others, before the number of combinations becomes prohibitive.
Single Hap	Standard Bayesian (Bimbam) Bonferroni		Branch	Haplotype methods have no ability to detect which SNPs upon a haplotype confer an increased association, unlike the Treescan based methods.

been described in section 1.5.2, and this is modified for the creation of the standard single SNP and standard single haplotype procedures. The flow chart of figure 4.6 shows the steps taken in determining whether any significant groupings have been found.

As each individual will contain two copies of a gene, there are three possible combinations of bases that are possible, assuming that only two variants are observable at each location. The combination of bases that an individual has is determined for each site, and the phenotype score from that individual is then assigned to the appropriate base grouping.

At this stage, the Bonferroni method and the single SNP version of Treescan both calculate an F statistic from an ANOVA comparing the distributions of

[1]Only applies once for each causative SNP to ensure that there are fair comparisons between simulations and to avoid declaring more correctly found haplotypes than there are causative SNPs

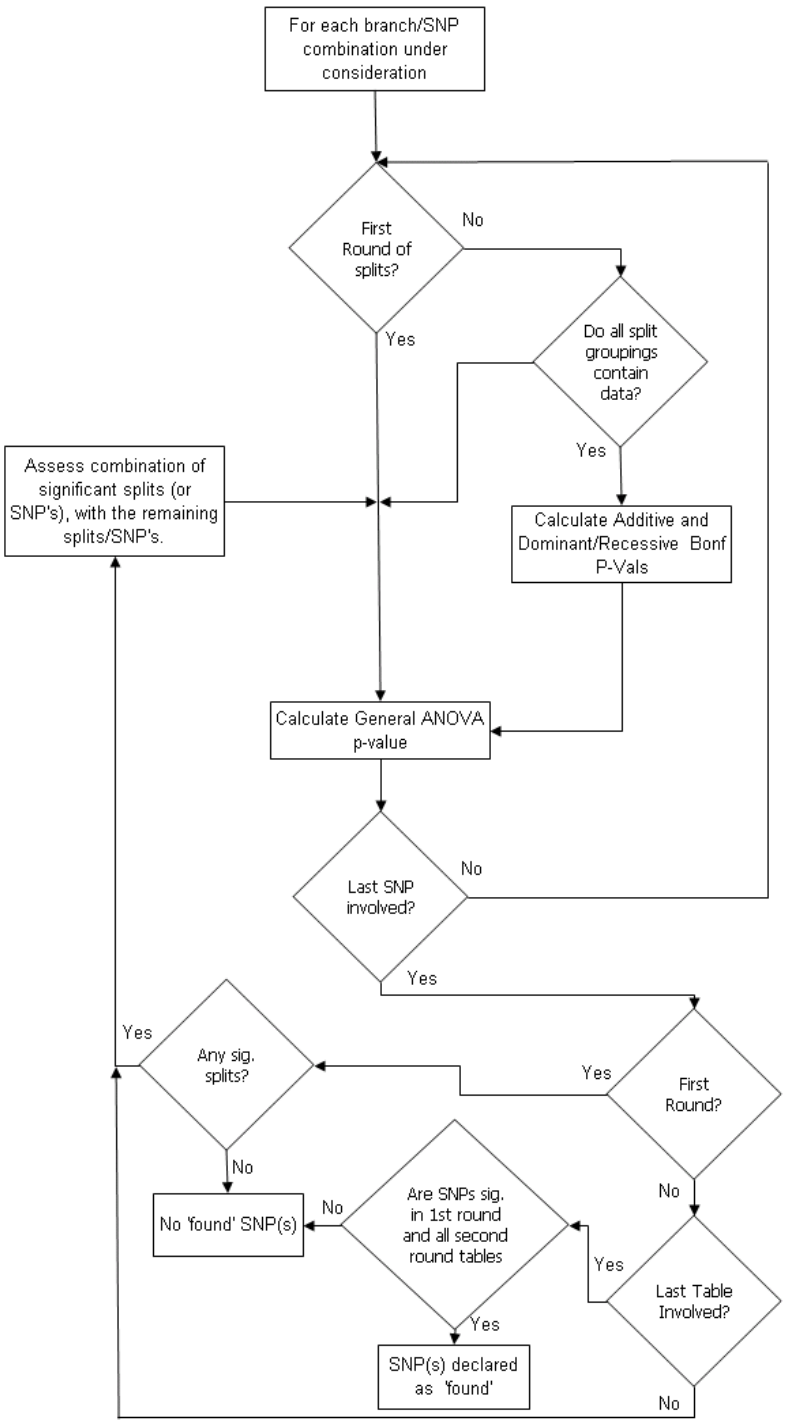


Figure 4.6: Flow chart of the decisions involved in frequentist methods. These are chosen so as to mirror the steps taken in the Treescan procedure (figure 1.3), with both first and second round tests of association being used. A SNP is declared as 'found' if it is found in the first round of tests, and all the other tables of second stage splits.

phenotypes in the three genotype classes. The Bonferroni method then simply adjusts the significance level α from the ANOVA using the standard Bonferroni correction for multiple testing, namely the equation $\alpha_{adj} = 1 - \frac{\alpha}{n}$, where n represents the total number of tests performed. Sites are subsequently adjudged to be significantly associated with a change in phenotype if the p-value is less than the adjusted significance level.

The Single SNP method adjusts the F-statistic by the use of the Boerwinkle-Singh (Boerwinkle and Sing, 1986) corrected estimator, so as to remain consistent with the methodology used in the Treescan program. This correction involves adjusting the observed phenotype variance, by the number of genotype classes present in that comparison:

$$s_G^2 = \sum_{i=1}^{k} \frac{n_i(\bar{Y}_i - \bar{Y})^2}{n} - \frac{k-1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(Y_{ij} - \bar{Y}_i)^2}{n-k}, \tag{4.9}$$

where n is the total sample size, k is the number of genotypic classes with observations, $\bar{Y}$ is the sample grand mean, n_i is the number of individuals in the ith genotypic class, and Y_{ij} is the phenotype of the jth individual with genotype i.

The final adjusted p-values from the single SNP method are then obtained by the permutation and correction methods detailed in the Treescan approach (section 1.5.2). As with the Treescan method, a SNP is adjudged to be significantly associated with the phenotype if it is found as significant in the first table of splits, and in all second round tables conditional on the other found SNPs (see figure 4.6).

The method of using Bayes factors for assessing the differences of phenotypes between the genotypes at each SNP will be discussed in detail in Chapter 3. In contrast to the frequentist approaches, the Bayes factor versions computes all second-round comparisons, as it may be that a combination of splits that are not found significant in the first round may have a significant effect when considered together. If second-round splits are also found as significant, then third-order splits are considered using only the combinations of SNPs that are found as significant in the previous round.

The set of all combinations in third and higher-order splits is not considered, as the number of comparisons increases dramatically for higher-order splits. For

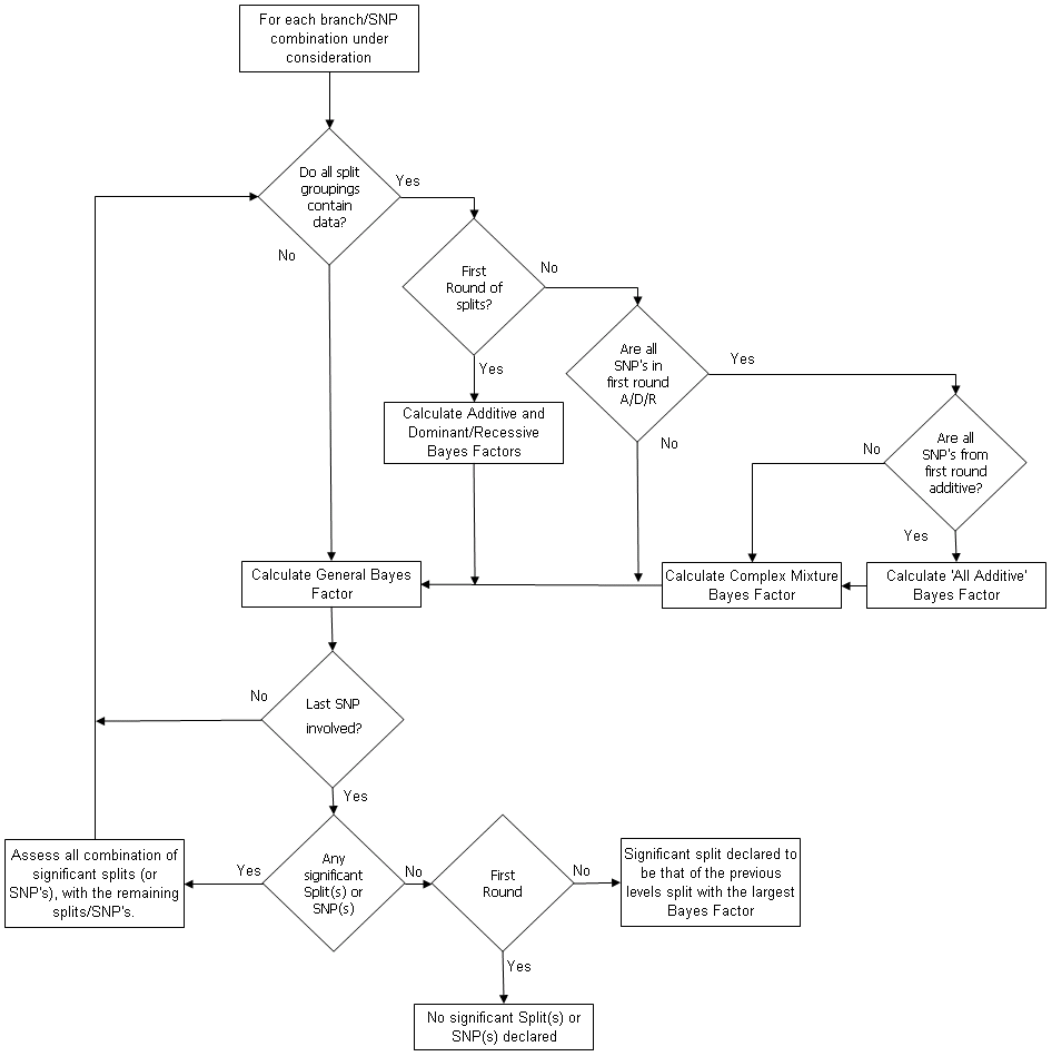


Figure 4.7: Flow chart of the decisions involved in Bayesian methods. The approach aims to test all first round associations, and then testing subsets of the splits that appear to be associated from the previous rounds of tests. At each round of splits all the relevant tests of association are used, and the maximum value is taken for use in determining if a SNP/split is declared to be found.

example, for n SNPs a third round comparison would involve $\binom{n}{3}$ comparisons, and therefore a typical data set of 70 SNPs, such as the blood pressure data set analyzed in Chapter 6, would involve $54,740$ comparisons at the third-level of splitting. This would increase the computation time considerably, and is unlikely to discover any SNPs that have not been indicated as significant to some extent in the first or second-round splits.

In order for a SNP or branch to be declared as significant in a Bayes factor approach, it has been decided that a Bayes factor of > 150 is required to obtain 'very strong' evidence against H_0, according to the categories of table 1.2 (Kass and Raftery, 1995). No choices are made for the prior odds of association, which would vary in practice depending on the strength of prior knowledge available. However, the choice of a region to use for a fine-scale study would in itself indicate that there is a suggestion, from a GWAS or from previous studies, that a variant in the region is to some extent associated with the phenotype. Figure 4.7 illustrates the steps used in calculation of the Bayes factors. It can be seen that for the first round of tests, a SNP or branch is to be designated as found if a Bayes factor is obtained that is significantly larger than the overall null model. At the second and subsequent level of splits, the Bayes factor is calculated according to the comparison between the marginal likelihood of the current level, and the highest marginal likelihood of the previous round of tests.

4.4.1 BimBam

BimBam (Bayesian IMputation Based Association Mapping, Servin and Stephens (2007)) is a method that implements exactly computed Bayes factors for use in phenotype-genotype association studies. Conceptually similar to the Bayes factors used in PheGe-Find, there are however some differences in the approaches that can have implications for the outcome and interpretation of the resulting Bayes factors.

As with PheGe-Find (Chapter 3), BimBam uses conjugate priors for defining the prior mean (μ) and the reciprocal of the variance (τ) of the phenotype data,

that is assumed to be Normally distributed. These priors are of the form:

$$\tau \sim \Gamma\left(\frac{\kappa}{2}, \frac{\lambda}{2}\right), \tag{4.10}$$

$$\mu|\tau \sim N\left(0, \frac{\sigma_\mu^2}{\tau}\right), \tag{4.11}$$

where κ, λ and σ_μ^2 are hyperparameters to be specified.

BimBam subsequently takes the limiting form of this distribution, that is letting $\kappa, \lambda \to 0$ and $\sigma_\mu^2 \to \infty$. Although this results in an improper prior distribution, the posterior can be shown to be proper, with Bayes factors which are reported as tending towards sensible limits. As a result of this, BimBam requires the specification of two other hyperparameters, σ_a and σ_d, specifying respectively a standard deviation of an effect size, and a measure of how close the effect of a mutation is to being additive. A further prior choice is made regarding a distribution concerning the number of SNPs that are likely to be defined as causative, denoted by the distribution $p(l)$, where l represents the number of specified SNPs.

An advantage of the approach taken by BIMBAM is that the prior mean and variance are not required to be specified, thus avoiding uncertainty about what represents suitable information to assign these values. However, the lack of specification can also be a negative feature, as if there are data available from another source then this will not be able to be incorporated into the Bayes factors of Bimbam. In real data sets there is likely to be some prior knowledge about what range of values is expected for a given phenotype, and basing the prior mean and variance on the sample of data is not always appropriate.

A further difference between the methods used in BIMBAM and those used in this thesis is that BIMBAM requires specification of the distribution of SNPs that are to be determined as causative. The prior used puts equal weight on any choice of up to four causative SNPs, as it is stated that the alternative of calculating each Bayes factor in comparison to a null model will lead to implicit prior assumptions about the relative plausibility of each multiple SNP model. However, the view taken in this thesis is that there is decreasing plausibility of larger numbers of SNPs being associated with a phenotype, and that the effect of

multiple SNPs should provide significant improvement on a simpler model with only a single variant. This is possibly most relevant in the context of this thesis, that is fine-scale studies as opposed to GWAS, as due to strong linkage between variants there will be multiple correlated SNPs associated with a phenotype as a result of one true signal. In this setting, it is preferable that strong evidence is required to suggest that the apparent effects of multiple variants is not simply the linked effects to the actual causative variant. An improvement suggested but not implemented by the BIMBAM approach is to use a prior with decreasing probability for larger numbers of causative SNPs. This is a possibility that needs further work but that could potentially be useful for the situation of assessing multiple linked SNPs.

As with the approach taken in this thesis, BIMBAM does not focus on the specification of the prior odds of an association. The choice of the prior odds will be context specific and, irrespective of the method that is chosen in calculation of the Bayes factors, should be chosen with care dependent on prior knowledge about the regions and variants being assessed.

4.5 PheGe-Find Output

In addition to the output files created for each of the methods of association and tree construction that have been described previously, PheGe-Find also constructs three plots relating to the data. However, PheGe-Find does not create the *sim_results* and *details* files that were created in PheGe-Sim, as in real data sets there will be no knowledge as to what are the 'true' causative mutations.

4.5.1 Linkage Plots

PheGe-Sim creates a PED file of the genotypes involved in a data set, which is then used as an input for the Haploview (Barrett et al., 2005) application. Haploview is automatically run through its command line options (section B.4), and will produce a linkage plot of the genotypes involved. The statistics used to illustrate the linkage between each pair of SNPs are D' and the LOD (Log Odds)

score. D' is defined by:

$$D' = \begin{cases} \frac{D}{min(p_1q_2, p_2q_1)} & \text{if } D \geq 0, \\ \frac{D}{-min(p_1q_1, p_2q_2)} & \text{if } D < 0, \end{cases} \tag{4.12}$$

with perfect linkage corresponding to a $D' = 1$, no linkage corresponds to $D' = 0$, and where:

$$D = p_{11}p_{22} - p_{12}p_{21} = p_1q_1, \tag{4.13}$$

where the p_{ij} values are the proportions of each entry from the following:

		SNP 2 allele 1	SNP 2 allele 2	
SNP 1 allele 1		p_{11}	p_{12}	q_1
SNP 1 allele 2		p_{21}	p_{22}	q_1
		p_1	p_2	

The LOD score is calculated according to the following equation (Ott, 1999):

$$\text{LOD} = \log_{10} \frac{\max_r \text{P}(\text{data}|r)}{\text{P}(\text{data}|r = \frac{1}{2})} \tag{4.14}$$

where r corresponds to the recombination fraction. The computation of this in Haploview relies upon a two-marker EM algorithm for estimating the maximum likelihood values required. Information regarding the D' and LOD scores are then visually summarized in Haploview according to the classifications of table 4.2.

Table 4.2: Key to colour combinations of linkage plots, reproduced from the help files of Haploview (Barrett et al., 2005).

	$D' < 1$	$D' = 1$
LOD < 2	white	blue
LOD ≥ 2	shades of pink/red	bright red

4.5.2 Manhattan Plots of Results

PheGe-Sim produces 'Manhattan' plots summarizing the results of the single SNP and Treescan analysis, for both the frequentist and Bayesian settings. Results are displayed according to the maximum $-\log_{10}$ p-value (i.e. the minimum p-value), or Bayes factor that has been observed in all the tests of association under consideration, to ensure that the scales of each of the plots are comparable. The logarithm of the p-values is taken to ensure that the results are strictly positive on a scale of zero to infinity. Treescan results are displayed as the association levels for each of the SNPs as opposed to branches, and as such there is liable to be multiple 'hits' for each SNP under consideration. If a valid *Site Positions* file has been specified, the results will be displayed according to the specified locations of each SNP. If this information is not provided, the plot will use a default spacing of one unit between each of the SNPs of the study.

4.5.3 Sensitivity Analysis

As with most Bayesian analysis, the Bayes factors that are to be developed in Chapter 3 and have been used for the association studies, require the specification of hyperparameters in the prior distribution. Having to make choices of prior distributions can be both a positive and negative feature of any analysis. However, it would be hoped that any conclusions from a data set would not be substantially altered by small changes in prior specifications. PheGe-Sim produces two plots that can provide visual summaries of the sensitivity to these prior choices; one of these corresponding to the prior for the mean of the data, and the other corresponding to the prior on the within-group variance.

The normal-inverse-chi-squared distribution used for the specification of the prior distribution (equation 3.11) consists of dependent normal and inverse-chi-squared distributions. The normal distribution describes the mean of the data, whereas the inverse-chi-squared distribution describes the variance.

The mean of the prior distribution for the phenotype mean, can be specified by the μ_0 parameter. As a default choice for this value, the mean of the sampled data can be used, however, μ_0 can ideally also be chosen according to the mean of a suitably similar data set to the one being assessed.

The choice of κ_0 relates to the information on μ, with κ_0 number of observations of mean μ_0 and variance σ^2. Choosing a large value of κ_0 would imply that there is a large degree of certainty about the mean of the data; and conversely choosing κ_0 as zero would imply that there is no knowledge of the potential mean of the data set. A default value of 20 is chosen, which appears to be reasonable in conveying a sufficient amount of confidence in the prior mean of the data. The effect of κ_0 on the Bayes factors is relatively minimal for reasonably sized data sets (Chapter 6), but for small data sets it can have a stronger impact and lower values may be more appropriate (Chapter 5). The mean and variance of the inverse-chi-squared distribution for the within-group variance, are respectively:

$$\mathrm{E}[\sigma^2] = \frac{\nu_0 \sigma_0^2}{\nu_0 - 2} \tag{4.15}$$

and

$$\mathrm{Var}[\sigma^2] = \frac{2\nu_0 \sigma_0^4}{(\nu_0 - 2)^2(\nu_0 - 4)}. \tag{4.16}$$

It can be seen that the hyperparameters of ν_0 and σ_0^2 are entangled, as they both appear in the formulations of the mean and the variance of the variance distribution. It should be noted that the mean is only defined if $\nu_0 > 2$, and likewise the variance is only defined if $\nu_0 > 4$.

Analogous to the choice of κ_0, the information on σ^2 is equivalent to the choice of the number of observations, ν_0, with variance σ_0^2. Distributions where ν_0 is less than 4, resulting in prior distributions with infinite variance, cause the Bayes factors to be unstable and unreasonably high. An increase in ν_0 above 4 results in a steady increase in the Bayes factor, assuming that the prior mean is reasonably close to the mean of the sample data. It has therefore been chosen that the default value of ν_0 shall be 20, and the choice of σ_0^2 shall be chosen as the sample variance. If there is therefore a reason for increased confidence in prior knowledge of the variance then ν_0 can be increased to reflect this, and the prior variance will become concentrated around σ_0^2.

Using the sample mean and variance as default choices of priors can result in the data artificially being used twice, and therefore may result in a misplaced confidence in a lack of uncertainty relating to the estimates. As a starting point

they can though be useful as the calculation of meaningful Bayes factors requires the prior values to be reasonably consistent with the sample, as otherwise the reported Bayes factors will be uniformly small. However, the use of previously reported information relating to inform the hyperparameter choices should be used whenever possible to avoid an overconfidence in the interpretation of results.

Chapter 5

Drosophila melanogaster Data

5.1 Background of Data

Drosophila melanogaster is a breed of fruit fly that has been extensively studied in genetics since the beginning of the twentieth century, and studies by Thomas Morgan of the phenotype characteristics of the fly formed the basis of many aspects of modern genetics. The use of *D. melanogaster* in genetic studies subsequently became popular due to the relative ease with which the flies could be maintained, and due to the short life cycle and high breeding rate of the flies. Further understanding of genetics led to the discovery that the DNA of the *D. melanogaster* fruit fly contains four chromosomes, three autosomes and an X-Y pair. The chromosome number is substantially smaller than that of mice and humans and so the fruit-fly was one of the first organism to have its DNA fully mapped, with the information currently being contained online in the FlyBase database (FlyBase, 2010).

Although the chromosome number is substantially lower than that of humans, there are some genes that are present in the fly that are directly related to an equivalent human gene. One such gene is ADH, that is located on chromosome 2L of *D. melanogaster* and on chromosome 4q of human DNA. This gene is responsible for the production of alcohol dehydrogenase, an enzyme involved in the breakdown of alcohol, that in the fly is present as a result of its diet consisting of rotting fruit. As humans also contain the ADH gene, numerous studies have

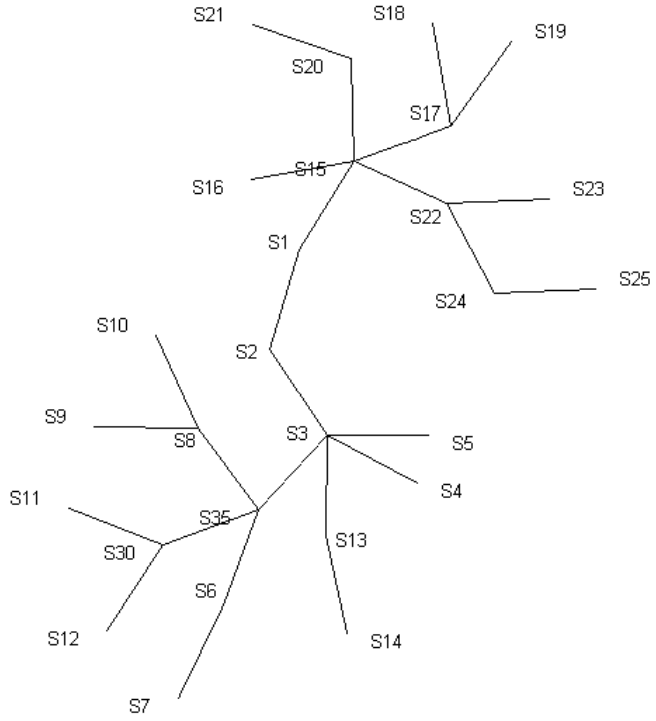


Figure 5.1: Reconstructed haplotype tree using the parsimony method.

been conducted using *D. melanogaster* in an attempt to gain an insight into potential risk factors of alcoholism. One such study was conducted by Aquadro et al. (1986), to gain an understanding of the relationship between different strains of the fly and the ADH gene as measured by the activity levels of the alcohol dehydrogenase enzyme.

Aquadro et al. (1986) constructed a phylogenetic tree using parsimony methods to describe the relationship between the various strains of *D. melanogaster* based on the ADH sequence data, and also found that there was an association between a variant at one of the sites in the ADH gene and alcohol dehydrogenase enzyme activity, when assessing each SNPs relationship with the phenotype individually. Templeton et al. (1987) subsequently suggested using the reconstructed tree implicitly in attempts to find associations, and this resulted in the nested-clade analysis (NCA) method. The fruit flies involved in the Aquadro study were bred to be homozygous for the ADH region. However, this is impossible for human studies, and so Treescan was developed which takes into account the

potentially heterozygous nature of most naturally occurring DNA. The Aquadro data set was reanalyzed by Treescan to provide a valid comparison with the NCA method, and in a similar way is being reanalyzed here to provide valid comparisons between the Bayes factors and frequentist-based methods.

The data available from *D. melanogaster* differs from the usual data sets required for Treescan-based analysis in a variety of ways. The first difference is that the *D. melanogaster* have been back-crossed so as to be homozygous, and so, unlike for most human data sets, it is not possible to test for any additive, general or recessive effects. As the flies are homozygous, there is also no requirement to phase the data set. A second difference is that insertions and deletions (indels) are present in the genotypes of the fruit flies, and these have been treated in a similar way to the treatment of SNPs by simply coding whether the indel is present or absent in a particular haplotype. There may be complex differences in the biological mechanisms that result in SNPs or indels that are not being accounted for in this approach. However, recoding in such a manner does not change the analysis methods and therefore allows for valid comparisons with the approach employed by Treescan (Templeton et al., 2005).

Figure 5.1 shows the haplotype tree relating the 48 *D. melanogaster* present in the study, that has been reconstructed using the parsimony method as detailed in Chapter 4. This corresponds exactly to the tree that was obtained by Aquadro et al. (1986) and subsequently by Templeton et al. (2005), and infers the same unobserved internal sequences that have been automatically labelled by the PheGe-Sim application. Although of small size, the data can be shown to be roughly normally distributed, as is required for some of the following analysis.

5.2 Linkage Plot

Figure 5.2 illustrates the linkage between the SNPs that are present in the *D. melanogaster* data set. There does not appear to be any clear pattern from the plot, although this is largely due to the relatively small data set in terms of sample size (48) and the small number of variants (22). In particular, the large number of blue squares illustrate regions where $D' = 1$ and $LOD < 2$ and are generally a result of low sample size and therefore low power to detect

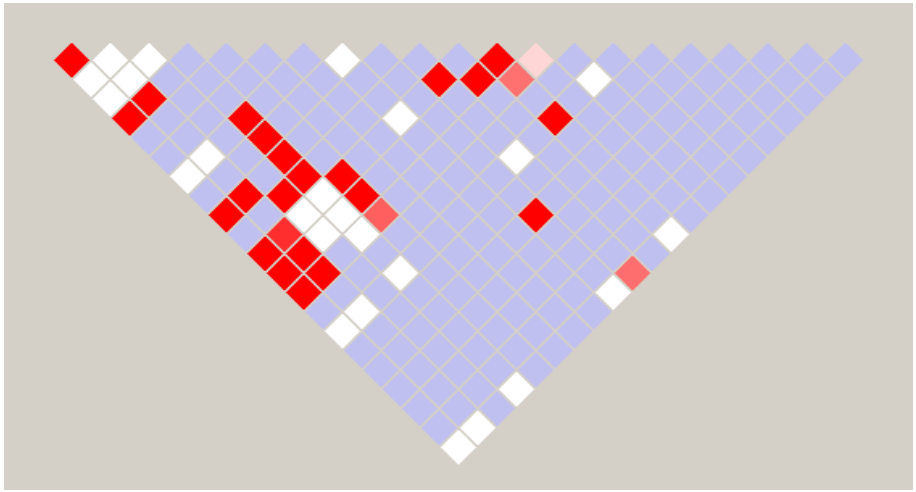


Figure 5.2: Linkage plot for *D. melanogaster* data. The plot is predominately blue due to evidence of linkage, but there being only a relatively small sample size. See table 4.2 for the details of the colouring used in the plot.

linkage. It should be noted that there are small differences between the estimates given by Haploview and those given in Aquadro et al. (1986), due to Haploview using Hedrick's (1987) estimator of D', whereas Aquadro uses Lewontin's (1964) estimate of linkage disequilibrium.

5.3 Estimating Recombination Rates & Hotspots

The *interval* program of the LDHat (McVean et al., 2002) package is a program that estimates recombination rates based upon imputed phased haplotypes. Although a newer version of the program has been written (*rhomap*), the *interval* program was used so that results in Chapter 6 can be comparable with the current estimates given in HapMap for human data. The application uses a composite likelihood estimation of the recombination rate, as calculation of the likelihood of the full data set is intractable. This method was first implemented by Hudson (2001) but adapted to include the possibility of a finite-sites model of mutation in the LDHat package. The inclusion of the possibility of a finite sites model is particularly relevant, as recurrent mutation and recombination can plausibly

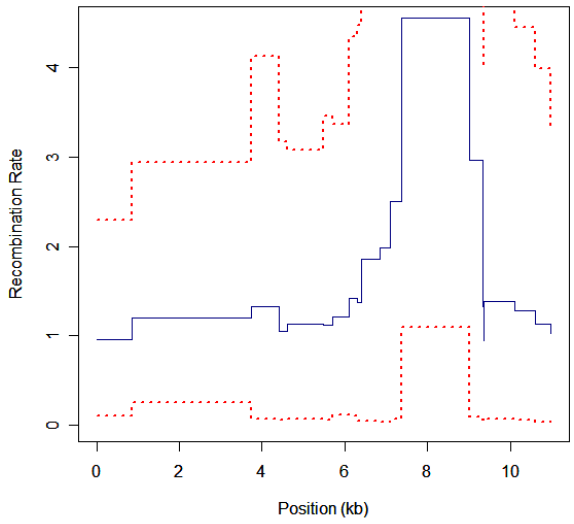
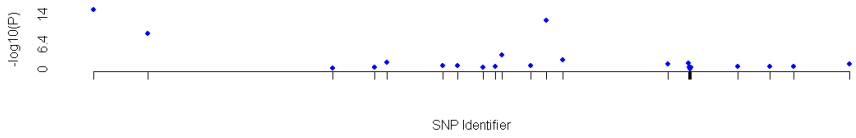


Figure 5.3: Recombination estimates directly from *D. melanogaster* data.

result in identical sequences of DNA.

The composite likelihood scheme involves four key stages, the first of which is to estimate the population mutation rate per site. All possible two-locus combinations of sites are then compared, and the likelihood of each comparison is calculated. The overall population recombination rate for the data set is then estimated by combining the pairwise comparisons. Further details of the scheme are given in the methods section of the paper by McVean et al. (2002).

Due to the small number of variants and small sample size of the number of flies under consideration, it is unlikely that clear estimates of recombination could be estimated using this data set. This proves to be the case as illustrated in figure 5.3, with recombination estimates suggesting some evidence of recombination, but with a large degree of uncertainty due to the small sample size. For the analysis in the following sections it is assumed that the data can be treated as if from one region of low recombination, because of the low sample size and relative simplicity of the reconstructed haplotype tree of figure 5.1.



(a)

(b)

Figure 5.4: *D. melanogaster* results, for the frequentist (a) and Bayesian (b) versions of the single SNP method. Site locations are approximate, based upon information in Aquadro et al. (1986).

5.4 Single SNP-based Analysis

Figure 5.4(a) illustrates the results from the single SNP method, with the uncorrected p-values from ANOVA tests being displayed. Initially it seems that there are three variants that are strongly associated with the phenotype, labelled SNP-1,SNP-12 and SNP-2, with p-values extremely low for such a small sample size (table 5.1). A Bonferroni correction for multiple testing would require that the p-values are lower than 2.272×10^{-3} for the 22 variants considered, and as such all three would be considered to be associated with the ADH phenotype. A further two SNPs (labelled SNP-10 and SNP-13) are associated with p-values of less than 0.05, which are not small enough to pass the Bonferroni-corrected level of significance. The linkage plot of figure 5.2 shows that the three variants with strong associations are also in strong linkage with each other, and so it may be that not all three are indeed causative. According to the procedure used by the single SNP method (figure 4.6), only SNP-1 would be declared to be significantly associated with the phenotype as none of the other SNPs considered account for any significant signal in addition to this SNP.

The results from the first round of association using Bayes factors are shown

Table 5.1: Selected single SNP (top), Treescan (middle), and Haplotype (bottom) results of analysis of *D. melanogaster* data. Note as homozygous bred there are no heterozygous measurements

SNP identifying label	nAA	nBB	μ_{AA}	μ_{BB}	General BF	General p-Value Uncorrected	p-Value Corrected
SNP-1	16	25	7.8	3.5	2446.0	4.8×10^{-16}	0.00
SNP-12	18	23	7.5	3.4	1302.4	2.2×10^{-13}	0.00
SNP-2	19	22	7.2	3.5	277.1	5.7×10^{-10}	0.00
SNP-10	8	33	7.7	4.6	1.8	2.6×10^{-4}	0.00
SNP-13	8	33	3.2	5.7	0.7	5.9×10^{-3}	0.00
SNP-1	16	25	7.8	3.5	2446.0	4.8×10^{-16}	0.00
SNP-12	17	24	7.6	3.5	1323.9	1.0×10^{-13}	0.00
SNP-2	19	22	7.2	3.5	277.1	5.7×10^{-10}	0.00
SNP-10	8	33	7.7	4.6	1.8	2.6×10^{-4}	0.00
SNP-13	6	35	3.3	5.5	0.4	2.6×10^{-4}	0.00
HAP-23	2	39	8.6	5.0	0.3	0.034	0.21
HAP-2	5	36	3.2	5.5	0.4	0.036	0.21

in figure 5.4(b). The chosen hyperparameters are the defaults of PheGe-Sim, namely: $\mu_0 = 5.204, \kappa_0 = 20, \sigma_0^2 = 5.383, \nu_0 = 20$. The values of κ_0 and ν_0 are relatively high for this setting due to the small sample size, and therefore may strongly affect the resultant Bayes factors, particularly as the mean and variance have been set at the true levels of the sample. The Bayes factors do indeed fluctuate to a reasonable extent depending on the combination of all four of the hyperparameters, however, remain reassuringly high at the same variants largely irrespective of the specific values that are chosen.

The pattern of association is similar to that for the frequentist approach, although variants SNP-10 and SNP-13 have comparatively lower association in the Bayesian setting[1]. The procedure used for Bayesian methods in determining the number of SNPs associated with the phenotype (figure 4.7) would determine that only variant SNP-1 is found to be associated, as at the second level of

[1]Note, however, unlike the p-values, the Bayes factors are not on the log scale.

(a)

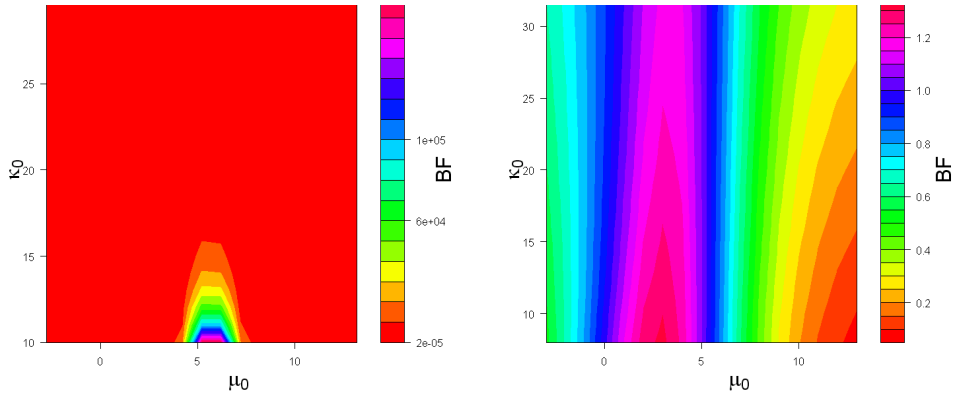
(b)

Figure 5.5: *D. melanogaster* results, for the frequentist (a) and Bayesian (b) versions of Treescan. Site locations are approximate, based upon information in Aquadro et al. (1986).

tests the addition of extra variants does not significantly improve the marginal likelihood of the alternative model.

5.5 Treescan & Haplotype Analysis

Treescan-based analysis can be performed upon the *D. melanogaster* data and results are shown using the standard approach in figure 5.5(a), and for the Bayesian alternative in figure 5.5(b), and are summarized in table 5.1. Unlike for most human data sets, there is no uncertainty in the haplotype reconstruction as the haplotype have been explicitly observed due to the flies being bred to be homozygous. The patterns of association for both Bayesian and frequentist approaches are strikingly similar to those of the comparable single-SNP analysis, due to there being only eight mutations occurring in more than one position on the reconstructed maximum parsimony haplotype tree. The Manhattan plots of figure 5.5 do however illustrate one of the potential problems with the use of Treescan, namely that SNPs can result in being declared as both associated and un-associated with the phenotype at different locations on the haplotype



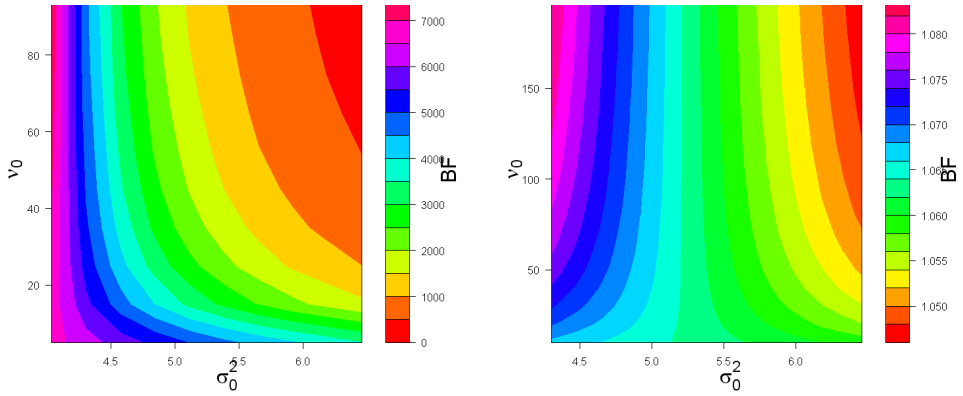
(a) *D. melanogaster* sensitivity to the μ_0 and κ_0 hyperparameters (SNP 1)

(b) *D. melanogaster* sensitivity to the μ_0 and κ_0 hyperparameters (SNP 6)

Figure 5.6: *D. melanogaster* sensitivity to hyperparameters.

tree. This is particularly true in this example, whereby the haplotype tree reconstruction has resulted in there being two instances SNP-12 occurring (table 5.1). Based upon this, it is not possible to determine whether a mutation is indeed responsible for a change in phenotype measurement, and it can only be declared as to the association of the specific haplotype with the phenotype. If it is indeed SNPs or other variants as opposed to haplotypes that are the driving force responsible for a phenotype, then the Treescan method fails to be useful in comparison with the single SNP approaches for this particular data set.

Testing the associations of haplotypes with the phenotype when ignoring the tree structure is also possible using the PheGe-Sim application, and the two haplotypes with strongest associations are presented in table 5.1. It is apparent that haplotype-based analysis for this particular data set is ineffectual, as a result of there being only 25 unique haplotypes for the 41 flies involved and therefore there is extremely low power to detect any differences that could exist. This is apparent in both the Bayesian and Frequentist settings, assuming that the p-value is corrected for multiple testing in an appropriate manner.



(a) *D. melanogaster* sensitivity to the ν_0 and σ_0^2 hyperparameters (SNP 1)

(b) *D. melanogaster* sensitivity to the ν_0 and σ_0^2 hyperparameters (SNP 6)

Figure 5.7: *D. melanogaster* sensitivity to hyperparameters.

5.6 Sensitivity Analysis

The Bayes factors that have been presented are equivalent to the posterior odds of association as they have not been adjusted by prior odds, and the sensitivity of these Bayes factors to hyperparmeter choices can be explored. Figure 5.6(a) represents the sensitivity of SNP-1 to the hyperparameters of the overall mean. It can be seen that the Bayes factors that are obtained are highly sensitive to the choice of both the μ_0 and κ_0 hyperparameters. The Bayes factors for this example are extremely sensitive to the prior choices mainly because of the small sample sizes involved, and so the choice of prior values in this situation can have an unreasonably high impact on the resultant Bayes factors.

Figure 5.6(b) illustrates the sensitivity to the Bayes factors of SNP-6, a variant that appears to have little or no association with the ADH enzyme activity. In this situation, unlike that for SNP-1, the choice of the mean related hyperparameters has only a marginal effect on the Bayes factors and that, irrespective of the values that are chosen, the Bayes factors remain small. The plots of figure 5.6 are reassuring in that with reasonable choices of prior values apparent true associations can be found, but irrespective of prior choices variants that show little association in the sample data will not be mistakenly interpreted as being in association with the phenotype.

Careful consideration must also be made for the hyperparameters ν_0 and σ_0^2, that are used for describing the prior variance. In order to produce reasonable Bayes factors the choice of ν_0 should be as low as possible, as a higher value represents increased confidence in the prior distribution of the within-group variance. Figure 5.7(a) represents the sensitivity of the Bayes factors to the variance hyperparameters for a SNP that appears to be associated with the phenotype. It can be seen that the plot appears to consist of curved segments, with it being suggested that an increase in Bayes factor would be obtained by using a lower variance of the within-group variance. This is a result of the within-group variance being substantially lower than the between-group variance, which has been used to fix the σ_0^2 hyperparameter of this test. This is in itself an indication that there may be a true effect at this SNP, however, the Bayes factors are generally suggestive of an effect at most combinations of ν_0 and σ_0^2. Figure 5.7(b) illustrates the sensitivity of the Bayes factors at a SNP with little apparent association with the phenotype, and displays a pattern that is more centred on the overall variance of the data than was displayed for SNP-1. Irrespective of the choices made for the prior variance, the Bayes factors remain reassuringly low at this SNP.

5.7 Conclusions

The analysis presented of the ADH data set results in similar conclusions to those previously reported by Templeton et al. (2005) and Aquadro et al. (1986). The three SNPs that have been found to be strongly associated with the ADH phenotype are found irrespective of whether a single SNP-based or Treescan approach is used, or whether the strength of association is tested using p-values or Bayes factors. It is also observed that the use of haplotype-based analysis is inappropriate for this data set, due to the small sample size that results in low power at detecting any potential associations.

The three SNPs that have been found to be associated with the phenotype are, however, in strong linkage with each other, and so there may be only one true causative variant. The most likely candidate for true association is the variant with the strongest association, namely SNP-1. There are no significant associations at the second round of splits in any of the forms of the analysis.

However, this could potentially be a result of there being a small sample size and therefore there is low power to detect multiple SNP effects.

Detailed differences between the Treescan and single SNP-based approaches are difficult to ascertain for this data, as there are only a few instances of repeat mutations on the reconstructed haplotype tree. It can be seen though that a stronger effect of SNP-12 is found using the Treescan approach, however, figure 5.5 shows that this SNP also appears to be unassociated with the phenotype at another location on the haplotype tree. The conclusion that can be made regarding this situation is open to interpretation, and it can reasonably be argued as a benefit or indeed a drawback of the Treescan method that the two instances of the mutation have been differentiated in this way. The argument of it being a drawback does though seem to be a more reasonable position, as a strong association at this SNP is nonetheless found using both p-values and Bayes factors when using a single SNP approach.

The comparison between the Treescan and single-variant approaches can also be useful in illustrating the potential benefits of a Bayes factor approach over the standard use of p-values. SNP-12 is found to have p-values differing by 220% between the two approaches, even though there is only a small difference in the allocation of phenotypes under the two scenarios. The Bayes factors, on the other hand, are more robust to such small adjustments of the observed data, and the Bayes factor only changes by 1.65% from 1302.4 to 1323.9. Although the Bayes factors are indeed sensitive to the hyperparameter values used for the prior distribution, the default choices of PheGe-Find have been demonstrated to result in reasonable values.

Chapter 6

ADRA1A Data

6.1 Background of Data

The α_{1A} adrenergic receptor gene (ADRA1A) is located on chromosome 8 of the human genome, and is involved in the contraction (vasodilation) and expansion (vasoconstriction) of blood vessels. As a result of these actions, the ADRA1A gene has been targeted in the treatment of hypertension and benign prostatic hyperplasia (BPH) through the use of alpha-adrenergic antagonist drugs such as Prazosin (Mancia et al., 1980). The effect of these drugs in the treatment of hypertension is that the drug inhibits the vasoconstriction effects of adrenaline and noradrenaline, thus reducing blood pressure.

As a consequence of the use of drugs that target the ADRA1A gene having the effect of reducing blood pressure, it is plausible that there may be genetic variants in the region of this gene that are positively associated with hypertension/blood pressure. A large collaborative GWAS attempting to find loci associated with blood pressure (Newton-Cheh et al., 2009) explicitly examined the association of common variants in the region of the ADRA1A gene due to the plausible association. However, no results exceeded chance expectations. Other smaller fine-scale studies (e.g. Gu et al. 2006) have, however, identified marginally significant, albeit of small effect size, associations with variants in the ADRA1A gene with the outcome of hypertension.

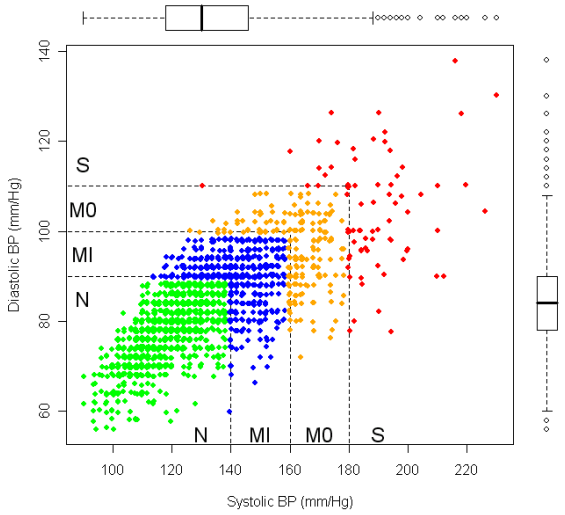
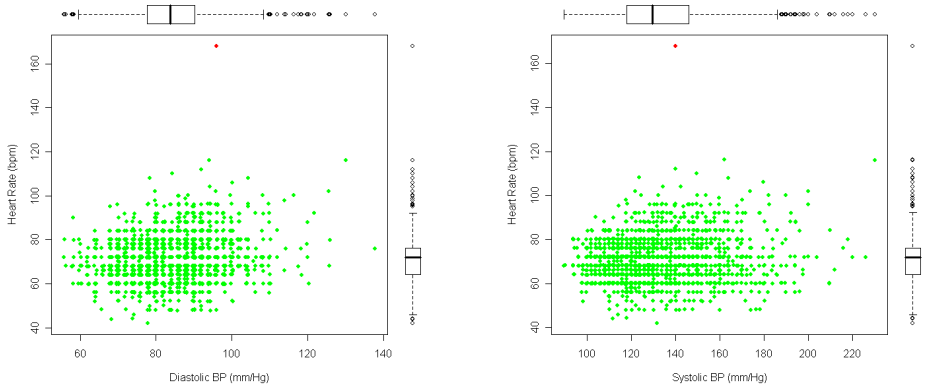


Figure 6.1: Correlation of systolic and diastolic blood pressure readings. The hypertension classifications are illustrated on the plot with the following categories: N (Green) = Normotensive, MI (Blue) = Mild Hypertensive, MO (Orange) = Moderate Hypertensive, S (Red) = Severe Hypertensive.

The data available for analysis in this case study are the genotypes of individuals from the PAMELA (Pressioni Arteriose Monitorate E Loro Associazioni) study, involving a random sample from the Monza region of northern Italy. In addition to the genotypes, continuous phenotype measurements of resting Heart Rates (HR), Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) have been recorded. The recorded blood pressure measurements are taken as the average of three readings over the course of an appointment with a trained physician, in order to try and reduce the within-patient variability of these measurements. Each measurement can be shown to be approximately normally distributed, as required for the analysis that follows. Further details of the PAMELA study and the data involved can be found in both Mancia et al. (1995) and Padmanabhan et al. (2010).

Individuals with missing/untyped genotypes were removed from the analysis, and individuals were also removed from analysis where a phenotype measurement had not been recorded. This resulted in a sample of 1895 individuals with



(a) Correlation of diastolic BP and heart rate

(b) Correlation of systolic BP and heart rate

Figure 6.2: Correlation of blood pressure and heart rate measurements. The five individuals with valid blood pressure readings but no available heart rate information are omitted from the plots.

genotypes and both systolic and diastolic blood pressure readings, however, only 1890 of this sample also had a valid heart rate measurement collected. Each individual had 70 SNPs genotyped, covering 132.045 kb of the region involving the ADRA1A gene.

Figure 6.1 illustrates the relationship between systolic and diastolic blood pressure readings for the available subjects. It can be seen that there is a reasonably strong significant positive correlation between the two measurements ($r = 0.74$, p-value $< 2.2 \times 10^{-16}$), and therefore using a single measurement that takes into account both the systolic and diastolic readings in combination may be reasonable. The colouring of the points in figure 6.1 represents a hypertension classification that is commonly used as a measure that combines the blood pressure readings into discrete categories. The classification of hypertension is given according to the World Health Organization Guidelines (1999) and Williams et al. (2004).

The relationships between heart rates and the blood pressure measurements is illustrated in figures 6.2(a) and 6.2(b). The correlations between HR and SBP ($r = 0.11$, p-value $= 1.0 \times 10^{-6}$) and between HR and DBP ($r = 0.16$, p-value $= 9.6 \times 10^{-13}$) are both significantly different from zero, albeit relatively low, suggesting that combining these values into a single measurement may not be

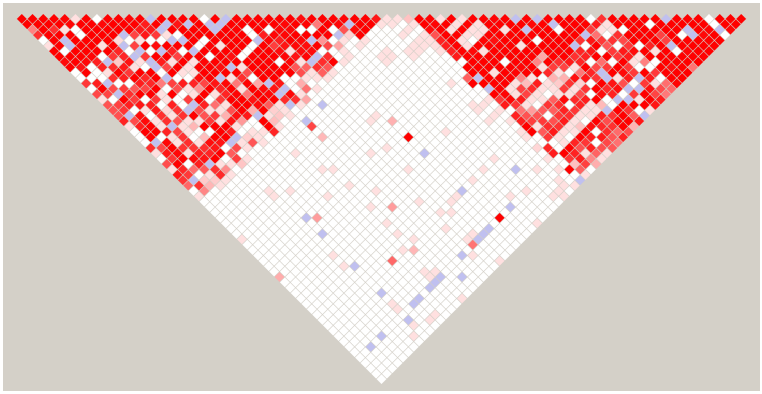


Figure 6.3: Linkage plot for individuals with systolic and diastolic readings.

appropriate. It is though evident from the plots that there is one individual, coloured red, with an unusually high heart rate measurement that could be the result of the data being incorrectly recorded. The analysis of this chapter assumes that the value has been incorrectly recorded, although if this assumption is not made and the data is analyzed including this point, the same general conclusions can be made (results not shown).

6.2 Linkage Plot

Initial phasing of the data using fastPHASE (Scheet and Stephens, 2006) results in 1252 haplotypes being estimated to cover the 132.045 kb of available DNA for the systolic/diastolic BP data sets, and 1276 haplotypes being estimated for the heart rate data set. This represents an unreasonably large number of haplotypes for the 70 SNPs of the data set, and so the data should be further explored to determine if there is an approach that can be taken to reduce the number of unique haplotypes.

Figure 6.3 shows a linkage plot for the systolic and diastolic blood pressure data sets. The linkage plot produced from omitting individuals who have systolic/diastolic measurements but no reading for heart rates is extremely similar to this plot (figure not shown). The shading of the plot is the default of the Haploview (Barrett et al., 2005) program, with the interpretation of the colours the same as that which was discussed previously in section 4.5.1.

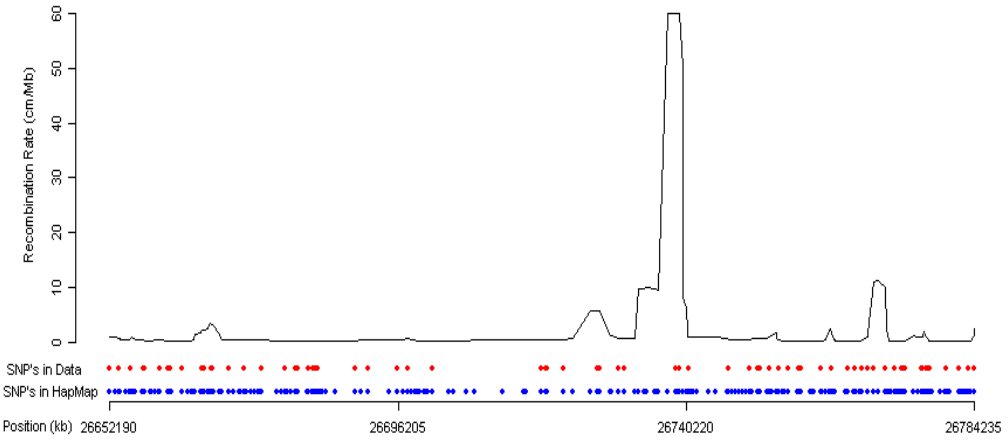


Figure 6.4: Recombination estimates from HapMap, and the SNPs typed in the study (red dots) and those in the HapMap CEU population (blue dots).

There appears to be two distinct high linkage regions shaded red, whilst linkage between the blocks is low, as reflected in the predominantly white shading. This would indicate that there are two haplotype blocks separated by a potential recombination hotspot (Gabriel et al., 2002). The association is not however perfect within each haplotype block, suggesting that there may be other recombination events occurring at a lower rate within these regions, or that there are sites exhibiting a degree of homoplasy. For Treescan-based methods to be appropriate, haplotypes should be found for areas of high linkage between SNPs since an underlying tree model of the haplotypes is assumed, and for this reason it is important to determine where recombination hotspots may exist.

6.3 Estimating Recombination Rates & Hotspots

6.3.1 HapMap

The PAMELA data does not include SNPs typed for all of the 237 SNPs that have been defined in the HapMap project (International HapMap Consortium, 2005).

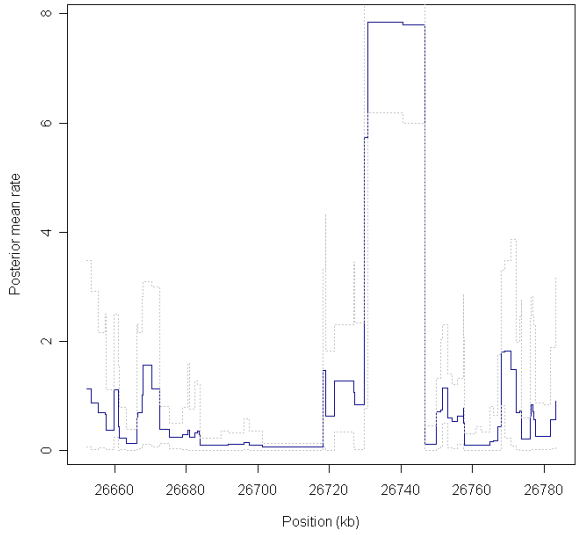


Figure 6.5: Recombination estimates directly from data.

It is however hoped that most of the essential features of the ADRA1A region can be captured by the 70 SNPs that are available in the data set. Figure 6.4 shows the pointwise recombination estimates given from the data in the HapMap project (International HapMap Consortium, 2010), which were calculated using the *interval* application from the LDHat (McVean et al., 2002) set of programs. The 70 SNPs in the data set (red dots) seem to capture most of the main features of the pattern of recombination, namely the three areas where there are peaks of recombination rates. There is, however, inevitably some comparative lack of detail, in particular there is a lack of information that could separate the two distinct peaks between the positions at approximately 26732.950 and 26739.722 kb.

6.3.2 LDHat

The *interval* program is also used to obtain recombination estimates directly for the 70 SNPs that were typed in this study. Due to the composite likelihood method requiring products of 2-SNP likelihoods across all pairs of SNPs, there is substantial computational time in estimating recombination rates between SNPs.

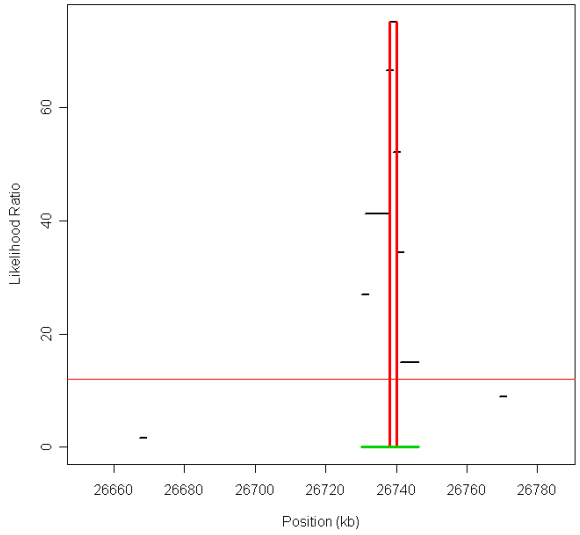


Figure 6.6: Hotspot determination by *sequenceLDhot*.

In order to lower the time required, various lookup tables are provided in the LDHat package which can be used instead of calculating a table for all pairwise comparisons when the sample size is large. fastPHASE estimated 1252 possible haplotypes when considering the full range of data, however, the largest lookup table available is for a sample size of 192 haplotypes. Generating a lookup table for this number of haplotypes would require an unreasonably long period of time, and so alternative strategies of handling the large number of haplotypes must be considered.

It has been chosen to sample the potential haplotypes according to their frequency of occurring in the data set so that the lookup table provided by the program can be used. Figure 6.5 shows the recombination estimates for the ADRA1A region using the available SNP data. It can be seen that the main features of the region are captured by the SNPs typed in the study, with the area of high recombination being evident.

Although the outputs from LDHat and the Haploview map strongly indicates that there is a recombination hotspot existing near SNPs rs4416829 and rs1390512 (positions 26732.950 to 26739.722 kb), the existence and location of any hotspot can be verified using the *sequenceLDhot* method (Fearnhead, 2006).

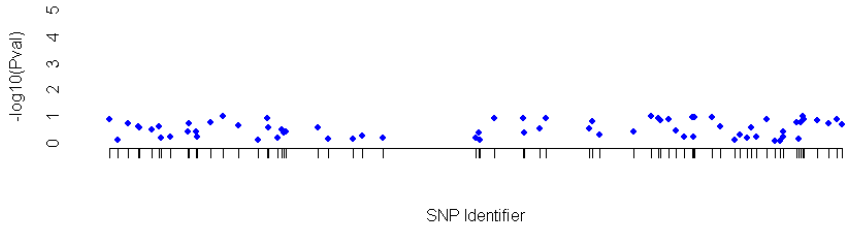
sequenceLDhot explicitly determines the locations of recombination hotspots, by using an approximate marginal likelihood method and performs likelihood ratio tests for each possible hotspot position, details of the method of which can be found in the paper of Fearnhead and Donnelly (2002). The results of applying *sequenceLDhot* to the ADRA1A data set are shown graphically in figure 6.6, where it can be seen that the recombination hotspot noticed previously is again evident.

Figures 6.3-6.6 show that the position of the recombination hotspot that has been identified covers SNPs at postions 37 and 38, and so the data is subsequently split into two haplotype blocks of low recombination, from SNP positions 1 to 36 and from SNP 39 to 70. This results in the total number of unique haplotypes reducing to 309 haplotypes for the systolic/diastolic data set, and 299 for the heart rate data. A problem with the haplotype or Treescan approaches is however illustrated, as it is not possible to include SNPs at positions 37 and 38 without substantially increasing in the number of unique haplotypes, and thus lowering the ability to detect any potential causative effects.

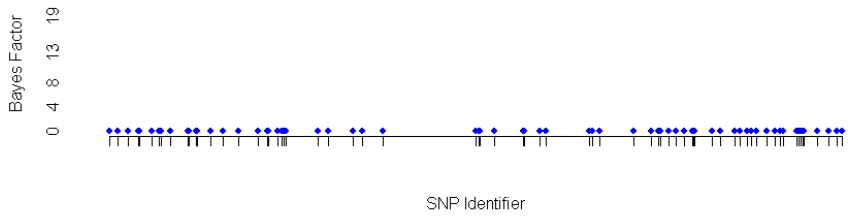
6.4 Categorized Analysis

In order to compare the analysis based upon the continuous measurements of systolic and diastolic blood pressure with the analysis performed by the WTCCC (2007), the data is dichotomized into hypertensive and normotensive classes. The classification of hypertension is given according to the World Health Organisation Guidelines World Health Organization (1999) and Williams et al. (2004), whereby a subject is classified as hypertensive with either a systolic BP reading of over 140, or a diastolic BP of over 90, or both. Figure 6.1 illustrates the hypertension classifications. This results in 799 people in the study being classified as hypertensive to some degree, and 1096 being classified as normotensive.

Hypertension is widely studied and is known to be a risk factor for a wide range of conditions, such as stroke, cardiac failure and renal conditions (Korner, 2007). Although hypertension is widely used as a categorical variable, the fact that hypertension is actually defined by two continuous measurements could lead to a substantial loss of information when an individual is classified as either



(a)

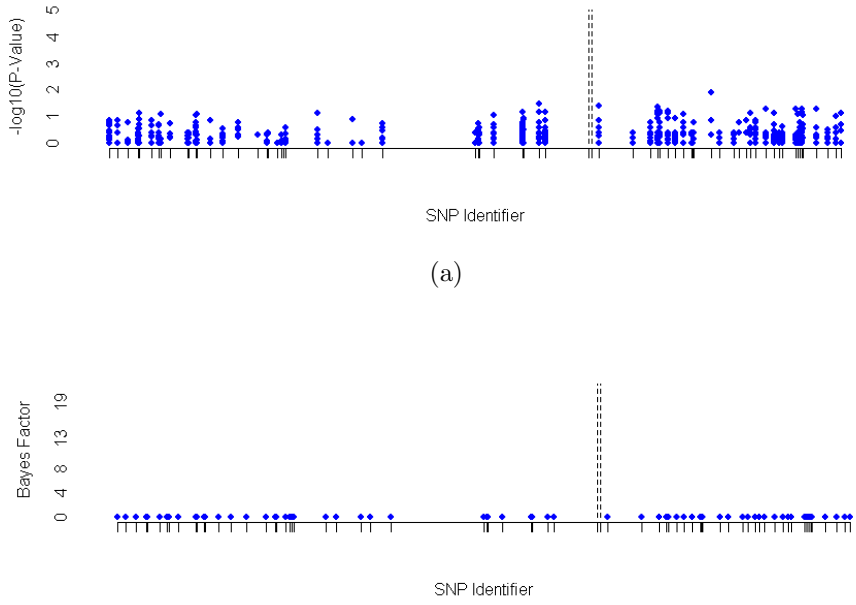


(b)

Figure 6.7: Single SNP analysis of the dichotomized hypertension data in frequentist (a), and Bayesian (b), settings.

hypertensive or not. A further complication to the use of hypertension as a categorical variable is the definition as to what constitutes hypertension, and indeed this definition has differed slightly over time (Korner, 2007). It could be argued that truly continuous variables should not be dichotomized (Senn, 1997), however, a two-state classification can be useful in practice in providing guidelines about levels of blood pressure that may be clinically significant.

In order to analyze the data in a categorical format, different methods are required compared to the analysis of continuous data. Chi-squared analysis can be used to calculate p-values to test for an association between the hypertension status and the genotypes at each SNP, or in a manner analogous to the Treescan method described in section 1.5.2. A Bayes factor equivalent to each of these methods can also be performed. Section 3.1 gave details of the Bayes factors that can be appropriate for analyzing categorical data. For the analysis that has been done here, a prior sample size for the Beta distribution of 5 has been



(a)

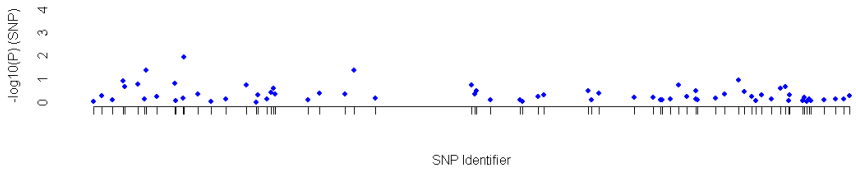
(b)

Figure 6.8: Treescan-based analysis of the dichotomized hypertension data in frequentist (a), and Bayesian (b), settings, with the recombination hotspot indicated by dashed lines.

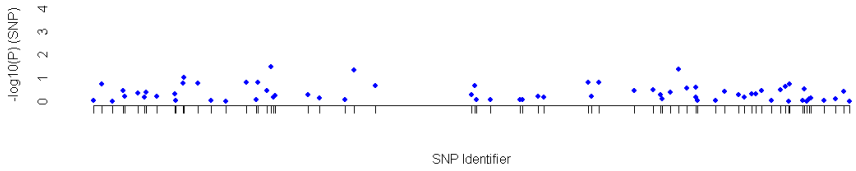
chosen. If the Uniform prior distribution suggested by Balding (2006) is instead used, this can result in some marginally higher Bayes factors, in particular for the Treescan-based analysis. This is as a result of very small groups of data being adjudged to be differentiated from the remaining data, whereas the the Beta prior can dampen down such effects that may be unlikely to be true associations.

Figures 6.7 and 6.8 illustrate, respectively, the Single SNP and Treescan analysis of the data, after it has been categorized according to hypertension status. The analysis that has been performed is for comparing the normotensive individuals to those with any degree of hypertension. Analysis can also be done comparing the normotensive group to individuals with severe hypertension, in order to find any differences between the two most extreme groups. Very similar conclusions are reached when the data is coded in this way (results not shown).

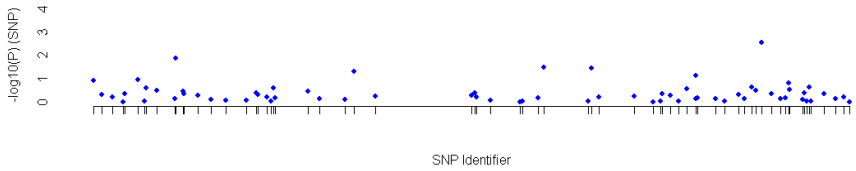
It can be seen that irrespective of the choice of analysis, there is very little



(a)



(b)



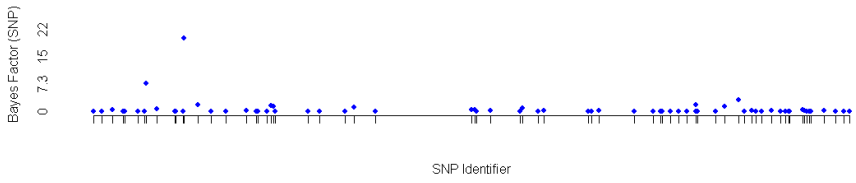
(c)

Figure 6.9: Standard single SNP tests for systolic (a), diastolic (b), and heart rate (c) phenotypes.

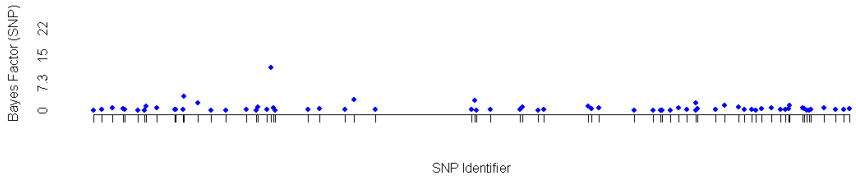
suggestion that there are any associations between the genotypes and hypertension status. However, the dichotomizing of the data has resulted in a huge loss of potentially useful information, and so any true associations between the genotype and the phenotypes may have been missed. The following sections treat the outcomes as separate continuous measurements, and so may be more powerful at finding any potential causative mutations.

6.5 Single SNP-based Analysis

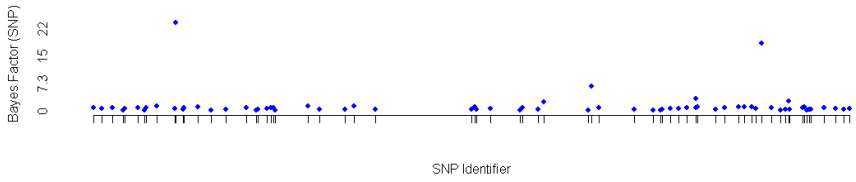
The single SNP analysis method has the advantage over the Treescan-based analysis in that it does not require the data to be phased and does not have issues in



(a)



(b)

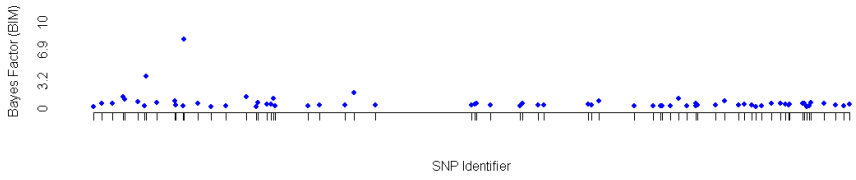


(c)

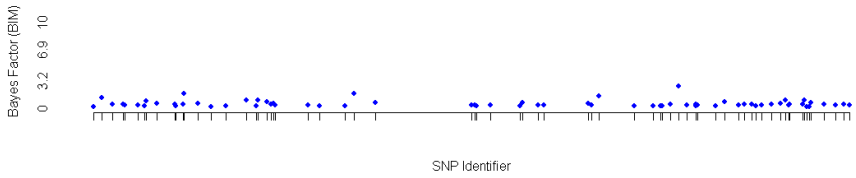
Figure 6.10: Bayesian single SNP tests for systolic (a), diastolic (b), and heart rate (c) phenotypes.

determining if there are recombination hotspots within the data. However, ignoring the tree structure could result in alleles being incorrectly grouped together if there is a large degree of homoplasy. Results of the frequentist single-SNP based approach is given in figure 6.9, and the corresponding Bayesian single-SNP analysis is illustrated in figure 6.10.

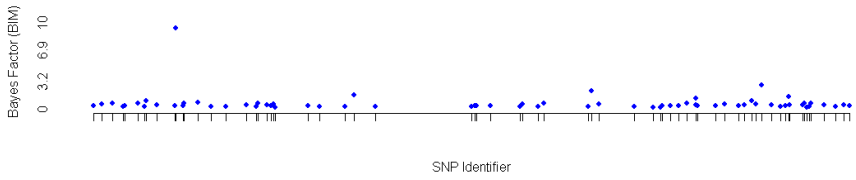
A Bonferroni correction to the 5% significance level applied to the 70 SNPs observed in this data set results in a threshold of 7.14×10^{-4}. This is considerably higher than a typical threshold used in GWAS, such as the critical level of 5×10^{-7} used in the WTCCC(2007) study. The lowest p-value that was observed in either of the three clinical measurements in the Bonferroni analysis was 0.0028, and this is substantially higher than the required significance level of 7.14×10^{-4}. As such,



(a)



(b)



(c)

Figure 6.11: BimBam single SNP tests for systolic (a), diastolic (b), and heart rate (c) phenotypes.

based upon a Bonferroni-corrected analysis, it would be concluded that none of the SNPs have a significant effect on any of the three clinical measurements in this data set.

As the uncorrected p-values for the single SNP analysis are identical to the p values obtained for the Bonferroni analysis, the lowest p-value obtained is again 0.0028 for the SNP rs11782159. This SNP results in the lowest corrected p-value, using the permutation procedure as detailed in sections 1.5.2 and 4.4, of 0.206. As this corrected p-value is higher than the threshold of 0.05 that is required for a SNP to be declared significant in this analysis, it would again be concluded that none of the SNPs have a significant effect on any of the clinical responses use in this data set.

Applying the Bayes factor method to the equivalent groupings used in the single SNP and Bonferroni-based methods, results in the 6 SNPs in table 6.1 being found that have a Bayes factor larger than 5. The Bayes factors have been calculated based upon the default settings of the PheGe-Sim program, with the hyperparameters being set according to the sample mean and variance of each data set. It can be seen in table 6.1 that the size of the Bayes factor varies considerably according to whether the additive, recessive/dominant or general model is used. For example, for the SNP rs17055925, there would appear to be very little effect if either the general or the recessive/dominant model is used, but an additive model suggests that there is some slight evidence of an effect of the SNP on the systolic blood pressure measurement. The means and size of each of the groups is shown in table 6.1, and it can be seen that, although the difference is small, there does seem to be an increasing blood pressure measurement with each copy of the B allele. In practice, however, the differences in effect size are likely to be too small to have any meaningful consideration in terms of the health of an individual.

The Bayes factors of BimBam were also calculated for each SNP individually, with the priors chosen being the default priors used by BimBam. Although consistently lower, the pattern of results (figure 6.11) is similar to the Bayes factors computed by PheGe-Find, with similar indications that there may be small effects at some of the SNPs.

6.6 Treescan & Haplotype Analyses

Analyses based on the Treescan methods are carried out separately on the two haplotype blocks of data; from SNPs rs4732845 (1) to rs11779546 (36), and from SNPs rs4732908 (39) to rs537220 (70). A tree is constructed for each block, using either the parsimony, fitch or maximum likelihood methods. The results are given for the parsimony method in figures 6.12 and 6.13, although the use of either of the other two construction methods gives comparable results. If there was no recombination and the infinite-sites model assumption was true, the results from the Treescan-based methods would be identical to the equivalent single SNP method. However, the trees constructed show a high degree of homoplasy

Table 6.1: Selected Single SNP results (above the line), and Treescan-based results (below the line) of the ADRA1A data. Summaries are also given of the counts and means of each of the three possible genotype groups from the first round of splits.

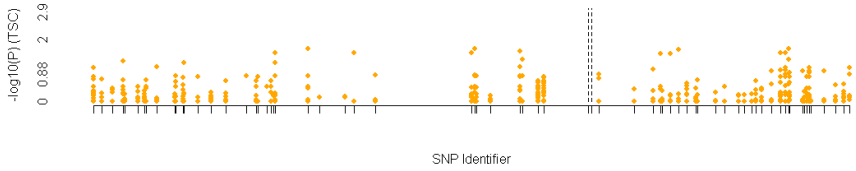
SNP identifying label	nAA	nAB	nBB	μ_{AA}	μ_{AB}	μ_{BB}	Clinical Score[a]	Additive BF	Dom/Rec BF	General BF	BimBam BF	Additive p-Value[b]	General p-Value Uncorrected	P-Value Corrected[c]
rs4732853	39	464	1387	68.5	70.2	71.5	H	20.2	9.5	7.2	9.5	0.0034	0.0134	0.614
rs11782159	234	856	800	70.4	72.0	70.4	H	0.7	6.2	19.4	2.8	0.1798	0.0028	0.206
rs11779546	68	627	1195	69.3	71.9	70.8	H	0.5	1.4	2.4	1.4	0.3584	0.0331	0.898
rs17055925	4	183	1708	115.5	129.2	133.4	S	10.5	9.1	6.5	1.8	0.0042	0.0116	0.574
rs11776470	3	154	1738	130.0	128.8	133.3	S	4.8	6.0	5.5	1.2	0.0132	0.0419	0.938
rs3739216	15	253	1627	77.2	84.5	83.9	D	0.5	2.8	1.2	1.9	0.7839	0.0322	0.916
rs11776470	1	77	1817	124.0	126.6	133.2	S	4.5	8.6	7.9	5.3	0.0070	0.0261	0.906
rs17055923	25	368	1497	67.6	70.1	71.4	H	11.9	5.3	4.7	6.0	0.0064	0.0211	0.869
rs4732853	25	373	1492	67.6	70.2	71.4	H	9.1	3.9	3.6	4.7	0.0090	0.0276	0.917

[a]H = Heart rate, S = Systolic blood pressure, D = Diastolic blood pressure

[b]Additive p-values have been calculated separately using regression, however, they are not used for the corrected p-values as only one test for each SNP is possible in the frequentist approach

[c]P-Values corrected according to all 70 SNPs for the single SNP analysis, or within the respective haplotype block for Treescan-based analysis.

(a)



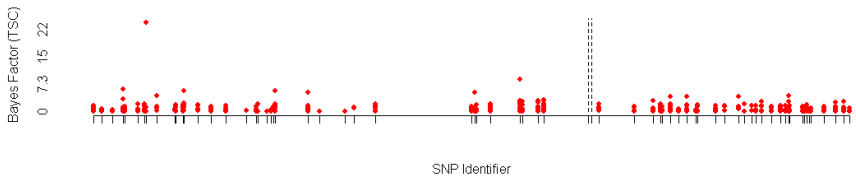
(b)

(c)

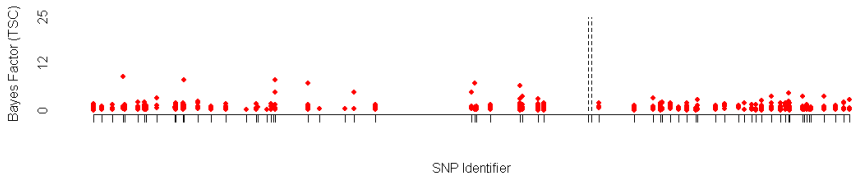
Figure 6.12: Standard Treescan results for systolic (a), diastolic (b), and heart rate (c) phenotypes, with the recombination hotspot indicated by dashed lines.

with SNPs mutating in many positions on the tree, and this results in different conclusions compared to the single SNP-based analysis. Only one of the three SNPs found with a Bayes factor greater than 5 (or corrected p-value less than 0.05) was also found with comparable criteria in the single SNP analysis, albeit with a stronger signal than was observed in the single SNP Bayes factor method.

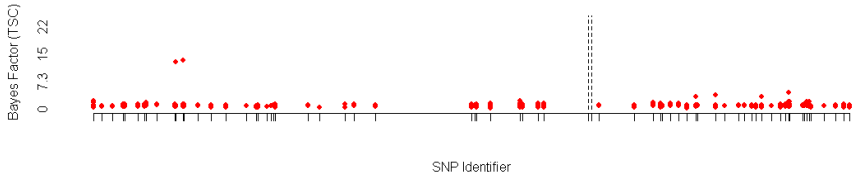
The large number of repeat mutations on the reconstructed haplotype trees make it difficult to ascertain which SNPs may indeed affect the phenotype responses. In addition, even if a SNP is adjudged to be associated with a response, it can be difficult to determine if this is an artificial feature resulting from the tree that was constructed. The strength of associations according to haplotype-based methods can also be assessed. However, because of the large number of



(a)



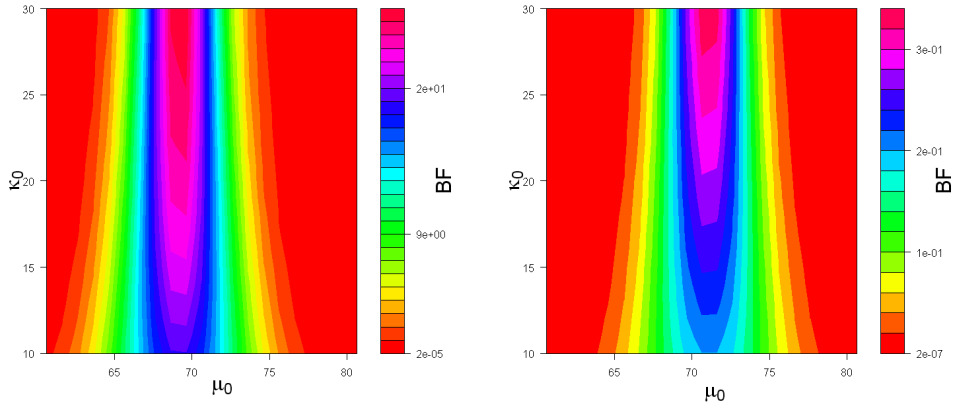
(b)



(c)

Figure 6.13: Bayes factor Treescan results for systolic (a), diastolic (b), and heart rate (c) phenotypes, with the recombination hotspot indicated by dashed lines.

low frequency haplotypes, there is low power to detect any differences and as such no strong effects are found. In the original Treescan paper, haplotypes of frequency less than five were removed in an attempt to decrease the risk of falsely reconstructing haplotype trees, and therefore increase the chance of finding causative associations. Such analysis has been applied to the ADRA1A data set, but there was also insufficient evidence of any associations based upon the resultant p-values and Bayes factors (results not shown).



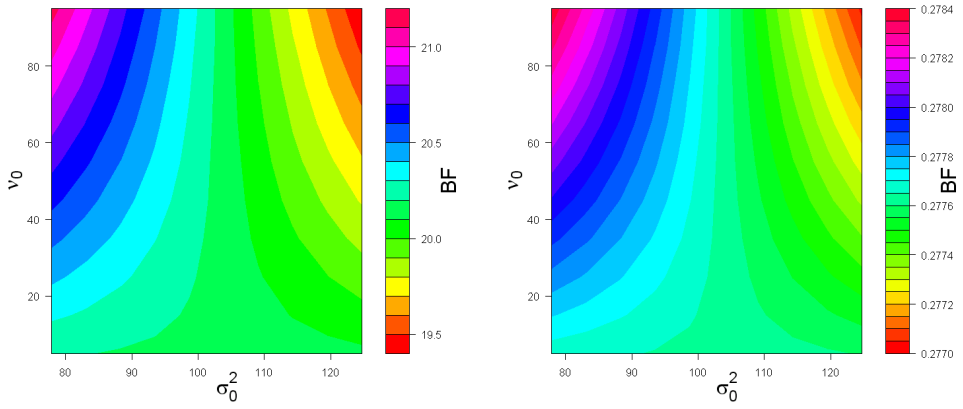
(a) SNP rs4732853 (11) (b) SNP rs1048101 (23)

Figure 6.14: Sensitivity to hyperparameters of overall mean.

6.7 Sensitivity Analysis

The sensitivity of the Bayes factors to the prior hyperparameter choices should be assessed, as it may be that the Bayes factors discussed above could be highly dependent on these choices. The prior values that have been used in the analysis are the default values of PheGe-Sim, however, sensible prior information from other sources could also be used for this data set.

The parameter μ_0 could be chosen according to guidelines about average values for the clinical measurement in question. For the systolic blood pressure a value of 120 mmHg would be reasonable, as this is defined as being approximately the upper level of the optimal category of systolic blood pressure (Williams et al., 2004; World Health Organization, 1999). The choice of this value is however subject to debate as it is also known that the blood pressure of an individual will be related to other non-controllable factors, such as age and gender. For similar reasons to the systolic blood pressure measurements, the prior mean for a diastolic blood pressure can be chosen to be the upper value of the optimal category of 80 mmHg. Resting heart rate measurements will also be subject to various uncontrollable factors, although a value of 70 bpm is reasonable as this represents the approximate average reading of a healthy adult (British Heart Foundation, 2010; American Heart Association, 2010). Analysis has been carried out using



(a) SNP rs4732853 (11) (b) SNP rs1048101 (23)

Figure 6.15: Sensitivity to hyperparameters of within-group variance.

these hyperparameter choices in place of the default setting for μ_0 in PheGe-Sim, and the patterns of association are similar for each of the SNPs being assessed. There are, however, slight differences in the Bayes factors, with the Bayes factors tending to be marginally lower than when using the values set according to the summaries of the sampled data.

The effects of the μ_0 and κ_0 hyperparameters for the mean are assessed together in figure 6.14, for two SNPs and their association with the heart rate phenotype. It is evident that the patterns of association are similar for both a SNP that appears to have some association with the phenotype (6.14(a)); and for a SNP with little apparent association (6.14(b)). It can be seen that the Bayes factors are strongly affected by the μ_0 parameter, and that choosing a prior value far away from the mean of the data can result in a substantial reduction of the Bayes factor. The hyperparameter κ_0 has only a relatively small influence on the Bayes factors, due to the large sample size of this data set.

Figure 6.15 illustrates the sensitivity of the Bayes factors to the ν_0 and σ_0^2 hyperparameters of the within-group variance σ^2, for the same two SNPs as used in figure 6.14. As the prior for the mean has been centred on the sample mean, reducing the σ_0^2 value indicates more confidence in the location of the distribution and therefore results in higher Bayes factors. The ν_0 parameter has a relatively small effect, but increasing it will steadily increase the Bayes factors due to this

indicating an increase in the belief about the accuracy of the prior distribution. If, however, the prior mean of the data were to be far enough away from the true data mean, then the Bayes factors would be increased by increasing the variance, so that the true mean is covered by the prior distribution.

Chapter 7

Results from Simulated Data

In this Chapter, the PheGe-Sim application of Chapter 2 is used to simulate genotype-phenotype data on which the methods of association detailed in Chapter 4 can be applied. In order to choose suitable parameters for the construction of the simulated data, the data set of Chapter 6 can be used to provide reasonable initial estimates. The details of the realism involved in the choices of coalescent parameters is though not of primary concern, as the most important consideration is that the generated data sets will display similar properties to that of real data with appropriately determined phenotypes. The estimates from the real data are therefore adjusted in an attempt to more closely mirror the true patterns of the resultant linkage plots.

7.1 Parameter Estimates

The choice of parameters for the simulations are intended to reflect the real data sets in Chapters 5 and 6, and as such, where possible, initial estimates are taken from these data sets. It should however be considered that, for both the estimation of the mutation rate and the parameters of the gamma distribution, estimates from the data are liable to underestimate the true values since they do not take into account the potential homoplasy involved. The primary consideration is to get a linkage plot that is similar in appearance to that of real data, and for a similar number of segregating sites (approximately 40) to be involved

for each variation in parameter choice.

7.1.1 Estimation of Mutation Rates

The choice of the mutation rate θ for use in the simulations is dependent on the other input parameters that can be chosen, as a change in the finite-sites distribution (section 7.1.3) or the rate of population expansion (section 7.1.4) will necessitate a different mutation rate to obtain the same number of segregating sites on average. The simulations are intended to generate approximately the same number of segregating sites as the real data sets, whilst also displaying similar patterns of linkage between these sites.

In order to estimate the mutation rate for each of the real data sets, the Watterson estimator (Watterson, 1975) can be used;

$$\hat{\theta}_W = \frac{S_N}{\sum_{i=1}^{n-1} \frac{1}{i}}, \tag{7.1}$$

where S_N represents the number of segregating sites and $\sum_{i=1}^{n-1} \frac{1}{i}$ is the $(n-1)$th harmonic number. This can be implemented through the use of the *theta.s* function of the R package *ape* (Paradis et al., 2004). In Chapter 6 it is found that there is strong evidence of a recombination hotspot existing in the ADRA1A data set. To avoid a misrepresentative mutation rate for the simulation of one haplotype region of low recombination, the estimate is therefore obtained separately for each haplotype block either side of the recombination hotspot. The estimators

Table 7.1: Mutation rate estimates. Simulated data results are averages over all the 2400 simulations.

DataSource	Number of Sites	Number of unique haplotypes	Mutation Rate Estimate
ADRA1A LHS	36	196	6.15
ADRA1A RHS	32	113	6.04
D. melanogaster	22	41	5.14
Simulated data	37.39	59.19	8.04

for each of the real data sets are displayed in table 7.1.

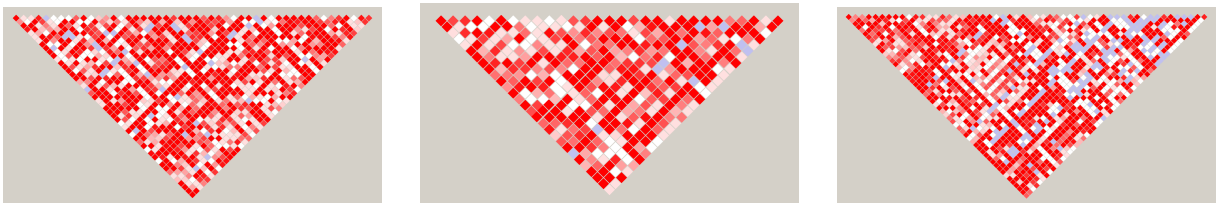
The estimates of the mutation rates are however likely to be underestimates of the true underlying mutation rates, as repeat mutations will not be accounted for by the estimates from the Watterson's estimator since it is derived assuming an infinite sites model. This situation is demonstrated in the simulated data sets, whereby a θ of 25 has been chosen to generate, in combination with the other parameters, representative linkage plots. Table 7.1 shows that the average estimate of the mutation rate by $\hat{\theta}_W$ is only 8.04, considerably less than the true value of 25. The estimates from the simulated data are slightly higher than those of the real data sets, but seem reasonable when evaluating the resultant linkage plots. It can also be seen from table 7.1 that there are a larger number of unique haplotypes than for the average of the simulated data. However, most of these haplotypes are of low frequency, and so there will be low power to detect differences between the phenotypes at these rare haplotypes.

7.1.2 Estimation of Recombination Rate

Estimation of the recombination rate ρ for the simulated data can be based directly upon estimates that are obtained using the real data sets of Chapters 5 and 6. The LDhat application by McVean et al. (2002) can be used to provide estimates of recombination rates for the *D. melanogaster* and ADRA1A data sets. A brief description of the method is given in section 5.3, and the resultant estimates obtained are displayed in table 7.2. Estimates were not obtained for the

Table 7.2: Recombination rate estimates. The 'area' is defined as the area underneath the estimated recombination rate plots for each data set (Chapters 5 and 6), and the 'length' is the length of the region of data being considered.

Data Source	Area	Length	Average Estimate
ADRA1A LHS	35.09	78.67	0.45
ADRA1A RHS	15.95	43.7	0.36
D. melanogaster	20.98	11.85	1.77



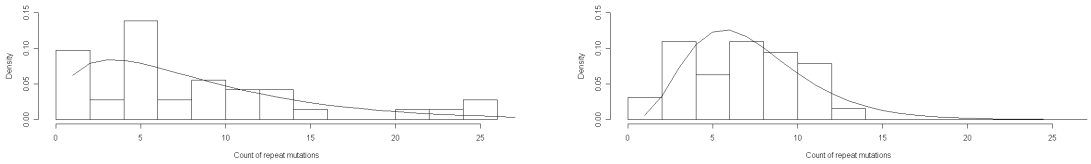
(a) $\rho = 0$ (b) $\rho = 0.5$ (c) $\rho = 2$

Figure 7.1: Simulated linkage plots for recombination rates of 0 (a), 0.5 (b) and 2 (c).)

simulated data set, as these rely on information being known about the relative site locations of each SNP, which has not been simulated, although this could in theory be done.

As noted in Chapter 6 and in the previous section, a recombination hotspot between two haplotype blocks of low recombination is found to occur in the ADRA1A data set. The simulations could be coded to mirror this effect by allocating a single position to be the location at which a large proportion of the recombination events occur. Although methodologically and computationally feasible, this approach would require a high recombination rate in an attempt to obtain the two near-independent haplotype blocks seen in the real data set. This would result in a large increase in simulation time, since the time taken to simulate even a single ARG substantially increases with a moderate increase in ρ. The effect of this is very similar to that of just pasting together two independent simulations as shown in section C.3. It was therefore decided to focus on creating a single haplotype block with relatively low recombination occurring uniformly along the gene segment. A recombination rate of 0.5 has been chosen to obtain, on average, 3.33 recombination events for each simulation, in order to obtain linkage consistent with the real data sets.

As with the estimation of θ discussed in section 7.1.1, the recombination estimate can be misinterpreted in the presence of a finite-sites model of mutation, as both features can result in the same apparent patterns of linkage. Figure 7.1 illustrates the linkage between sites that is obtained, using three different values of the recombination parameter. As expected, it appears that there is some tendency towards lower linkage between sites as the recombination parameter increases as



(a) Gamma Left (b) Gamma Right

Figure 7.2: Estimated Gamma distributions fitted to the ADRA1A data of Chapter 6; for the left (a) and right (b) of the recombination hotspot.

indicated by the increase in lighter coloured squares. The effect is however relatively small, and section C.2 shows that recombination alone struggles to recreate sufficiently realistic patterns of linkage.

7.1.3 Estimation of the Gamma Distribution for Finite Sites

Approximate values of the shape (α) and rate (β) parameters of the Gamma model for finite sites can be obtained directly from the real data sets of Chapters 5 and 6. The number of repeat mutations that occur on each reconstructed haplotype tree is counted, and the resultant distribution of the counts can be used to estimate the required parameters of the Gamma distribution. As with many of the other simulation parameters, the ADRA1A data set of Chapter 6 must be split into two haplotype blocks either side of the apparent recombination hotspot. If the hotspot is ignored, the Gamma distribution becomes extremely

Table 7.3: Estimates of the parameters involved in the Gamma distribution for finite sites. Simulated data results are averages over all the 2400 simulations.

Data Source	Shape ($\hat{\alpha}$)	Rate ($1\backslash\hat{\beta}$)	Scale ($\hat{\beta}$)
ADRA1A LHS	1.609	0.187	5.353
ADRA1A RHS	4.383	0.597	1.676
D. melanogaster	8.8	6.4	0.16
Simulated data	5.636	2.028	0.530

(a) Infinite sites (b) Reasonable rate of homoplasy (c) Rate of homoplasy too high

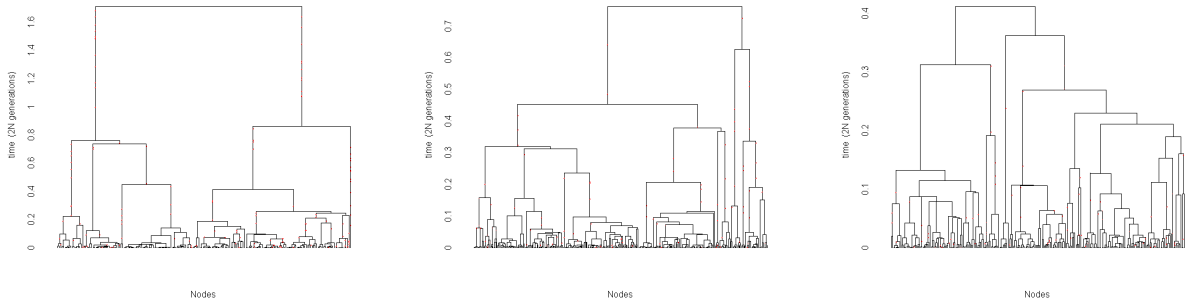
Figure 7.3: Simulated linkage plots for finite sites, as the Gamma model changes.

right skewed, with an unrealistically high number of repeat mutations being estimated to occur on any reconstructed haplotype tree. Figures 7.2(a) and 7.2(b) illustrate the distribution of mutation counts for the left and right-hand side of the recombination hotspot. Two of the SNPs from the data set are omitted, as their inclusion into either of the haplotype blocks results in the unrealistic situation of upwards of 50 instances of repeat mutations at a single site.

The R function *fitdistr* by Venables and Ripley (2002) is used to obtain maximum likelihood estimates of the (α) and (β) parameters, and the fitted distributions are shown with the solid lines on figure 7.2. A similar process was employed for the *D. melanogaster* data, although there are far fewer repeat mutations and a substantially lower sample size. The estimated parameters for the real data sets are given in table 7.3.

It can be seen that there is some difference between the shape and rate parameter estimates for either side of the recombination hotspot of the ADRA1A data set. The accuracy of the method of obtaining estimates of the parameters of the Gamma distribution through using the reconstructed haplotype trees can be checked using the simulated data sets. For the simulated data sets, the α and β parameters can be estimated using the same procedures as for the real data, and the averages of these over all the simulations are displayed in table 7.3. It can be seen that the estimates correspond closely to the values that were used for the simulations, whereby the shape and rate parameters were specified as 5.5 and 0.5, respectively.

The usefulness of the Gamma model of finite sites can be seen when comparing the linkage plots that can be obtained to those of an infinite sites model, as in



(a) $\beta = 0$ (b) $\beta = 10$ (c) $\beta = 50$

Figure 7.4: Simulated coalescent trees for various rates of population expansion. There are different scales on the y-axis for each of the plots, due to an increasing level of population expansion resulting in the coalescence process completing earlier. Changing the level of population expansion also changes the shape of the plots, with an increasing level of expansion resulting in a smaller ratio of the terminal and internal branch lengths.

figure 7.3. The model of infinite sites shown in figure 7.3(a) shows an overabundance of blue squares in comparison to that of both haplotype block regions of the ADRA1A data set. There is also a much more block-like structure imposed upon the linkage, and there are very few squares that are intermediate shades of red. The other extreme (figure 7.3(c)) is where the rate is too high resulting in far too many light red and white coloured squares, suggesting that the gamma distribution is set too high.

7.1.4 Estimation of the Population Expansion Parameter

The β parameter relating to a theoretical exponential population expansion cannot simply be directly estimated from real data, and is therefore determined, in combination with the other parameters, so that the simulated data most closely matches the real data. It is likely that there are complicated demographic events that have occurred in human history, such as migrations between populations and bottlenecks of population size. Other simulators of the coalescent process can accommodate such events, however, only a simple exponential population expansion is assumed in these simulations.

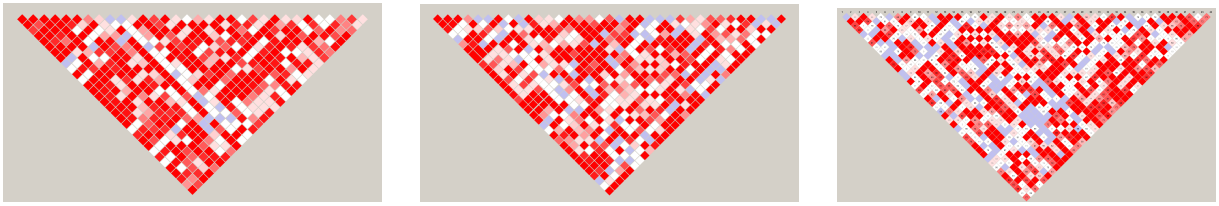
Figure 7.4 represents three realizations of the coalescent process (with no recombination) for different values of the population expansion parameter. Figure 7.4(a) illustrates the extreme scenario of no population expansion occurring at all, and it can be seen that the total length of the tree is dominated by long ancestral branches. This results in many of the mutations occurring on these branches and therefore there will be many mutations, particularly for an infinite sites model, that are perfectly correlated with each other (figure 7.5(a)).

The other extreme situation is that of a high rate of population expansion, as illustrated in the coalescent plot of figure 7.4(c) and the linkage plot of figure 7.5(c). It can be seen that the scale on the y-axis is shorter than for the situation of no population expansion, reflecting the decreased time (in coalescent units) that is taken to reach a most recent common ancestor. In addition, the terminal branches are longer in comparison to the internal branches than is the case when there is no population expansion. In an extreme scenario, the tree will become more and more star-like as the population expansion parameter increases, until all the lineages are effectively independent from each other. It can however be seen from figure 7.5(c) that even this level of population expansion ($\beta = 50$) results in SNPs becoming more independent from each other, as indicated by the increasing prevalence of white coloured squares in the linkage plot. A further consideration when increasing the value of β is that an increase in the mutation rate, and possibly also the number of terminal nodes, will be required to obtain approximately the same number of SNPs. For the examples in figure 7.4, table 7.4 shows different choices of the mutation rate θ, and number of terminal nodes n, for the different β values.

Table 7.4: Adjustment of mutation rate and the number of terminal nodes for different values of population expansion.

Figure	β	Mutation Rate (θ)	Terminal Nodes (n)
7.4(a)	0	10	250
7.4(b)	10	25	250
7.4(c)	50	50	500

The coalescent tree of figure 7.4(b), and the linkage plot in figure 7.5(b),



(a) $\beta = 0$ (b) $\beta = 10$ (c) $\beta = 50$

Figure 7.5: Simulated linkage plots for population expansion rates of 0 (a), 10 (b) and 50 (c).

illustrate a situation where the value of β has been set as 10. In this situation, in combination with suitable choices for the other parameters, the linkage plot most closely visually resembles that of the real ADRA1A data set of Chapter 6. This can also be seen when comparing the proportions of each colour present in the real and the simulated data sets.

7.1.5 Choice of Other Parameters

In order for the comparisons between the methods involved in finding causative mutations to be fair, all the association methods and tree construction options of Chapter 4 are tested on every simulated data set.

The choices as to how many causative mutations there are and their corresponding effect size(s) are allowed to vary, with the effect size being given as a percentage of the within group standard deviation, σ. The effect sizes chosen represent mutations that are; $20\%, 40\%, 60\%$ or 80% of the standard deviation. The effect of two causative mutations is also explored, with the effect size of each mutation being allocated to be the same. Under the finite-sites assumption, the effect of the choice as to whether a mutation is causative in all parts of the tree (AC), or only on a specific lineage (NAC), is also assessed. It would be expected that Treescan-based methods would perform better than single-SNP type analysis in the NAC situation, but would perform worse if the AC assumption is specified.

The realized number of segregating sites is a result of a complex relationship between many of the parameter choices, including: the mutation rate, the rate of population expansion, and the degree of homoplasy that has been specified.

Table 7.5: Summary of simulation parameters.

Variable	n	θ	β	ρ	Γ_α	Γ_β	$num.pop$
Value	250	25	10	0.5	5.5	0.5	2000

where n = Number of terminal nodes, θ = Mutation rate, β = Population expansion parameter, ρ = Recombination rate, Γ_α = Shape parameter of Gamma distribution, Γ_β = Scale parameter of Gamma distribution, $num.pop$ = Simulated population size

In addition, there is also the stochastic variation between different runs of the simulation. A change in any one of these parameters will require a change in the others to obtain the same average total number of segregating sites, and so the calibration of all the parameter choices must be made so as to obtain an average number of mutating sites similar to that of the real data sets. This has been achieved with the parameter choices used in the preceding sections, as the average number of SNPs for all the simulations that have been performed is found to be 37.39. This is slightly higher than the corresponding number of sites of the real ADRA1A data set, which has 36 and 32 segregating sites for the left and right hand sides of the recombination hotspot. However, this is, in part, due to there also being a slightly higher rate of mutation in the simulated data (section 7.1.1).

An additional parameter, n, must also be specified. This parameter controls the number of non-unique (in terms of distinct haplotypes) terminal nodes of the ARG, and as such its specification depends on the combination of the other chosen parameters. The resultant linkage patterns are however reasonably constant at different values of n, as each additional terminal node will on average account for only a relatively small increase in the total length of the ARG, since it typically attaches close to the tips of the existing ARG if n is sufficiently large and β is comparatively small. There will therefore only be a small chance of an increased number of mutations occurring that could affect the linkage within the haplotypes. For the simulations that follow a value for n of 250, in combination with the other specified variables, results in reasonable patterns of linkage.

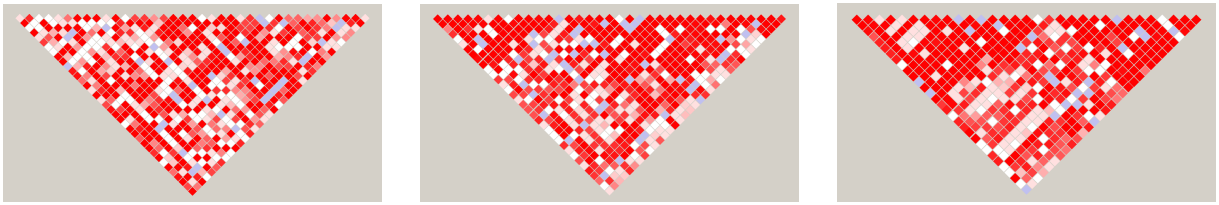
To remain consistent with the ADRA1A data set a sample size of 2000 is chosen for each of the simulations. Unlike the standard coalescent, whereby the sample size is used to construct the ARG for each terminal node, the sample size used in PheGe-Sim is used to sample the terminal nodes of the ARG according to the frequency with which they occur. This is more efficient computationally as less information is being retained in the construction of the ARG, in this case for 250 terminal nodes as opposed to the 2000 that the standard coalescent model would require. The resultant linkage plots in this approach are largely unaffected compared to the standard coalescent approach, as few mutations will be predicted to occur on the short distances of additional terminal branches.

The sample size used will have an impact on the time taken to run the simulations, particularly for the permutation procedure used in the single SNP method of association. Changes of the sample size would require alterations of the other parameters to obtain similar patterns of linkage, primarily altering the proportion of blue squares in the linkage plots that can arise through small sample sizes of particular SNP combinations. The power of all of the methods at detecting causative mutations will clearly increase with an increase in sample size, and correspondingly decrease with a reduction of sample size.

Table 7.5 summarizes the choices of parameters used for the simulations that follow. As noted throughout this chapter, different combinations of parameters could reasonably be used to obtain similar patterns of linkage. However, it will been shown that the parameters that have been chosen provide a good description of the real data.

7.1.6 Comparisons with Real Data

Although substantial efforts have been made in order to make the simulated data as realistic as possible, there are inevitably some differences between the analysis of the real and simulated data sets. The first difference is that the simulations have used *known* haplotypes, whereas the real data required phasing of the genotypes to obtain *inferred* haplotypes. The reasoning for the omission of phasing the simulated genotypes in the simulations is that, as commented on previously in section 1.5.1, the reconstruction of haplotypes is generally a reasonably accurate



(a) (b) (c)

Figure 7.6: Simulated linkage plots for the final parameters chosen for the simulations (a); and the linkage plots of the LHS and RHS of the ADRA1A data set.

procedure for tightly-linked SNPs. A more practical consideration in choosing not to phase simulated data is that the haplotype-estimation procedure can involve substantial computational time.

In addition to the assumption of perfectly phased haplotypes, there is also the assumption in the simulated data that the boundaries between distinct haplotype blocks have accurately been determined. In the real data set available to this study, the recombination hotspot and block-like structure of the haplotypes is clearly apparent, both visually in a linkage plot and through the criteria used when applying the program SequenceLDhot (Fearnhead, 2006) in determining hotspots. However, this will not always be the case for data sets, and in the available data there is little indication as to how accurate the hotspot determination procedure could be. In a hypothetical situation, it can be seen that the effect of joining together two entirely independent haplotype blocks will result in linkage patterns similar to that of the true observed data (section C.3).

A pragmatic view is taken of the parameter choices, as the specific choice of parameters is taken to be an issue that is not of interest in itself, but merely useful as a mechanism for producing simulations that reflects some features of

Table 7.6: Summary of colours used in linkage plots.

	$D' < 0.4$	$0.4 \leq D' < 0.7$	$0.7 \leq D' < 1$	$D' = 1$
LOD < 2	white	white	white	blue
LOD ≥ 2	green	orange	pink	red

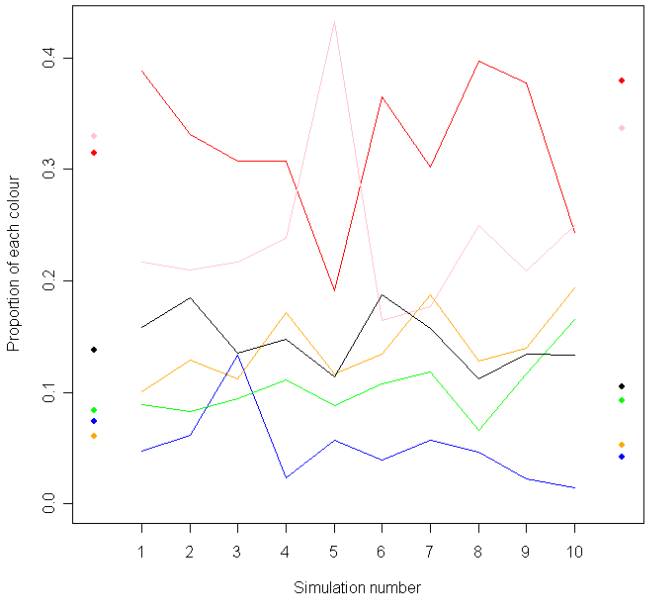


Figure 7.7: Percentages of each category of linkage for an example of ten simulated data sets. The true proportions for the LHS and RHS of the recombination hotspot are indicated at the LHS and RHS of the plot, respectively.

the real data. The linkage plot of figure 7.6(a) does however appear to support the belief that features of the real data are indeed being captured, as visually at least it appears that the simulated data set is similar in comparison to each of the haplotype blocks of the ADRA1A data set. A summary of the proportion of each colour present in the linkage plots can be obtained, and this can provide a further illustration of the similarities between the real and simulated data sets. Figure 7.7 illustrates the comparison between the averages of each colour (table 7.6) for each simulation, and the true proportion given in each haplotype block of the ADRA1A data set. Although broadly similar, there are some differences between the real and simulated data: most notably, in the simulations, the proportion of the pink group tends to be underestimated and the proportion of the orange group is slightly overestimated.

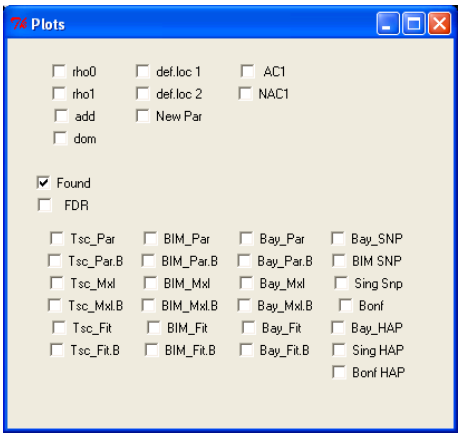


Figure 7.8: Rpanel for illustrating simulation results.

7.2 Results of Simulations

The various association methods, and approaches in determining if a mutation has been correctly identified, have been summarized through the use of another interactive Rpanel, as shown in figure 7.8. The use of Rpanel in this setting enables different approaches to be easily compared, and facilitates the collation of results of simulations that have been run on multiple computers. Figure 7.9 shows the key used for plotting the results of each method, where the colouring, line types and shape of the points have been chosen to represent the connections between the different approaches.

In addition to the parameters chosen above, the results of each association method is compared under four different criteria. In the plots of the results in section 7.2, the models of causative mutations used for each of the plots is as follows:

(a) Single 'All Causative' and Additive mutation

(b) Single 'All Causative' and Dominant mutation

(c) Single 'Not All Causative' and Additive mutation

(d) Single 'Not All Causative' and Dominant mutation

(e) Two 'All Causative' and Additive mutations

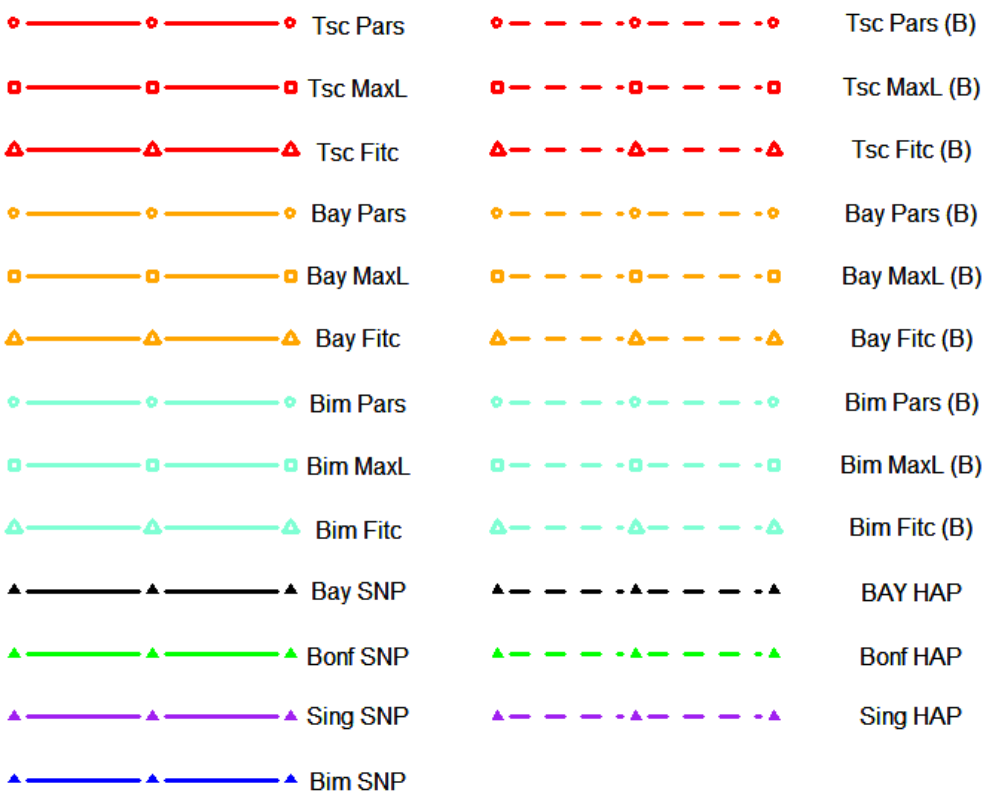


Figure 7.9: Plotting key for the results of the simulations, where the abbreviations used are: Tsc = Treescan, Pars = Parsimony, MaxL = Maximum Likelihood, Fitc = Fitch, Bay = Bayes factor approach, BIM = Bimbam approach, SNP = Single SNP method, Bonf = Bonferroni Correction, HAP = Haplotype association method, (B) = the 'branch correction' approach. Note that the found results for both the 'Treescan' and the 'Treescan branch' approaches will be identical, and thus only appear as one line in the relevant plots.

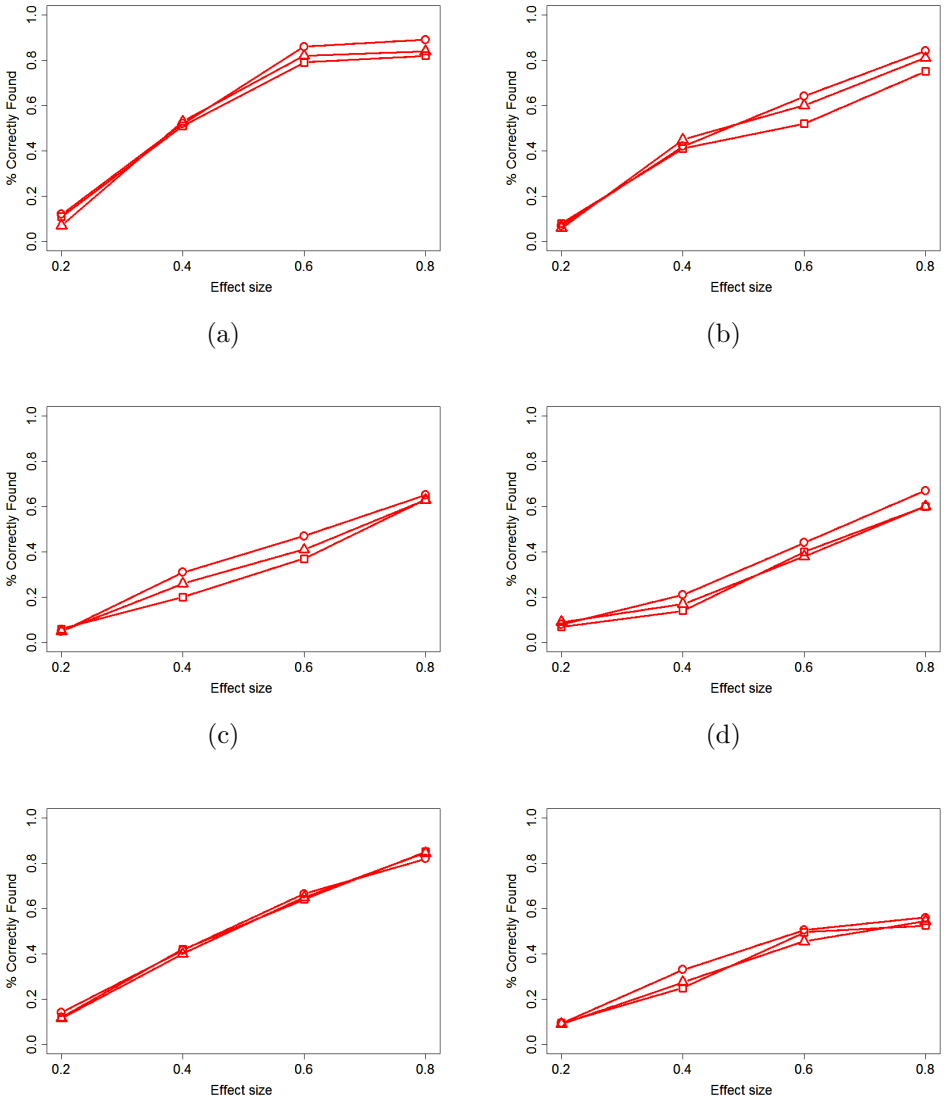


Figure 7.10: Proportion of correctly found SNPs across the six causation models, for the Treescan method when using the parsimony (circles), maximum likelihood (squares) and fitch (triangles) methods.

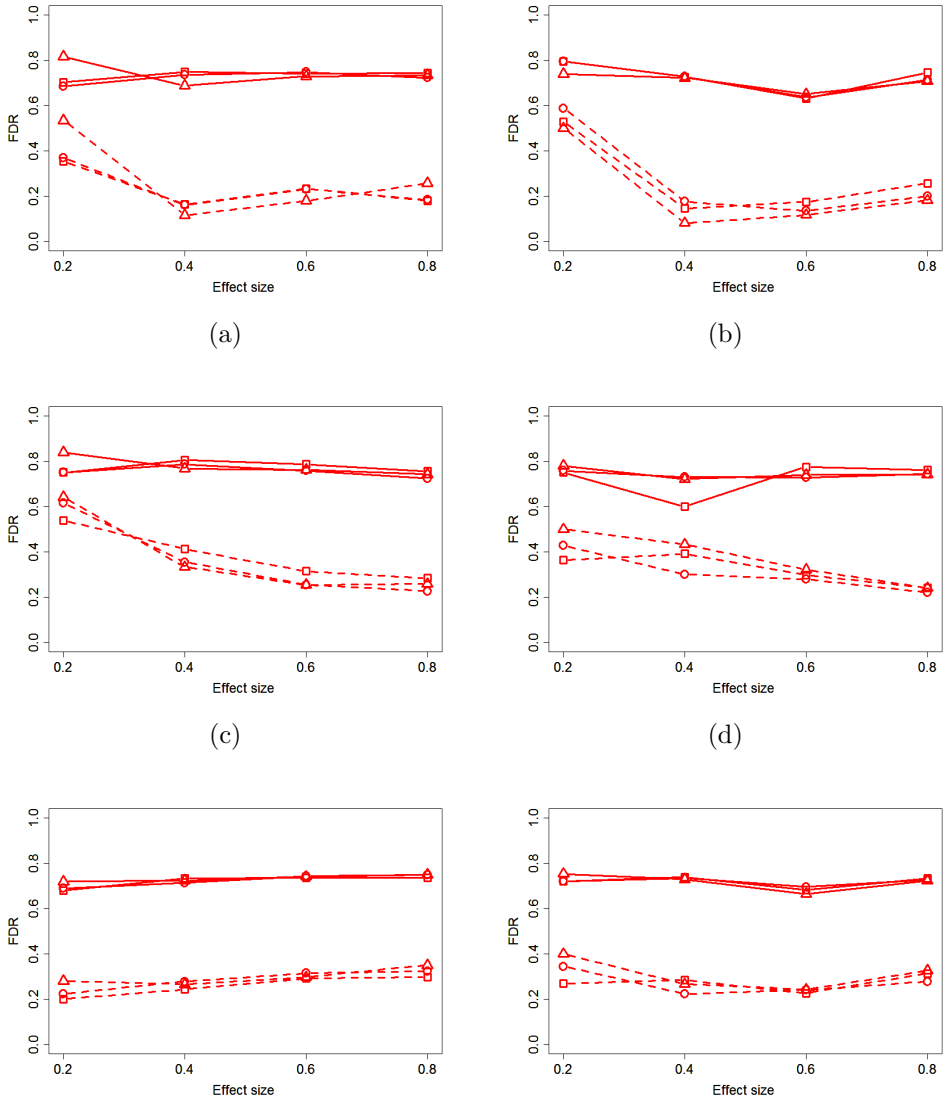


Figure 7.11: False discovery rates for found SNPs across the six causation models, for the Treescan method when using the parsimony (circles), maximum likelihood (squares) and fitch (triangles) methods.

(f) Two 'Not All Causative'and Additive mutations

Comparisons can be made across each of the different models of causative mutations. As would be expected, the ability of the methods at finding causative mutations increases according to the size of the causative mutation, and also increases when a mutation is allocated to be causative at all locations at which it appears on the ARG. In general, additive models appear to be easier to detect than the dominant models, as a result of there being fewer sampled individuals carrying both forms of a causative mutation. It is also true that, on average, there is slightly more difficulty at detecting two causative mutations as compared to the similar one causative mutation models.

The simulations were run on multiple computers using R version 2.9.1, along with the PHYLIP, Treescan and BimBam applications. Each computer used for the simulations were equipped with the Windows XP operating system, with 2.19GHz processor speed and 1GB of RAM. For each method and effect size 100 simulations were executed. The time taken for the simulations increased according to the effect size and whether one or two causative mutations were chosen, as this will cause more tests of two and three-way associations for both the Bayesian and Frequentist approaches. The approximate time for single simulation with all methods of association selected varied from about 30 minutes to 2 hours, depending on the chosen simulation parameters.

7.2.1 Tree Construction Approaches

Figure 7.10 illustrates the results of the correctly found mutations for each of the three construction methods applied to the original Treescan method. It can be seen that there is very little difference in the chance of finding the true causative mutations between the three phylogenetic methods for each of the six causation models that are displayed. In order for a fair comparison between the methods, the False Discovery Rates (FDRs) of each tree construction approach must also be taken into consideration. As with the percentages of correctly found mutations, the False Discovery Rates of each tree construction method are broadly in agreement with each other; regardless of whether the 'branch' or the SNP correction method is applied.

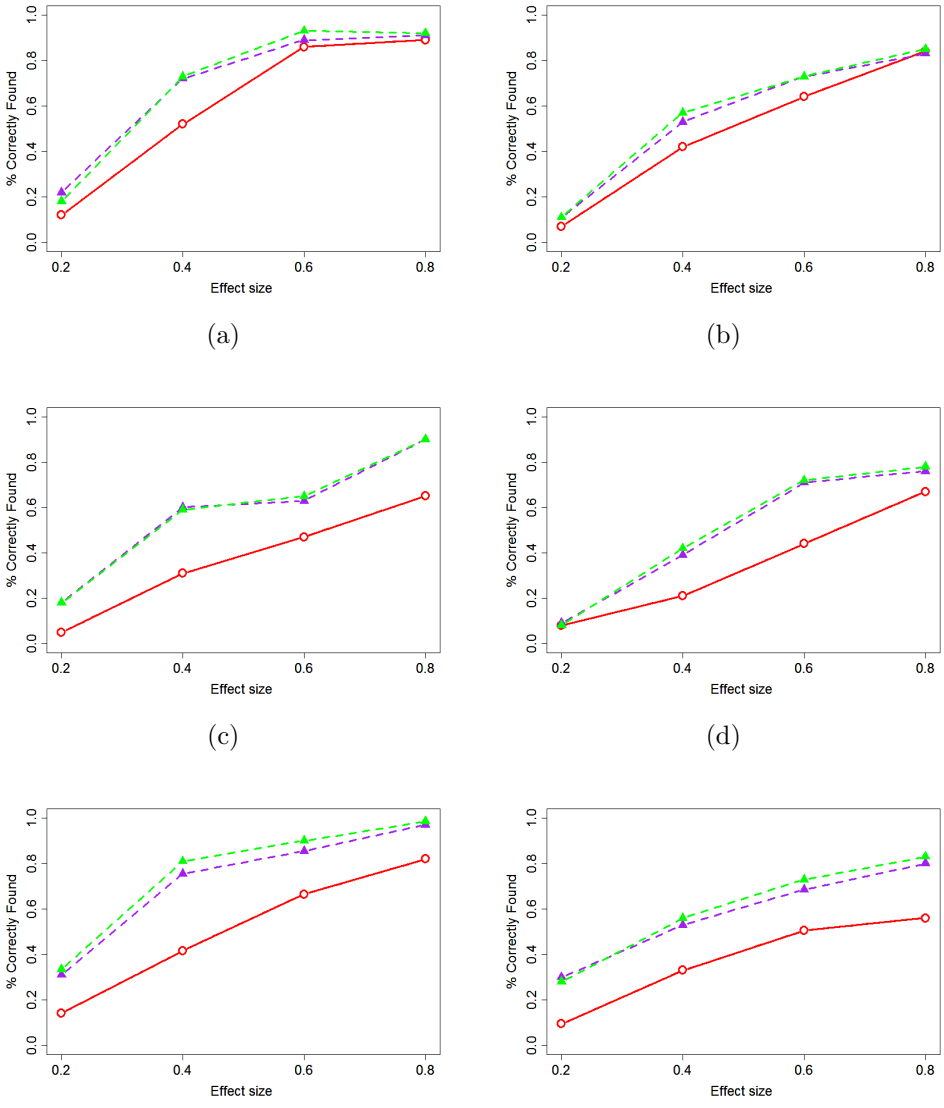
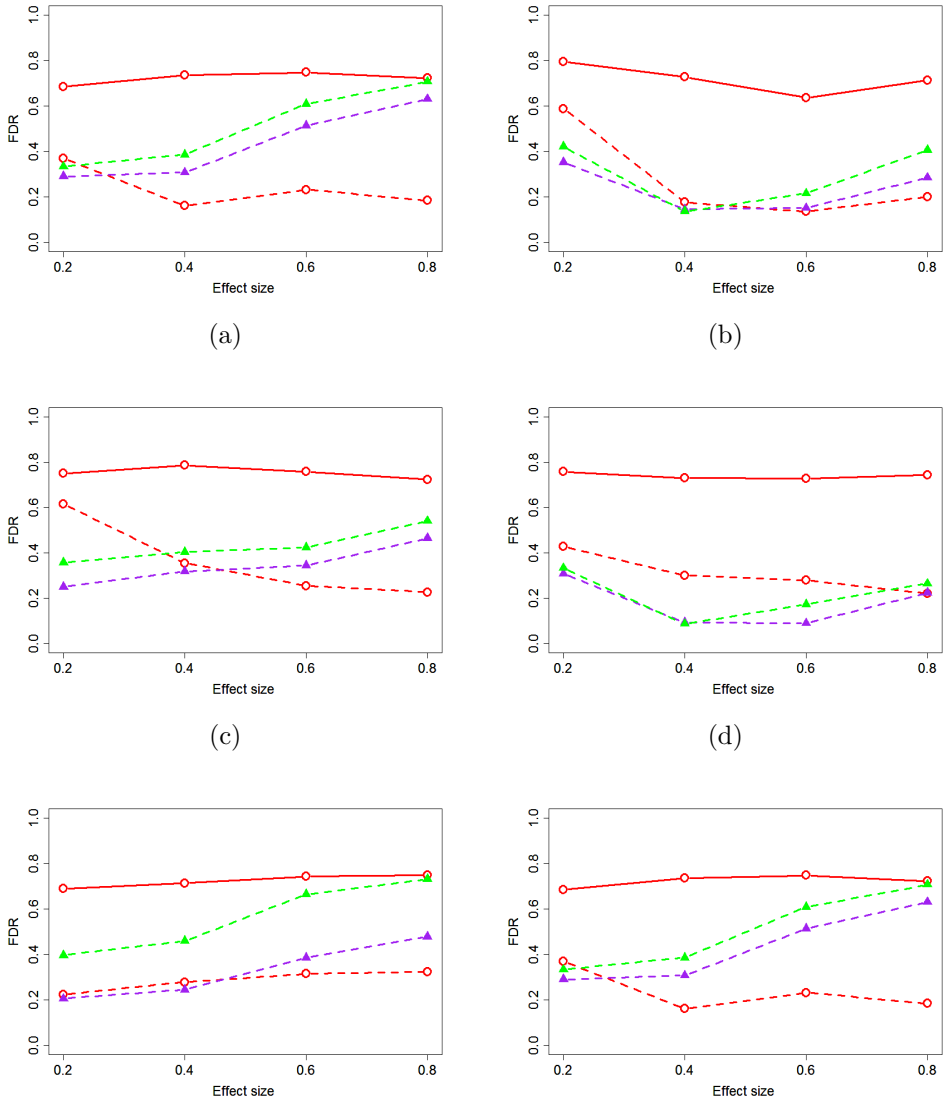


Figure 7.12: Proportion of correctly found SNPs across the six causation models, for Treescan Parsimony method (solid red line with circles), Bonferroni haplotype (dashed green line with triangles) and 'single' haplotype methods (dashed purple line with triangles).



(a) (b)

(c) (d)

(e) (f)

Figure 7.13: False discovery rates for found SNPs across the six causation models, for Treescan Parsimony method (solid red line with circles), Bonferroni haplotype (dashed green line with triangles) and 'single' haplotype methods (dashed purple line with triangles)

As expected the FDR of the branch correction method is substantially lower than that of the single SNP approach, due to multiple mutations occurring on branches of the reconstructed haplotype trees. There is, however, the added consideration that, for the branch correction method, there is no information relating to which of the SNPs that occur on a branch is indeed causative.

7.2.2 Treescan and Haplotype Comparisons

The Treescan and haplotype-based analysis can be compared in the frequentist setting, and the results are displayed in figure 7.12. The comparison is to a parsimony-reconstructed tree (red circles), as the previous section illustrated that the difference in phylogenetic approaches is small. The initial impression from the plot is that the haplotype-based approaches are more powerful at finding the true causative mutations for each of the simulation scenarios that have been considered. This result is surprising in that it would be expected that the extra information supplied by the reconstructed tree would provide more power for the detection of causative mutations. It would however appear to be the case that the tree-construction methods are reasonably poor at reconstructing the tree, as a result of the recombination and homoplasy events that occur in the simulated data. The False Discovery Rates attributable to the haplotype-based methods are however generally higher than the comparable branch corrected Treescan approach, apart from possibly for figures 7.13(b) and 7.13(d) representing the two dominant models of inheritance, when it may be easier to detect differences using the haplotype approaches. The FDR of the non-branch-corrected SNP assessment from Treescan is unreasonably high for all the simulations, as a result of the trees that have been reconstructed containing multiple SNPs on each branch making it difficult to distinguish between potential causative and non-causative mutations.

Similar comparisons between the different approaches can also be obtained in the Bayesian-based association methods (results not shown). The interpretation of the results are similar to that of the frequentist approaches, with comparable differences in both success and FDRs between the methods.

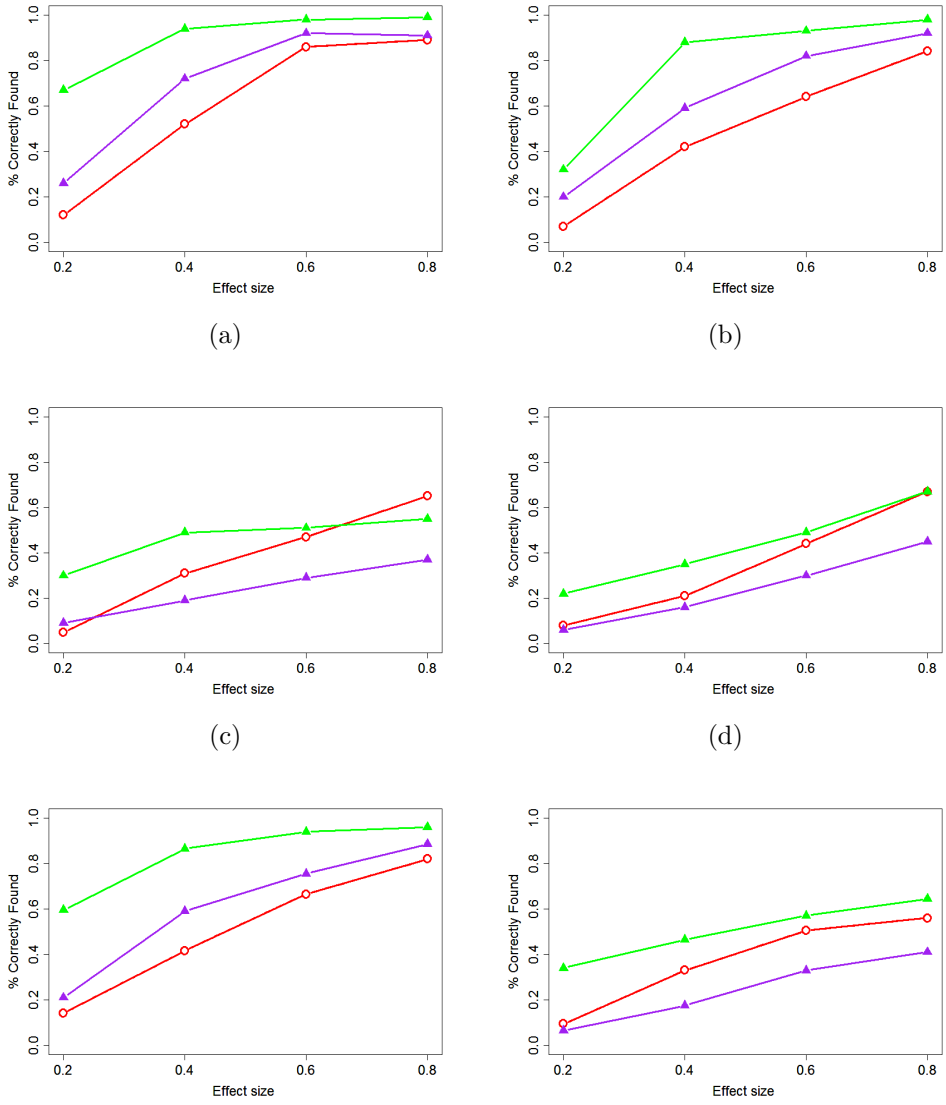
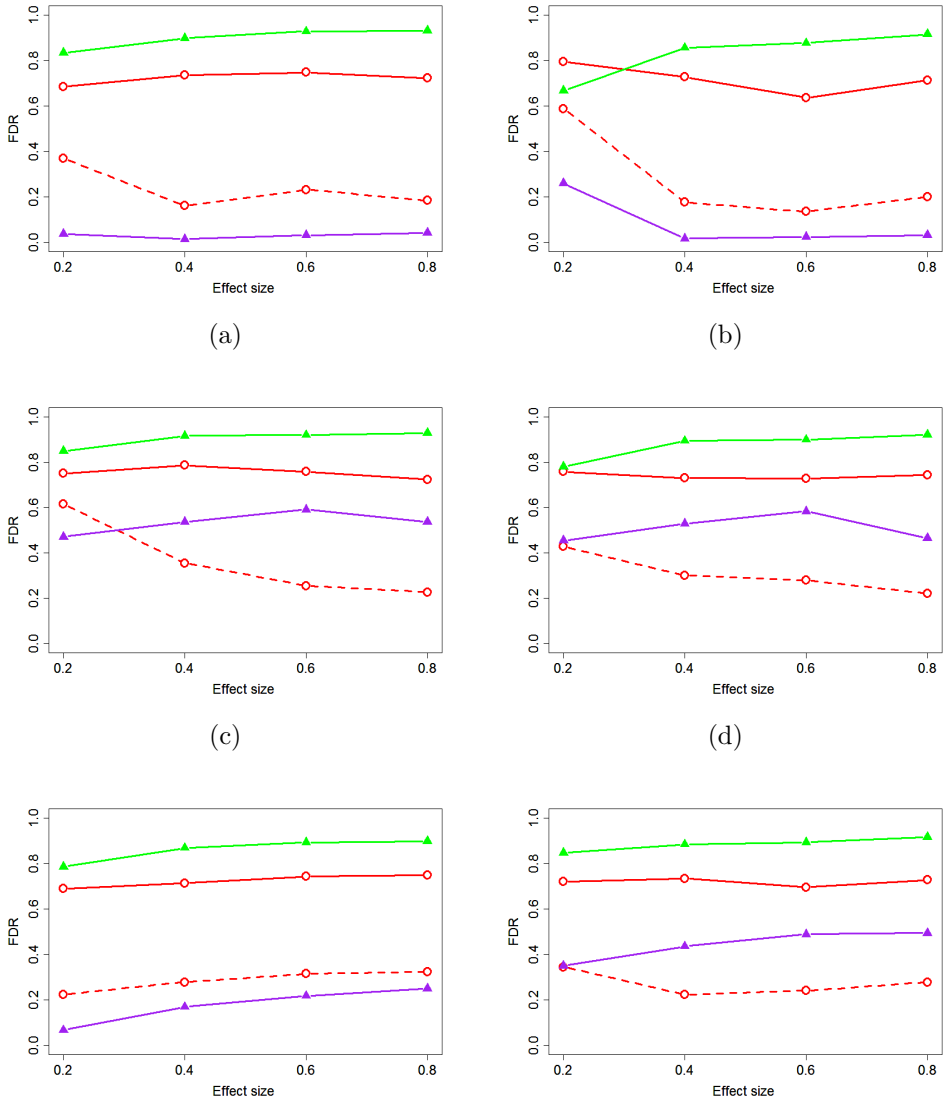


Figure 7.14: Proportion of correctly found SNPs across the six causation models, for parsimony based Treescan (solid red line with circles), Bonferroni SNP (solid green line with triangles) and single SNP methods (solid purple line with triangles).



(a) (b)

(c) (d)

(e) (f)

Figure 7.15: False discovery rates for found SNPs across the six causation models, for parsimony based Treescan (solid red line with circles), Bonferroni SNP (solid green line with triangles) and single SNP methods (solid purple line with triangles).

7.2.3 Treescan and Single SNP Comparisons

Figure 7.14 illustrates comparisons that can be made between Treescan and the single SNP correction methods in the frequentist setting. The initial impression from the plot is that the Bonferroni method is superior to all the other approaches. However, figure 7.15 illustrates that there are exceptionally high False Discovery Rates associated with the Bonferroni approach. Although the Bonferroni approach is generally considered to be over-conservative in its correction levels when applied to GWAS, in the fine-scale context the reverse is apparently true. This is due to the linkage that exists between SNPs resulting in the 'spill over' effects of closely-linked SNPs also being strongly associated with the phenotype of interest. However, unlike the other approaches being considered there is no natural extension of the Bonferroni correction that could allow for the consecutive assessment of both SNPs individually and in combination. The Bonferroni method therefore correctly finds many causative SNPs, but cannot successfully separate the effects of a causative mutation from other closely linked non-causative SNPs.

A second feature of the illustrated results is the differences between figures 7.14(a and b) and 7.14(c and d). The differences are largely due to the choice of whether a mutation is causative in all locations that it occurs in the ARG, or whether only one lineage is affected by the occurrence of a mutation resulting in the causative form of a SNP. This approach is in some sense intended to reflect the penetrance of a mutation, in that not all carriers of a causative form of a mutation will necessarily have an increased phenotypic measurement. As would be expected, the performances of each of the methods tends to be worse in the situation of lower penetrance of the not-all-causative models. There are though further comparisons to be made regarding the relative effectiveness of each approach under the two simulated criteria.

In the situations of figures 7.14(a) and 7.14(b), where a mutation is causative at all locations in an ARG, it can be seen that the single SNP-based methods perform better than the Treescan approach in terms of finding the true causative SNPs. It can also be seen from figures 7.15(a) and 7.15(b) that the lowest False Discovery Rates are also obtained when using the single SNP-based approach.

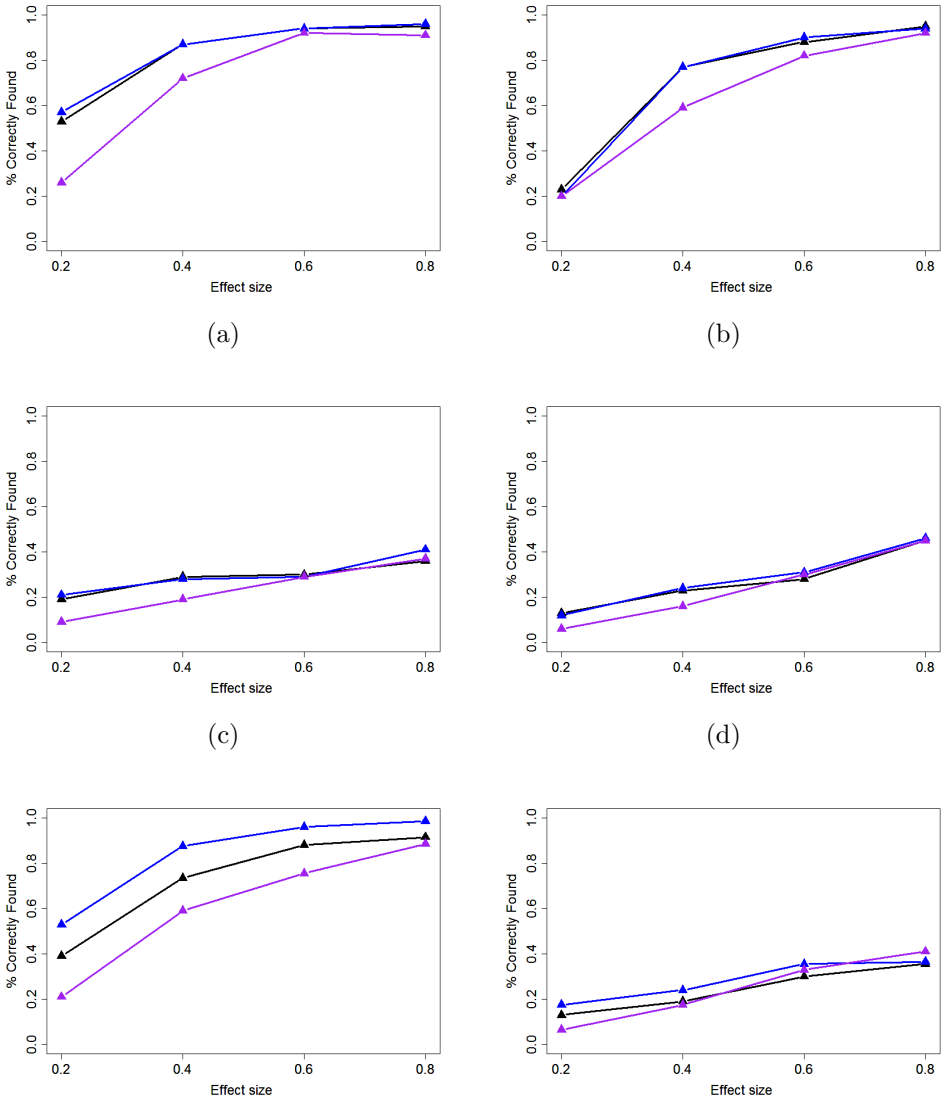


Figure 7.16: Proportion of correctly found SNPs across the six causation models, for the single SNP (solid purple line with triangles), Bayes factor SNP (solid black line with triangles) and BimBam SNP (solid blue line with triangles) methods.

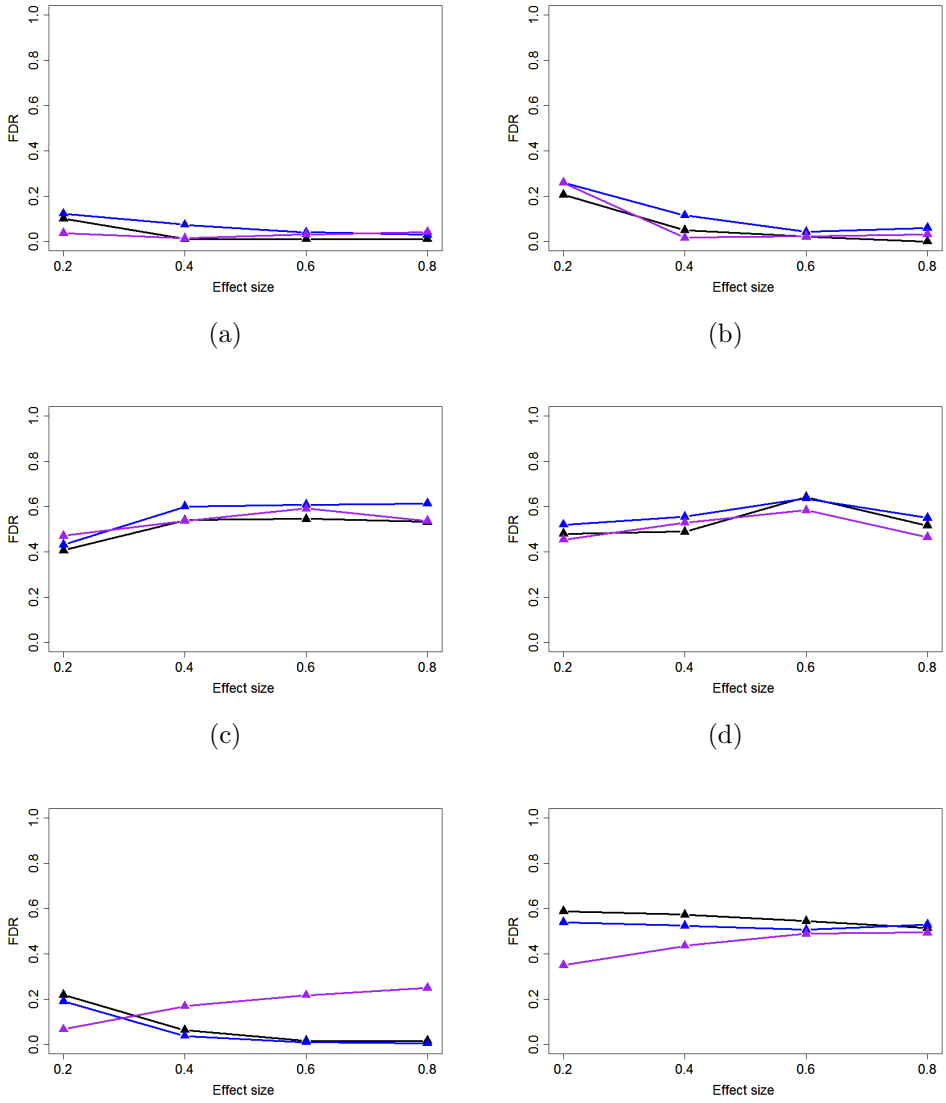


Figure 7.17: False discovery rates for found SNPs across the six causation models, for the single SNP (solid purple line with triangles), Bayes factor SNP (solid black line with triangles) and BimBam SNP (solid blue line with triangles) methods.

In the alternative situation where a mutation is only causative at one lineage in the ARG, it can be seen that the single SNP method performs worse than the Treescan-based approaches. This would be expected, as the single SNP method would be collating together individuals with causative and non-causative variants of the same mutation. Treescan would however construct a haplotype tree that may be able to separate the same mutant allele into multiple mutations on different branches. Differences in the groupings of phenotypes may therefore be possible to detect, and thus the Treescan approach could be more powerful at finding causative mutations in this situation.

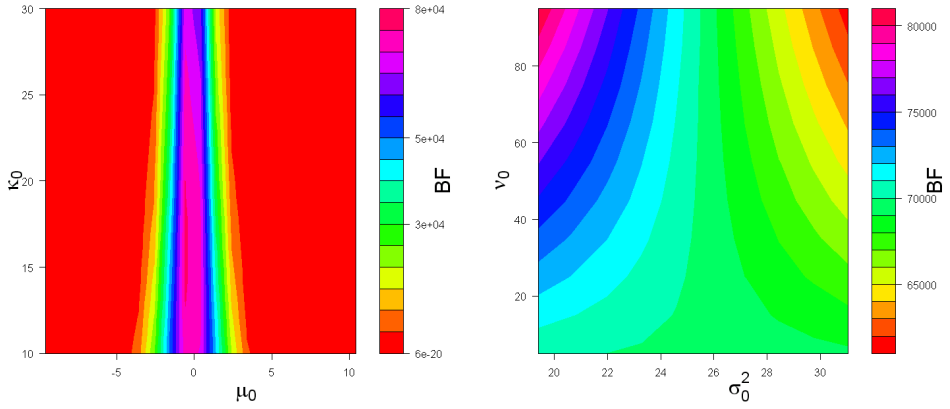
The differences observed between the frequentist approaches are also observed in the Bayesian setting, albeit with two slight differences. The first of these is that of marginally higher success rate for the Bayes factor Treescan approach as opposed to the frequentist method, however this is offset by there also being slightly higher False Discovery Rates. This is likely in part due to the different decision procedures used in the two approaches (as described previously in figures 4.6 and 4.7), as much as it is due to inherent differences in the Frequentist and Bayesian models.

A second difference is that the Bayesian haplotype association method appears to perform considerably worse than the frequentist Haplotype version. This is due to the haplotype analyses resulting in rare groupings to be tested for association, which can result by chance in low p-values. However, these will not be reflected in large Bayes factors since the priors on the Bayes Factors can be used to reduce the impact of such situations.

7.2.4 Frequentist and Bayesian Comparisons

Figures 7.16 and 7.17 show the comparisons between the single SNP-based methods, for the Bayesian and frequentist settings. The results appear to be broadly similar to each other in all of the simulation settings. There is, however, an indication that the Bayes factors methods are slightly better at finding causative mutations; particularly for lower effect sizes.

The Bayes factors of Chapter 3 and those used by BimBam are both similar in their design, although there are some differences in the prior specifications



(a) Sensitivity of Bayes factors to the hyperparameters of the overall mean

(b) Sensitivity of Bayes factors to the hyperparameters of the within-group variance

Figure 7.18: Sensitivity of simulations to prior choices of the Bayes factors.

that have been made in each approach. Despite this, the results obtained are broadly similar for the success rate and for the FDR. There is some indication that BimBam tends to perform better than the Bayes factor approach of PheGe-Sim. However, as illustrated by figures 7.17(a) to 7.17(d) this is balanced by there also being higher rates of False Discovery for the BimBam approach. In figures 7.17(e) and 7.17(f) there is a stronger suggestion that BimBam is preferable to the Bayes factors of PheGe-Sim, which is as a result of the BimBam approach being more willing to accept a model with two causative SNPs.

7.2.5 Sensitivity Analysis

Although in the PheGe-Sim program the results of the Bayes factors sensitivity to hyperprior choices is not automatically explored, it is useful to produce the sensitivity plots for some of the simulations where the true underlying effects are known.

Figure 7.18(a) demonstrates the sensitivity of the Bayes factors in the simulated data to the hyperparameters, for the prior of the mean for a SNP known to be causative for the phenotype. It is apparent that the choice of μ_0 has a substantial effect on the resultant Bayes factor, although the Bayes factor remains high

within a reasonable range about the sample mean of the data. In this situation it appears that the prior sample size suggested by κ_0 has relatively little impact on the Bayes factors, as a result of the large sample size of 2000 being for the simulations.

For the same true causative SNP, the effect of the ν_0 and σ_0^2 hyperparameters are shown in figure 7.18(b). Changing the σ_0^2 value corresponds to a reasonable change in the Bayes factor, although the Bayes factor remains high throughout. In a similar manner to the value of κ_0, the effect of ν_0 is relatively small due to the large sample size that has been used.

The illustration of the sensitivity of the Bayes factors to the prior values for the simulated data set is similar to that observed in the real data set of Chapter 6. As would be expected, a reduction of the sample size will result in there being more effect of the priors on the Bayes factors, and this would be apparent if chosen for the simulations, and is also shown in the small data set of chapter 5.

7.3 Conclusions

It has been shown that the simulation of data sets using PheGe-Sim can reasonably represent the real data set of Chapter 6. Fine adjustments could be made to make the linkage plots of the simulated data more closely resemble that of the ADRA1A data set. However, given that there will be fluctuations between real data sets, the simulated data would appear to be within reasonable range of the real data. If further data sets were to become available, reasonable ranges of the parameter values (as opposed to just point estimates) could potentially be ascertained, and the accuracy of the method of simulation could be more accurately verified.

The results of the association methods in detecting causative mutations illustrates that there is no method that is most appropriate for all of the six choices of scenarios. Each method considered has advantages and disadvantages, and the eventual choice of which could be most useful would be dependent on the prior beliefs in the various scenarios.

Methods that rely on the construction of haplotypes are accompanied by the inherent problem that it is likely that it is SNPs as opposed to haplotypes that

cause changes in phenotype measurement. Identification of haplotypes containing a causative form of a mutation can be useful, however, if the causative SNP is not identified then it would be more difficult to explain the underlying basis of genetic conditions. However, imputation of SNPs that are not genotyped Servin and Stephens (2007) based upon linkage with other SNPs, is a possible approach that can be useful in identifying non-genotyped causal variants. In situations of many low frequency haplotypes there will also be low power to detect any potential genetic effects. Treescan analysis can allow for the identification of the SNP upon a haplotype that causes a change in phenotype. However, in the simulations performed it has been shown through the high FDR that this approach can struggle to differentiate causative and non-causative mutations that could be predicted to occur on the same branch of a haplotype tree.

Single SNP methods may have the advantage of identifying specific causative mutations, however there are also potential problems with their use. The method of applying a Bonferroni correction has been shown to be flawed for tightly-linked SNPs, as it cannot differentiate between causative SNPs and those in strong association. It could be argued that the causative SNP will be the one with the highest strength of association as determined by the p-value. However, there is no appropriate method of determining if more than one SNP results in a change of phenotype. A single SNP permutation approach can be useful in determining causative associations, however, the use of p-values limits the flexibility in testing various models of association that can have an impact particularly at small effect sizes. This is a real issue since most SNPs so far detected to be involved in complex diseases appear to have relatively small effects.

The use of single SNP Bayes factors tests from Chapter 3, and also those of BimBam, appear to be relatively successful at identifying true causative mutations without also identifying large numbers of false positives. This is in part due to two or more SNPs being tested together, and also due to it being possible to test for multiple models of inheritance. However, for complex diseases it is plausible that there is incomplete penetrance of causative mutations, as has been simulated using the 'Not All Causative' setting. In this situation, a Treescan-based approach can be useful at separating the causative and non-causative forms of mutations. It seems, based upon the simulations, that single SNP and Bayesian

based methods may be slightly preferable compared to the other approaches. However, any differences in approaches should also be assessed using real data sets, to assess any differences that are not apparent in the simulated data.

7.4 Conclusions

The primary conclusion that can be made regarding the data available on the ADRA1A gene is that there are no strong associations present in the gene with any of the phenotype measurements of heart rate or blood pressure. This conclusion is similar to that made by the WTCCC's (2007) GWAS, where the association had been tested with the categorical outcome of hypertension. Newton-Cheh et al. (2009) also found no significant effects with hypertension or blood pressure readings, commenting on the lack of association at the ADRA1A gene in particular due to the region being known to be targeted by anti-hypertensive drugs. There are numerous reasons why no strong associations have been identified, and the explanation that there are indeed no true associations is entirely plausible. However, it is also known that the phenotypes, of blood pressure in particular, are notoriously difficult to measure accurately. This is as a result of the blood pressure and heart rate of an individual being highly variable between measurements, and even though the study protocol has tried to minimize the effects of this a moderately sized genetic effect can be masked by this variability. Further to this, there are known environmental effects on hypertension and blood pressure, such as stress, diet and exercise (Korner, 2007), such that real but small genetic effects can effectively be lost within a sea of noise. It is also feasible that the ADRA1A gene interacts with other genes in some way, and assessment of an effect in each gene individually does not yield sufficient information for positive associations to be discovered.

Although no strong associations have been identified, the analysis in this chapter can provide an insight into some of the various strengths and weaknesses of the association methods under consideration that would not be possible using only simulated data. Arguably the most relevant of these is the extra effort required to construct a haplotype tree for the Treescan-based methods, with the first related issue being that regions of low recombination must be identified

prior to phasing of the haplotypes. If reasonable haplotype blocks have been identified, then there is a further complication in the accuracy of the method used in the construction of the haplotype tree. In this data set, irrespective of the tree method used, there existed a large number of repeat mutations that resulted in small and potentially unsuitable haplotype groupings. On balance, for the PAMELA data, the potential benefits of a Treescan-based approach are outweighed by the extra effort and difficulty required in running such an analysis.

In addition to the preference of using single SNP-based methods as opposed to Treescan-based approaches, there is also some indication that the Bayes factor approach may be more useful compared to the standard use of p-values. This is particularly evident in the SNPs that do show some degree of association with any of the phenotypes, in that the strength of a reported association can be sensitive to the choice of the mutation model that has been used. A Bayes factor approach can allow for selection over competing models, possibly by averaging of the Bayes factors for each model, whereas in the frequentist approach this is more difficult due to issues in dealing with adjustment of p-values as a result of multiple testing considerations. The Bayes factors can also be useful in their assessment of groups with small numbers of individuals, in that such associations are unlikely to be over-interpreted, as the low power to detect associations will be properly reflected in the size of the Bayes factor. Although the use of Bayes factors requires the explicit specification of prior information, for this data set these can be sensibly chosen, and when this is the case, they can have a relatively minor impact on the results.

Chapter 8

Conclusions & Future Research

The objective of this thesis was to explore methods that can be used for fine-scale phenotype-genotype association studies. This chapter will provide a summary of the preceding chapters and the results that have been obtained, before describing features of the thesis that could potentially be extended or improved upon.

Chapter 1 illustrates the historical context of genetic association studies, and presents some of the specific challenges that are faced in the context of fine-scale studies. The coalescent model is introduced as a tool that can be used for developing methods that are appropriate for both the simulation and analysis of fine-scale genetic data sets. Methods that have been suggested for the specific use in fine-scale studies were then introduced. In particular, haplotype-based approaches and the Treescan method are discussed. The potential to implement a Bayesian approach as an alternative to the commonly used frequentist methodology is then suggested, and possible advantages and disadvantages of such an approach in the fine-scale genetic association context are discussed.

Chapter 2 then introduced the PheGe-Sim (Phenotype Genotype Simulation) program that has been developed in the R language, to simulate data with some of the specific features of fine-scale genotype-phenotype data sets. This program extends the basic coalescent process, by allowing the possibility of recombination, population expansion, and a finite-sites model of mutation to be modelled. It is hoped that inclusion of such features will improve the accuracy of the simulated data sets, when compared to the real data of Chapters 5 and 6 and fine-scale

association study data sets in general. The procedures involved in allocating mutations on to the Ancestral Recombination Graph are subsequently discussed, in addition to the methods used for creating the haplotypes generated as a result of the mutations.

The novel formulation of the Bayes factors used for PheGe-Sim and PheGe Find is presented in Chapter 3. All the methods presented use conjugate prior distributions to obtain exactly computed Bayes factors, which allows for fast calculation of marginal likelihoods compared to some other possible approaches. Methods of considering SNPs in combination with each other are also presented, that retain information relating to the causation model involved.

The PheGe-Find (Phenotype Genotype - Find) application is introduced in Chapter 4, in which different approaches that can be used for the association studies are discussed. The methods used for association are also implemented in the PheGe-Sim application, however, PheGe-Find is capable of reading in, checking for inconsistencies, and analyzing real data sets from a variety of input file formats. Methods of constructing the haplotype trees required by Treescan are briefly explored, as well as the methods of determining if causative mutations have been correctly found when using simulated data sets.

The association methods that are under consideration for the simulated data are also used upon the real data sets of Chapters 5 and 6. The data of Chapter 5 represents data from the ADH gene of the *Drosophila melanogaster* fruit fly that was initially analyzed by the Nested Clade Analysis (NCA) method, a precursor to Treescan. As such, this data set can be used to compare the association approaches upon data with previously detected genotype-phenotype associations. The results from analysis of this data set are consistent with the original findings, with strong associations being found at three of the sites being considered. Only one of the SNPs is considered found after corrections for correlations between the SNPs and for multiple testing. The similarity between the results confirms that the Bayes factors approaches are at least broadly in agreement with the standard frequentist approaches when considering data with strong phenotype-genotype associations.

Chapter 6 used the association methods upon a human data set concerning the ADRA1A gene and three separate phenotype measurements, and is a data set

that has not previously been analyzed using either the Treescan or Bayesian-based approaches. Some significant associations have been found using the frequentist approach, however, these do not pass any of the correction criteria used to address multiple testing. The related Bayes factors also suggest that there are no overly strong associations between the SNPs and any of the recorded phenotype measurements. This data set does though illustrate some of the potential drawbacks of the haplotype and Treescan approaches, that are not apparent when using simulated data. In particular, the additional difficulties in phasing and constructing haplotype trees were highlighted, issues that negate any benefit that Treescan may have had when using the simulated data.

Chapter 7 presents the results of a simulation study, assessing both the accuracy of the simulated coalescent data and the relative strengths of the association methods that have been considered. As would be expected, the conclusions that can be made from the simulations depends heavily on the choice of parameters that are used. However, some general conclusions can be obtained. It is shown that the parameters used to obtain similar patterns of linkage to the real data for the simulations are reasonably similar to those estimated from the real data sets. It is also shown that without having the possibility of a finite-sites model and recombination that is contained within PheGe-Sim, it is not possible to generate suitably realistic data sets.

The first conclusion that can be made regarding the association approaches on the simulated data, is that the method of using the Bonferroni-correction to assess for significance is unsuitable in fine-scale studies. Although the Bonferroni approach can detect many causative SNPs, there is no suitable procedure in differentiating causative SNPs from those that are in strong linkage, and similarly there is also no appropriate procedure for detection of multiple SNPs in combination.

Haplotype and Treescan-based approaches are shown to be useful for detecting SNPs that are not causative in all locations in which they occur. However, there can also be high false discovery rates associated with these approaches under certain scenarios. Single SNP methods seem to have a slight advantage over the approaches using haplotypes, and the difference is particularly clear for SNPs that are causative in all locations, where the single-SNP based methods are powerful

at finding causative mutations while also being associated with low false discovery rates. Imputation of un-genotyped SNPs can also potentially be useful to further identify possible causative variants if they are in reasonable linkage with SNPs that have been directly genotyped. The simulations also indicate that there may be a slight advantage of using Bayes factors as opposed to the standard use of p-values, however, there does not appear to be a large degree of difference in the ability of the single SNP permutation method and the two formulations of Bayes factors.

There are a variety of ways in which the research that has been presented in this thesis can be extended, and to some extents improved. The first such consideration is that all the models that have been presented, both Bayesian and frequentist, have used normally distributed data. Although normality is usually reasonable for natural data (or transformations of the data), the potential of non-normal data sets could be further explored. As noted in section 3.9, there is potential to include covariates into the Bayes factor procedures, and therefore subsequently into the PheGe-Find program. This would be useful for considering genetic and environment considerations together, however, would introduce issues of model selection that would have to be addressed. The stepwise approach of adding one SNP at a time has been shown to be useful for taking into account the linkage between multiple SNPs. However, this procedure may not be as appropriate for when there could be multiple potential covariates, and methods such as the Lasso approach may be more suitable in such situations.

There is flexibility in the construction of PheGe-Sim such that many more models of associations could be tested, with multiple effects and interactions. Assessment of such models may provide further insight into the potential benefits and drawbacks of the methods used, however, it is not known as to how realistic the simulated causation models would be. If more data sets similar in size to the data of Chapter 6 were to become available, the parameters of the simulations could be adjusted to take account of the additional information.

A further improvement that could be made to the simulation of the coalescent is to simulate the haplotype block structure, with regions of low recombination separated by heavily recombining regions. In order to achieve this, the methods

currently in use would have to be adapted to be more efficient in dealing with recombination events. However, as noted in section C.3, the rate of recombination would have to be so high as to make the haplotype blocks close to being independent. Improvements would also subsequently have to be made to the detection of causative SNPs if a haplotype block structure was implemented, as the detection of hotspots would also have to be automated and integrated into PheGe-Find.

Although there is some scope for improvement in aspects of the analysis that has been presented, the methods of the PheGe-Sim and PheGe-Find applications have been demonstrated to be useful in the context of fine-scale genetic association studies; an area which will become ever more relevant as GWAS identify further areas of interest, and as the SNP map of the human genome becomes ever more detailed.

Appendix A

Supplementary Figures and File Formats

A.1 Example Fasta Output

Example of output Fasta File format:

>Sequence1
AAAGGGCGGACCTATCTTGT
TGTTCTATCCAGGCGGGAAA
>Sequence2
AAAGGGCTGACCTATCTTGT
TGTTCTATCCAGTCGGGAAA
>Sequence3
GGGAAACTGCAGTAATTTTC
CTTTTAATGACGTCAAAGGG

Sequence1: Label of sequence, preceded by > character
Sequence itself can continue over as many lines as required, until a new sequence label is introduced.

A.2 Example PHYLIP output

Example of output Phylip file format:

```
4 40
Sequence1 AAAGGGCGGACCTATCTTGT
Sequence2 AAAGGGCTGACCTATCTTGT
Sequence3 GGGAAACTGCAGTAATTTTC

TGTTCTATCCAGGCGGGAAA
TGTTCTATCCAGTCGGGAAA
CTTTTAATGACGTCAAAGGG
```

4: Number of Sequences
40: Number of SNPs
Sequence1: Label of sequence. Will take the first ten characters (including spaces) as the title of the sequence.

A.3 Example PED file format

Example of .Ped file format:

```
000001 000001 0 0 1 0 T A A A A A G G G A A G G T C C T C T G
000002 000002 0 0 1 0 A A A T A C G G G A G G T G C T T T T G
000003 000003 0 0 1 0 T A A A A C C G G G A G G G C T T C T G
```

Of the format:
000001 000001: Individual Name and Family Name (not relevant in this thesis)
0 0 1 0: Father ID, Mother ID, Sex, Affected Status
Base at SNP 1 on first haplotype,
Base at SNP 1 on second haplotype,
Base at SNP 2 on first haplotype,

....,
(Note can be phased or unphased haplotypes)

A.4 Example Sim Results File

```
Sun Apr 18 16:09:37 2010 Order: --TSC.PARS-- --BF.PARS-- --Bay.SNP--
Sim_1 True:       121               |121               |121|
      Found:      55,121,148,156,175|55,121,148,156,175|121|
      Corr.Find:  121               |121               |121|
      False +:    55,148,156,175    |55,148,156,175    |   |
      False -:                      |                  |   |
      Linked +:                     |                  |   |

Sim_2 True:       397|397|397|
      Found:      397|   |397|
      Corr.Find:  397|   |397|
      False +:       |   |   |
      False -:       |397|   |
      Linked +:      |   |   |


Number of correctly found mutations [Tsc Pars] 2/2 (100%)
Equivalent to 1 correctly found mutations per simulation (1 per sim
true number) False positive rate of 2 (4 Total), per simulation
False Discovery Rate 0.6666667

Number of correctly found mutations [Tsc Pars BRANCH] 2/2 (100%)
Equivalent to 1 correctly found mutations per simulation (1 per sim
true number) False positive rate of 0 (0 Total), per simulation
False Discovery Rate 0

Number of correctly found mutations [Bayes Factor Pars] 1/2 (50%)
Equivalent to 0.5 correctly found mutations per simulation (1 per
```

```
sim true number) False positive rate of 2 (4 Total), per simulation
False Discovery Rate 0.8

Number of correctly found mutations [Bay Pars BRANCH] 1/2 (50%)
Equivalent to 0.5 correctly found mutations per simulation (1 per
sim true number) False positive rate of 0 (0 Total), per simulation
False Discovery Rate 0

Number of correctly found mutations [Bayes Factor SNP] 2/2 (100%)
Equivalent to 1 correctly found mutations per simulation (1 per sim
true number) False positive rate of 0 (0 Total), per simulation
False Discovery Rate 0

Sun Apr 18 16:11:40 2010
```

Of the form:
Summary of the chosen methods of association, and the order in which they appear in the tables.
For each simulation and each causative method used, summaries are given as to whether the simulated causative mutations have been found (see table 2.4).
For each method of association used, summaries are given of the found and false discovery rates across all of the simulated data sets.

A.5 Example Details File

```
Sun Apr 18 16:06:50 2010
Sim_1

Selected causative Branches
431

Individuals at risk and defective mutations
```

```
73 1



Observable Haplotypes
1 2 3 4 5 6 7 8 9 12 14 15 17 19 20 23 24 26 28 30 38 39 40 41
43 45 47 54 55 56 64 72 73 74 75 80 83 85 91 92 93 101 105 111
112 114 115 116 118 125 134 148 150 153 158 162 164 168 175 195
198 204 207 216 229

Caus.sites   121
-------------------------------------------------------------------------
Sun Apr 18 16:09:38 2010
Sim_2

Selected causative Branches
197 282

Individuals at risk and defective mutations
197 1
229 1



Observable Haplotypes
1 2 3 4 5 6 7 9 10 11 12 15 19 20 22 24 27 29 31 33 38 43 48 49
50 52 53 60 61 63 100 102 103 138 140 154 157 159 171 178 182
197 200 210 211 213 222 228 229 233


Caus.sites   397
-------------------------------------------------------------------------

Beta paramater = 10
```

```
Number of defective loci = 1
Interaction effects = 0
Mutation Effects = 3
Number of non distinct haplotypes = 250
Number of individuals in population = 2000
Number of Simulations executed = 2
Standard deviations of genotypes = 5
Theta 25
Recombination Rate = 0.5
Mutation Types = A
Finite Sites
AC
Gamma Shape = 5.5
Gamma Scale = 0.5
Average Number of SNPS = 33

Sun Apr 18 16:11:40 2010
```

Selected causative branches: The branches of the ARG upon which causative mutations occur.
Individuals at risk and defective mutations: Details of the haplotypes affected by each of the simulated causative mutations.
Observable Haplotypes: The unique haplotypes occurring as a result of the simulated ARG.
Caus.sites: The site(s) that have been simulated to result in a change in phenotype measurements
A summary is then given recapping the parameters that have been chosen for simulating the data.

A.6 Example Single SNP output File

```
Tue Jun 01 15:49:43 2010
SNP#    F            B-W          P(F)      PCorr     PMon
```

40	10.40165	0.22373	3e-05	0	0
271	4.25547	0.07794	0.01432	0.118	0.118
97	4.81562	0.04574	0.02832	0.35	0.35
197	2.64955	0.03956	0.07093	0.398	0.398
6	1.70707	0.01697	0.18166	0.698	0.698
398	0.6702	-0.00792	0.51172	0.976	0.976
136	0.34279	-0.0158	0.70983	0.992	0.992
273	0.16387	-0.0201	0.84886	0.994	0.994
337	0.15939	-0.02021	0.85267	0.98	0.994
342	0.051	-0.02282	0.95028	0.95	0.994

SNP#	F	B-W	P(F)	PCorr	PMon
271	6.21732	0.30927	1e-05	0.154	0.154
197	3.95722	0.24607	0.00027	0.806	0.806
97	6.0167	0.23861	8e-05	0.842	0.842
6	3.72457	0.22689	0.00052	0.936	0.936
273	4.33676	0.19872	0.00063	0.998	0.998
136	4.29148	0.19604	0.00069	0.994	0.998
398	3.21859	0.18508	0.00214	0.996	0.998
337	3.1661	0.18074	0.00247	0.988	0.998
342	3.11869	0.17681	0.00282	0.94	0.998

Tue Jun 01 15:50:10 2010

SNP#: SNP site position label
F: F statistic as calculated from an ANOVA
B-W: The Boerwinkle-Singh Corrected Estimator
P(F): The p-value calculated from the standard ANOVA
PCorr: The p-value calculated from the permutation correction procedure, using the Boerwinkle-Singh estimates
PMon: The enforced Monotonic p-values from the corrected p-values, ensuring

that a higher B-W value does not result in a lower permutation corrected p-value, as can happen by chance.

A.7 Example Bayes factor File

```
Tue Jun 01 15:50:10 2010
Bayes Factor Results, under the following prior choices
mu_0 =  0  , kappa_0 =  20  , nu_0 =  20  , sigma_sqrd_0 =  20

Split  Null LogL  R_D LogL    Add LogL    Alt LogL    Bay Fac   Group

40   -3209.8     -3203.4     -3205.9     -3203.1      830.49     3
271  -3209.8     -3209.2     -3211.9     -3209.5      1.9027     3
197  -3209.8     -3209.5     -3210.7       -3211      1.3606     3
6    -3209.8     -3210.5     -3211.9     -3211.8     0.48385     3
273  -3209.8     -3210.8     -3211.1     -3211.5     0.35874     3
337  -3209.8       -3211     -3211.5     -3212.2     0.29501     3
136  -3209.8     -3211.1     -3211.7     -3212.5     0.28668     3
97   -3209.8          ..          ..     -3211.2     0.25356     2
342  -3209.8     -3211.2     -3211.7     -3212.6     0.24555     3
398  -3209.8     -3211.8     -3211.5     -3212.8     0.18347     3


Split Level =  2
Split   Null LogL   Mix LogL    Add LogL    Alt LogL    Bay Fac   Groups

40,271  -3203.1     -3206.9          ..     -3202.5      1.7454     6
40,273  -3203.1     -3205.9          ..     -3203.6     0.57764     6
40,136  -3203.1     -3211.9          ..     -3205.4      0.1033     6
40,6    -3203.1          ..          ..     -3208.7    0.003834     8
40,337  -3203.1          ..          ..     -3208.9    0.003102     8
40,197  -3203.1          ..          ..     -3209.3   0.0020395     8
40,342  -3203.1          ..          ..     -3210.5  0.00058009     8
```

```
40,398  -3203.1          ..          ..      -3210.6  0.00053231      8
40,97   -3203.1          ..          ..      -3211.6  0.00020269      5


Best Model found
40

Tue Jun 01 15:50:14 2010
```

Date and time of start of simulation
Hyperparameter choices for the Bayes factors
Split: Branch(es) or SNP(s) under consideration
Null LogL: Null log marginal likelihood of the data set if in the first level of splits, or the highest marginal from the previous round of splits if in the second round of splits or higher.
R_D LogL: The Dominant or Recessive log marginal likelihood (first round of splits)
Mix LogL: The complex mixture marginal log Likelihood (second or higher level of splits)
Add LogL: The additive log marginal likelihood
Alt LogL: The general alternative log marginal likelihood
Bay Fac: The Bayes factor calculated according to the highest possible log marginal likelihood of the possible alternative models, compared to that of the Null
Groups: The number of non-empty groups as a result of the splits
Date and time at the end of the simulation

A.8 Example Phenotype File

```
IID S_CLIN  D_CLIN  H_CLIN
640001  120 88  72
640003  118 68  66
```

```
640004  168 98  72
640005  102 78  66
640007  146 92  68
640009  118 84  60
...     ... ... ...
...     ... ... ...
```

Tab delimited file with column names corresponding to:
IID: Individual Identification Labels. Should correspond to those used in the genotype files.
S_CLIN, D_CLIN, H_Clin: Names for each phenotype measurement. Will be used in labelling the output files and plots of any analysis.

A.9 Example Position File

```
rs4732845   26652190
rs17055869  26653565
rs17055880  26655432
rs1472346   26657311
rs7821479   26657574
rs9314327   26659872
rs4732639   26661027
...         ...
...         ...
```

Tab delimited file, with the site name followed by the base position.

A.10 Gamma Correction

The process aims to match the specified finite sites distribution as closely as possible, without impacting on the number and positions of the mutations that have already occurred as a result of the mutation rate and other parameters. In order to achieve this, the options displayed in the flow chart are taken in an

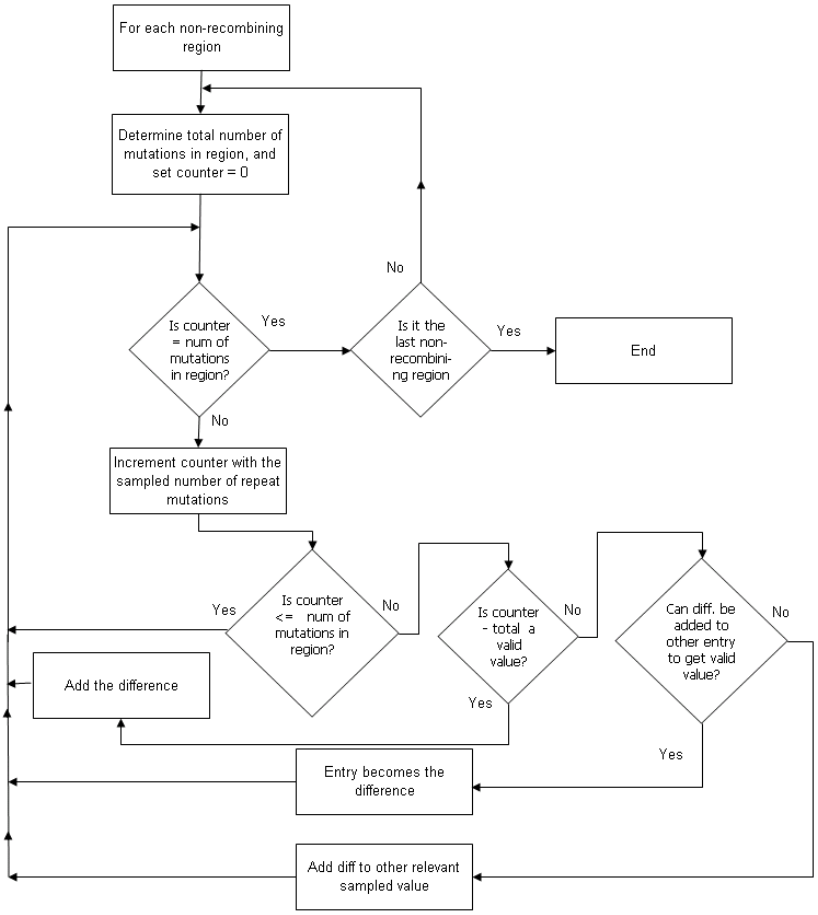
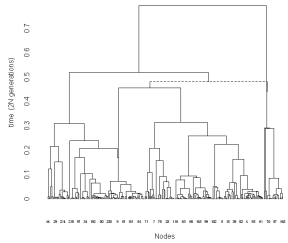
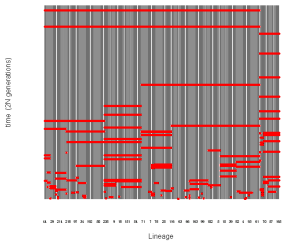


Figure A.1: Flowchart of the decisions involved in allocating finite sites using the Gamma distribution.

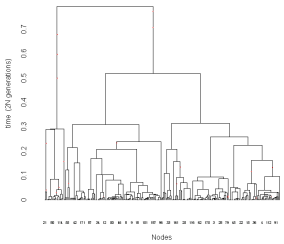
attempt to ensure that sampled values are not simply discarded if they represent a potentially valid level but do not fit into the mutation structure that already exists. In most situations where the initial parameters have been sensibly chosen, the corrections applied will have minimal impact and the finite sites distribution will be closely matched. In figure A.1, a 'valid' value is a number that is within the range of those previously sampled up until that stage in the algorithm



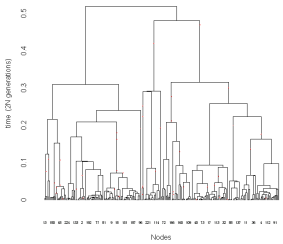
(a) Example ARG



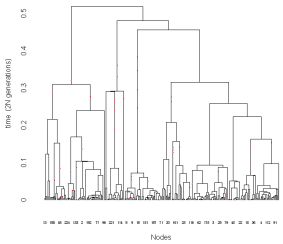
(b) ARG Line Plot



(c) Coalescent Plot A



(d) Coalescent Plot B



(e) Coalescent Plot C

Figure A.2: Example output plots.

A.11 Example Output Plots

Figure A.2(a) illustrates an Ancestral Recombination Graph that has been simulated according to the parameters used for the simulations in chapter 7. In addition to the ARG, the line plot of figure A.2(b) has also been produced, which illustrates the terminal nodes carrying each mutation. In this example, there are two recombination events that occur, resulting in the three separate coalescent regions illustrated in figures A.2(c), A.2(d) and A.2(e). It can be seen in figures A.2(d) and A.2(e) that the heights of the plots are lower than that of the ARG, indicating that the most recent common ancestor for the set of terminal nodes has been reached at an earlier time for this coalescent region. The differences in coalescent structure as a result of the recombination events can also be clearly seen across the three regions.

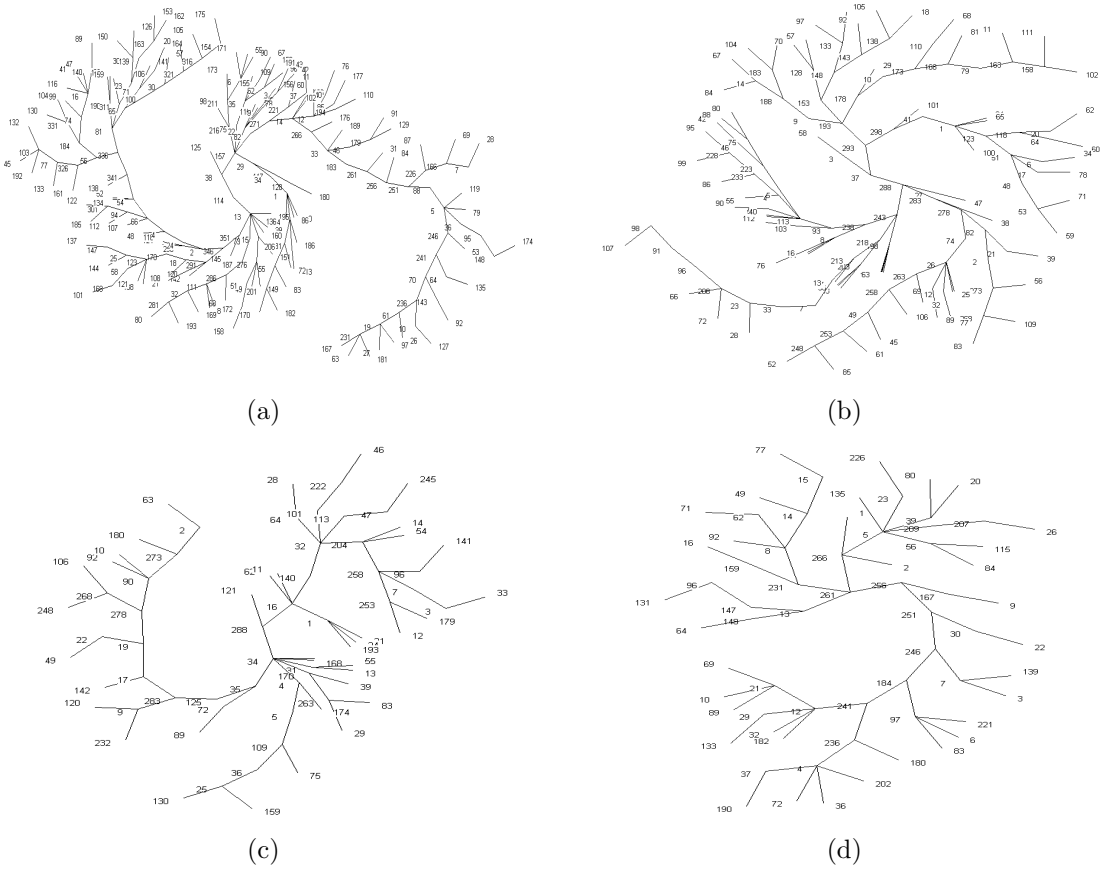


Figure A.3: Reconstructed parsimony haplotype trees for the LHS (a) and RHS (b) of the recombination hotspot of the ADRA1A data set, and two haplotype trees reconstructed from simulated data (c,d)

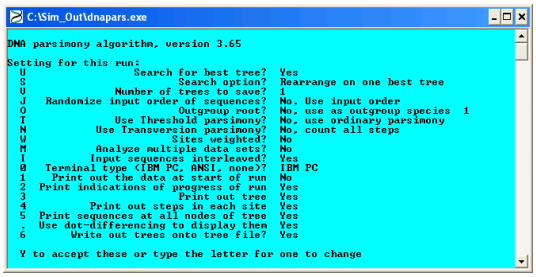
A.12 Reconstructed Haplotype Trees

Figure A.3 illustrates reconstructed haplotype trees for both real and simulated data sets. It can be seen that the trees for the real data sets appear to have more haplotypes than for the simulated data, particularly in figure A.3(a) for the left of the recombination hotspot. The sample size are, however, approximately the same for the real and simulated data, and so the increase in the number of haplotypes implies that these haplotypes exist at lower frequencies on average. This is indeed the case with many of the haplotypes of the real data set being estimated to occur at low frequencies, and so there will be a corresponding low power to detect any associations for the mutations that define these rare haplotypes.

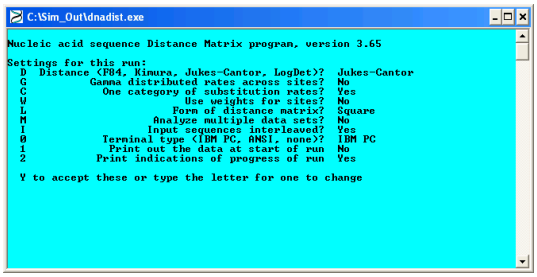
Appendix B

Inputs to External Programs

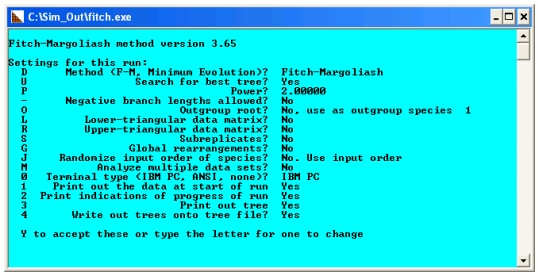
B.1 PHYLIP Input Options



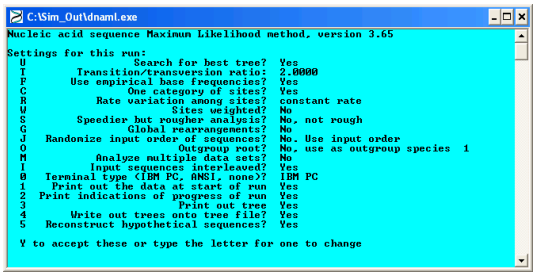
(a) dnapars



(b) dnadist



(c) fitch



(d) dnaml

Figure B.1: Inputs for the PHYLIP program that have been used as defaults for PheGe-Sim and PheGe-Find. Details of the options can be found at: http://evolution.genetics.washington.edu/phylip.html

B.2 Treescan Options

```
system(paste('"C:/Sim_Out/treescan.exe"',
             '"C:/Sim_Out/treescansim_pars"',
              "-b", "-f", "-k", "-p1000"), wait = TRUE, invisible = TRUE)
```

"C:/Sim_Out/treescansim_pars": Input Treescan file (Parsimony, Maximum Likelihood or Fitch)
"C:/Sim_Out/treescan.exe": Path of Treescan application
"-b": Forces the program to print the ANOVA tables to the log file
"-f": Results are sorted by the F statistic
"-k": Uses the Boerwinkle-Singh correction for the tests of association
"-p1000": Specifies 1000 permutations in order to calculate the p-values

B.3 BimBam

```
setwd("C:/Sim_Out")
system(paste("bimbam","-g","bimbam_tsc_geno.txt","-p",
             "bimbam_pheno.PHENO",
             "-pos","bimbam_pos.txt", "-o", "bimtest.out","-l 2","-sort"),
                wait = TRUE, invisible = TRUE)
```

setwd("C:/Sim_Out"): Set path of BimBam application and files.
"-g","bimbam_tsc_geno.txt": Specify BimBam genotype file
"-p","bimbam_pheno.PHENO": Specify BimBam phenotype file
"-pos","bimbam_pos.txt": Specify BimBam site locations file
"-o", "bimtest.out": Specify BimBam output folder
"-l 2": Calculates multi-SNP Bayes factors for all subsets of size 2
"-sort": Sort results according to the highest single SNP Bayes factor

B.4 Haploview

```
setwd("C:/Sim_Out")
Q.PED <- paste("-pedfile Sim_PED_", zorb, ".PED", sep = "")
```

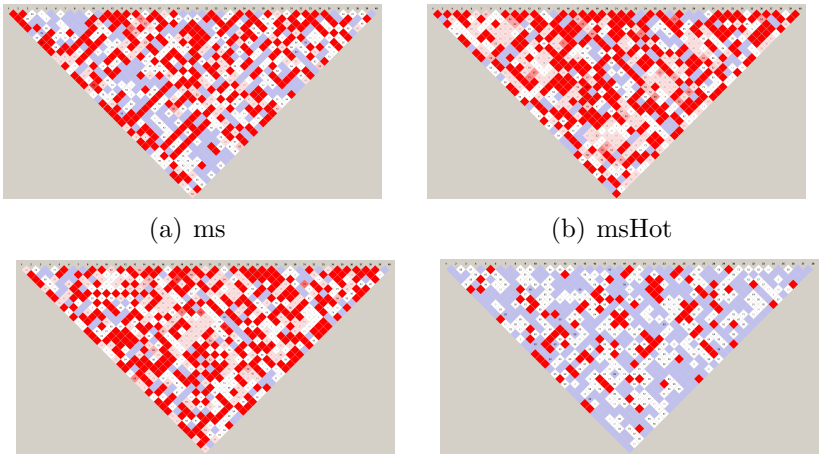
```
system(paste("java","-jar","Haploview.jar"  , Q.PED,
         "-nogui", "-png"), wait = TRUE, invisible = TRUE)
```

`setwd("C:/Sim_Out")`: Set location of Haploview application and files.
`Q.PED`: Location of PED file for current simulation
`"-nogui"`: Instructs Haploview not to open the Graphical User Interface
`"-png"`: Produce the linkage plot in *png* file format

Appendix C

Additional Simulation Results

C.1 Linkage Plots from other Simulators



(a) ms (b) msHot

(c) Genome (d) SimCoal

Figure C.1: Examples of linkage plots from other coalescent simulators. Where possible the parameters have been matched with those used for the parameters for the simulations of Chapter 7.

There are numerous programs that simulate genotype data, and examples of the resultant linkage plots are given in figure C.1. It was not possible to generate data using the CoaSim application, as the program could not be used on a Windows operating system, and there were unidentified problems associated with its use under Linux.

Figure C.1 illustrates that the simulators can generally recreate similar patterns of linkage compared to that of real data. It is evident, however, that without the use of a finite-sites model, there is an overabundance of blue and dark red squares in the linkage plots. It appears to be the case that recombination alone struggles to obtain sites displaying intermediate shades of red, as was displayed in the linkage plots of the ADRA1A data set of Chapter 6 (figure 6.3).

C.2 Illustration of Linkage Plots without Finite Sites

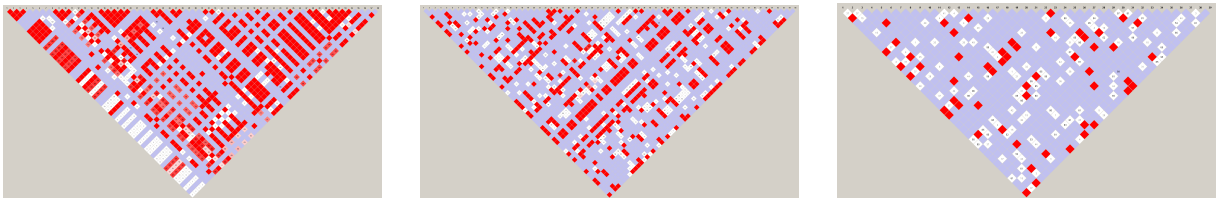
The apparent importance of including finite sites into the PheGe-Sim program can be explored using the linkage plots under various scenarios where the finite sites assumption is not used, and the other parameters are altered.

(a) $\beta = 0$ (b) $\beta = 10$ (c) $\beta = 50$

Figure C.2: Simulated linkage plots for various combinations of parameters (infinite-sites).

Figure C.2 uses an infinite-sites model and no recombination, where each plot corresponds to a different rate of population expansion, β. As a consequence of the reduction in height of the trees as β increases, the mutation rates must also be increased to obtain approximately the same total number of mutations for each situation. It can be seen that there is a block-like structure to the situation where $\beta = 0$, in that many sites appear to be perfectly correlated to each other as indicated by the strong presence of blue and dark red squares. This is a result of the ancestral branches being long in comparison to the more recent branches, and so mutations will either be perfectly associated with either a high frequency (dark red), or low frequency (blue). As the rate of population

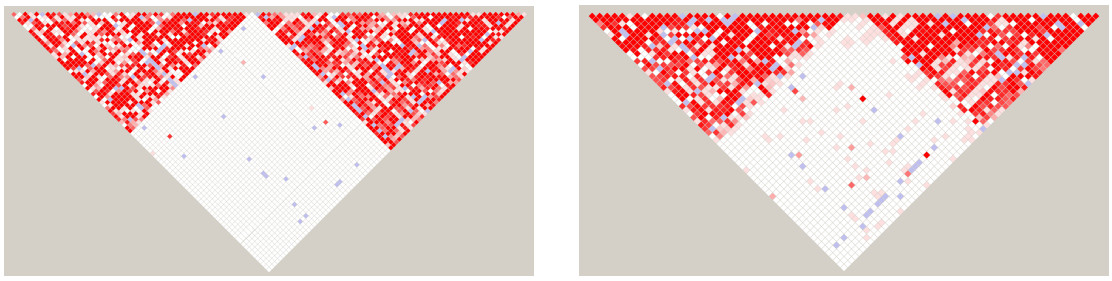
expansion increases, more mutations will only occur on terminal nodes and so the haplotypes will exist in lower frequency, and as a consequence there will be an increase of blue squares in the linkage plots. Figure C.3 illustrates a similar setting, but where a recombination rate of 2 has been chosen. It can be seen that there is a similar pattern of the linkage plots as observed previously, and that the increased recombination rate does not appear to substantially alter the observed linkage plots.



(a) $\beta = 0$ (b) $\beta = 10$ (c) $\beta = 50$

Figure C.3: Simulated linkage plots for various combinations of parameters (infinite sites, recombination rate = 2).

C.3 Two Independent Simulations



(a) Simulated (b) Real

Figure C.4: Simulated data of two independent ARGs, and the real data that it is intended to mimic.

Figure C.4(a) represents a simulation of haplotypes from two independent Ancestral Recombination Graphs that have subsequently been pasted together. It can be seen that there are a few positions where linkage is apparent in between

the two haplotype blocks, which are a result of chance. There does, however, appear to be more instances of linkage being apparent in the real data set in comparison to that of the simulated data, suggesting that the two haplotype blocks may not be entirely independent from each other. Rates of recombination in the simulations can be increased in an attempt to more closely the real data set. However, in the current formulation of the program the run-time will dramatically increase; and there also are memory limitations in the current formulation of the application. There are areas in which the efficiency of the PheGe-Sim application could potentially be improved, and with this increased efficiency larger regions of genetic data containing haplotype blocks and recombination hotspots could subsequently be simulated.

C.4 Code

The PheGe-Sim and PheGe-Find applications are available on request, along with copies of the other applications that are required. These application run on Windows XP, Vista and 7, although have only been extensively tested on Windows XP. A ReadMe file accompanying the applications details the specific procedures of running the code.

Bibliography

American Heart Association (2010). Average resting heart rate. www.americanheart.org/presenter.jhtml?identifier=3003997.

Aquadro, C., S. Desse, M. Bland, C. Langley, and C. Laurie-Ahleberg (1986). Molecular Population Genetics of the Alcohol Dehydrogenase Gene Region of Drosophila Melanogaster. *Genetics 114*, 1165–1190.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics 11*, 375–386.

Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics 7*, 781–791.

Balding, D., M. Bishop, and C. Cannings (2001). *Handbook of Statistical Genetics* (1st ed.). John Wiley and Sons Ltd.

Barrett, J. et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics 41*, 703–707.

Barrett, J., B. Fry, J. Maller, and M. Daly (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics 21*, 263–265.

Boerwinkle, E. and C. Sing (1986). Bias of the contribution of single locus effects to the variance of a quantitative trait. *American Journal of Human Genetics 39*, 137–144.

Bowman, A., E. Crawford, G. Alexander, and R. Bowman (2007). rpanel: simple Interactive Controls for R Functions Using the tcltk Package. *Journal of Statistical Software 17*, 1–18.

British Heart Foundation (2010). Average resting heart rate. http://www.bhf.org.uk.

Chadeau-Hyam, M., C. Hoggart, P. O'Reilly, J. Whittaker, M. De Iorio, and D. Balding (2008). Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics 9*, 364.

Clark, A. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology 27*, 321–333.

Clement, M., D. Posada, and K. Crandall (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology 9*, 1657–1660.

Donnelly, P. and S. Tavaré (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics 29*, 401–421.

Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature 450*, 203–218.

Durrant, C. and A. Morris (2005). Linkage disequilibrium mapping via cladistic analysis of phase-unknown genotypes and inferred haplotypes in the genetic analysis workshop 14 simulated data. *BMC Genetics 6*, (Suppl 1):S100.

Durrant, C., K. Zondervan, L. Cardon, S. Hunt, P. Deloukas, and A. Morris (2004). Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *American Journal of Human Genetics 75*, 3543.

Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Research 18*, 653–660.

Excoffier, L., J. Novembre, and S. Schneider (2000). Computer note. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity 91*, 506–509.

Fearnhead, P. (2006). SequenceLDhot: detecting recombination hotspots. *Bioinformatics 22*, 3061–3066.

Fearnhead, P. and P. Donnelly (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society Series B 64*, 657–680.

Felsenstein, J. (1988). Phylogenies from Molecular Sequences: Inference and Reliability. *Annual Review of Genetics 22*, 521–565.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. http://evolution.genetics.washington.edu/phylip.html.

Felsenstein, J., J. Archie, W. Day, W. Maddison, C. Meacham, F. Rohlf, and D. Swofford (2010). The Newick Tree Format. http://evolution.genetics.washington.edu/phylip/newicktree.html.

Felsenstein, J. and G. Churchill (1996). A hidden Markov Model Approach to Variation Among Sites in Rate of Evolution. *Molecular Biology and Evolution 13*, 93–104.

Fisher, R. (1930). *Genetical Theory of Natural Selection* (1st ed.). Clarendon Press.

Fitch, W, M. (1971). Toward defining the course of evolution: Minimum change for a specified tree topology. *Systematic Zoology 20*, 406–416.

Fitch, W. and E. Margoliash (1967). Construction of phylogenetic trees. *Science 155*, 279–284.

FlyBase (2010). FlyBase: A database of *Drosophila* genes & genomes. http://flybase.org/.

Gabriel, S. et al. (2002). The structure of haplotype blocks in the human genome. *Science 296*, 2225–2229.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian Data Analysis* (2nd ed.). Chapman and Hall.

Griffiths, R. and P. Marjoram (1996). Ancestral inference from samples of DNA sequence with recombination. *Journal of Computational Biology 3*, 479–502.

Gu, D., S. Shaoyong, D. Ge, S. Chen, J. Huang, B. Li, R. Chen, and B. Qiang (2006). Association study with 33 single-nucleotide polymorphisms in 11 candidate genes for hypertension in Chinese. *Hypertension 47*, 1147–1154.

Hedrick, P. (1987). Genetic disequilibrium measures: Proceed with caution. *Genetics 117*, 331–341.

Hein, J., M. Schierup, and C. Wiuf (2005). *Gene Genealogies, Variation and Evolution. A Primer in Coalescent Theory* (1st ed.). Oxford University Press.

Hellenthal, G. and M. Stephens (2007). msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics 23*, 520–521.

Hoggart, C., M. Chadeau-Hyam, T. Clark, R. Lampariello, J. Whittaker, M. De Iorio, and D. Balding (2007). Sequence-level population simulations over large genomic regions. *Genetics 177*, 1725–1731.

Howie, B., P. Donnelly, and J. Marchini (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics 5*, e1000529.

Hudson, R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology 23*, 183–201.

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

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

Huelsenbeck, J. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics 17*, 754–755.

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell 72*, 971–983.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature 437*(430), 1299–1320.

International HapMap Consortium (2010). Recombination rates. http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/.

Jeffreys, H. (1939). *Theory of Probability* (1st ed.). Clarendon Press.

Jukes, T. and C. Cantor (1969). *Evolution of protein molecules in H.N.Munroe, ed., 'Mammalian Protein Metabolism'* (1st ed.). Academic Press.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kerem, B., J. Rommens, J. Buchanan, D. Markiewicz, T. Cox, A. Chakravarti, M. Buchwald, and L. Tsui (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science 245*, 1066–1073.

Kimmel, G., R. Karp, M. Jordan, and E. Halperin (2008). Association mapping and significance estimation via the coalescent. *American Journal of Human Genetics 83*, 675683.

Kingman, J. (1982a). The coalescent. *Stochastic Processes and their Applications 13*, 235–248.

Kingman, J. (1982b). On the genealogy of large populations. *Journal of Applied Probability 19*, 27–43.

Korner, P. (2007). *Essential Hypertension and its Causes: Neural and Non-Neural Mechanisms* (1st ed.). Oxford University Press.

Lander, E. (1996). The new genomics: global views of biology. *Science 274*, 536–539.

Lewontin, R. (1964). The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics 49*, 49–67.

Li, Y., C. Willer, J. Ding, P. Scheet, and G. Abecasis (2010). Mach: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology 34*, 816–834.

Liang, L., S. Zöllner, and G. Goncalo (2007). Genome: a rapid coalescent-based whole genome simulator. *Bionformatics 23*, 1565–1567.

Lowe, C., J. Cooper, T. Brusko, N. Walker, D. Smyth, R. Bailey, K. Bourget, V. Plagnol, S. Field, M. Atkinson, D. Clayton, L. Wicker, and J. Todd (2007). Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nature Genetics 39*, 1074–1082.

Mailund, T., M. Schierup, C. Pederson, P. Mechlenborg, J. Madsen, and L. Schauser (2005). CoaSim: A flexible environment for simulating genetic data under coalescent models. *BMC Bioinformatics 6*, 252.

Mancia, G., A. Ferrari, L. Gregorini, M. Ferrari, C. Bianchini, L. Terzoli, G. Leonetti, and A. Zanchettie (1980). Effects of prazosin on autonomic control of circulation in essential hypertension. *Hypertension 2*, 700–707.

Mancia, G., R. Sega, C. Bravi, G. De Vito, F. Valgussa, G. Cesana, and A. Zanchetti (1995). Ambulatory blood pressure normality: results from the PAMELA study. *Hypertension 13*, 1377–1390.

Marchini, J., D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, S. Qin, Z, H. Munro, G. Abecasis, and P. Donnelly (2006). A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics 78*, 437–450.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics 39*, 906–913.

McVean, G., P. Awadalla, and P. Fearnhead (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics 160*, 1231–1241.

Minichiello, M. and R. Durbin (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *American Journal of Human Genetics 79*, 910–922.

Morgan, T., A. Sturtevant, H. Muller, and C. Bridges (1915). *The Mechanism of Mendelian Heredity* (1st ed.). Henry Holt.

Newton-Cheh, C. et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genetics 41*, 666–676.

Ott, J. (1999). *Analysis of Human Genetic Linkage* (3rd ed.). The John Hopkins University Press.

Padmanabhan, S., C. Menni, W. Lee, S. Laing, P. Brambilla, R. Sega, R. Perego, G. Grassi, G. Cesana, C. Delles, G. Mancia, and A. Dominiczak (2010). The effects of sex and method of blood pressure measurement on genetic associations with blood pressure in the PAMELA study. *Hypertension 28*, 465–477.

Page, R. (1996). Treeview: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences 12*, 357–358.

Paradis, E., J. Claude, and K. Strimmer (2004). APE: analysis of phylogenetics and evolution in R language. *Bioinformatics 20*, 289–290.

Pritchard, J. and N. Cox (2002). The allelic architecture of human disease genes: common disease-common variant . . . or not? *Human Molecular Genetics 11*, 2417–2423.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Roeder, K., S. Bacanu, L. Wasserman, and B. Devlin (2006). Using linkage genome scans to improve power of association in genome scans. *American Journal of Human Genetics 78*, 243–252.

Rommens, J., M. Iannuzzi, B. Kerem, M. Drumm, G. Melmer, M. Dean, R. Rozmahel, J. Cole, D. Kennedy, N. Hidaka, M. Zsiga, M. Buchwald, J. Riordan, L. Tsui, and F. Collins (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science 245*, 1059–1065.

Rung et al. (2009). Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature Genetics 41*, 1110–1115.

Schaid, D. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology 27*, 348–364.

Scheet, P. and M. Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics 78*, 629–644.

Senn, S. (1997). *Statistical Issues in Drug Development* (1st ed.). John Wiley and Sons.

Servin, B. and M. Stephens (2007). Imputation-based analysis of association studies: candidate genes and quantitative traits. *PLoS Genetics 3*, 1296–1308.

Sladek, R. et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature 445*, 881–885.

Spencer, C., Z. Su, P. Donnelly, and J. Marchini (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics 5*, e1000477.

Stephens, M. and D. Balding (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics 10*, 681–690.

Stephens, M., N. Smith, and P. Donnelly (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics 68*, 978–989.

Swofford, D. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. http://paup.csit.fsu.edu.

Templeton, A., E. Boerwinkle, and C. Sing (1987). A cladistic analysis of phentotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic Theory and an analysis of alcohol dehydrogenase acivity in *Drosphila*. *Genetics 117*, 343–351.

Templeton, A., K. Crandall, and C. Sing (1992). A cladistic analysis of phentotypic associations with haplotypes inferred from restriction endonuclease mapping. III. Cladogram Estimation. *Genetics 132*, 619–633.

Templeton, A., T. Maxwell, D. Posada, J. Stengard, E. Boerwinle, and C. Sing (2005). Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics 169*, 441–453.

Templeton, A., C. Sing, A. Kessling, and S. Humphries (1988). A cladistic analysis of phentotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics 120*, 1145–1154.

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

Trégouët,D.A and others (2009). Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nature Genetics 41*, 283–285.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4th ed.). Springer.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology 33*, 79–86.

Wakeley, J. (2008). *Coalescent Theory: An Introduction* (1st ed.). Roberts & Company Publishers.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology 7*, 256–276.

Williams, B., N. Poulter, M. Brown, M. Davis, G. McInnes, J. Potter, P. Sever, and S. Thom (2004). British Hypertension Society guidelines for hypertension management 2004 (BHS-IV): summary. *British Medical Journal 328*, 634–640.

World Health Organization (1999). 1999 World Health Organization - International Society of Hypertension Guidelines for the Management of Hypertension. *Journal of Hypertension 17*, 151–183.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics 16*, 97–159.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature 447*, 661–678.

Zöllner, S. and J. Pritchard (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics 169*, 1071–1092.