STATISTICAL METHODS FOR THE ANALYSIS OF GENETIC ASSOCIATION STUDIES

D.PHIL THESIS AND ABSTRACT

Zhan Su

Balliol College, University of Oxford.

September 22, 2008

First and foremost, I must thank Jonathan Marchini, whose dedication I have the utmost respect for and I am very grateful for his guidance. I would also like to thank Peter Donnelly for his input into my work.

The people at the Statistics department of Oxford University have made it a wonderful place to work. In particular, I am indebted to Niall Cardin, who has been extremely generous with his time, to correct my various misconceptions and to answer every tedious question. I would also like to thank Teresa Ferreira and Joanne Gale, for brightening up the office and for graciously putting up with me.

I would like to thank all my friends from the Balliol College MCR and Boat Club, of whom there are too many to name here, for four fantastic years.

Finally, I would like to dedicate this thesis to my parents, who, at every chance, have provided me with the very best opportunities in life.

Abstract

One of the main biological goals of recent years is to determine the genes in the human genome that cause disease. Recent technological advances have realised genome-wide association studies, which have uncovered numerous genetic regions implicated with human diseases. The current approach to analysing data from these studies is based on testing association at single SNPs but this is widely accepted as underpowered to detect rare and poorly tagged variants. In this thesis we propose several novel approaches to analysing large-scale association data, which aim to improve upon the power offered by traditional approaches. We combine an established imputation framework with a sophisticated disease model that allows for multiple disease causing mutations at a single locus. To evaluate our methods, we have developed a fast and realistic method to simulate association data conditional on population genetic data. The simulation results show that our methods remain powerful even if the causal variant is not well tagged, there are haplotypic effects or there is allelic heterogeneity. Our methods are further validated by the analysis of the recent WTCCC genome-wide association data, where we have detected confirmed disease loci, known regions of allelic heterogeneity and new signals of association. One of our methods also has the facility to identify the high risk haplotype backgrounds that harbour the disease alleles, and therefore can be used for fine-mapping. We believe that the incorporation of our methods into future association studies will help progress the understanding genetic diseases.

Contents

1 Introduction **1**

2 Detecting Association: an Overview **5**

2.1 Single marker association tests 7

2.1.1 Frequentist tests 8

2.1.2 Bayesian tests 13

2.1.3 Comparison of P values and Bayes Factors 16

2.2 Haplotype tests 18

2.2.1 Haplotype sharing tests 20

2.2.2 Clustering haplotypes 22

2.2.3 The Multimarker Predictor Test 25

2.2.4 Imputation 26

2.2.5 Estimating haplotype phase 32

2.3 The coalescent model 32

2.4 Our approach 36

3 HAPGEN: A Method for Simulating Realistic Association Data 37

3.1 Introduction 37

3.2 Single mutation model 40

3.2.1 Step 1: Selecting a disease SNP 41

3.2.2 Step 2: Simulating alleles at the disease SNP 41

3.2.3 Step 3: Simulating alleles at other SNPs 43

3.3 Two mutation model 45

3.3.1 Step 2: Simulating alleles at the disease SNPs 46

3.3.2 Step 3: Simulating alleles between disease SNPs 47

3.4 Illustration of HAPGEN 49

3.5 Comparing HAPGEN to coalescent simulations 51

3.6 Discussion 53

4 Methods for Detecting Association in Haplotype Data 56

4.1 Introduction 56

4.2 Notations 57

4.3 A haplotype similarity measure 58

4.3.1 Similarity at typed SNPs 59

4.3.2 Similarity at untyped SNPs 61

4.3.3 The similarity matrix 61

4.4 HAPSEARCH: a non-parametric association test 63

4.5 HAPCLUSTER: a parametric association test 66

4.5.1 Constructing haplotype trees 68

4.5.2 Constructing basis clusters and groupings 71

4.5.3 Analysing groupings 72

4.5.4 Model choices 76

4.5.5 Model priors 78

4.5.6 The test statistics 81

4.6 Simulation study 83

4.6.1 Simulation of case-control data 83

4.6.2 Assessing HAPSEARCH 85

4.6.3 Assessing HAPCLUSTER 88

4.6.4 Comparing to SNP based methods 95

4.6.5 Stratifying by type of disease SNP 97

4.7 Discussion 99

5 GENECLUSTER: A Powerful Method for Analysing Association Data 103

5.1 Method 105

5.1.1 Notations 105

5.1.2 Step 1: Constructing genealogy for the panel haplotypes . . . 106

5.1.3 Step 2: Constructing genealogy for the case-control sample . . 107

5.1.4 Step 3: Testing for association 114

5.2 Simulation Study 132

5.2.1 Calculating power 133

5.2.2 Computation cost 135

5.2.3 Detecting association in a single mutation disease model . . . 136

5.2.4 Detecting association in a two mutation disease model 150

5.2.5 Detecting regions of allelic heterogeneity 167

5.3 Discussion 168

6 Analysis of the WTCCC Data 177

6.1 Introduction 177

6.2 Method 179

6.3 Results 181

6.3.1 Comparing GENECLUSTER with SNP based approaches . . 187

6.3.2 Detecting new signals of association 189

6.3.3 Detecting regions of allelic heterogeneity 196

6.4 Discussion 202

7 Application to Fine-mapping **218**

7.1 Introduction 218

7.2 Sampling strategies 219

7.2.1 Strategy A: random sampling 220

7.2.2 Strategy B: sampling based on single SNP tests 221

7.2.3 Strategy C: sampling based on GENECLUSTER 224

7.3 Simulation study 227

7.3.1 Calculating power 227

7.3.2 Sampling in a single mutation disease model 228

7.3.3 Sampling in a two mutation disease model 230

7.4 Discussion 235

8 Summary and Future Work **239**

8.1 Summary 239

8.2 Future Work 242

A Supplementary Figures For Chapter 4 **255**

B Supplementary Figures For Chapter 5 **260**

C TREESIM **270**

Chapter 1

Introduction

It has long been established that genetics is a major determining factor of human phenotype. One of the main biological goals of recent years is to determine the genes that impact specific phenotypes, in particular the ones that cause disease. It is hoped that the discovery of these genes will increase our understanding of disease, which can lead to cures, and allow the identification of highly susceptible individuals, which can lead to preventions.

The traditional approach to identifying chromosomal regions that contain disease genes is linkage analysis in families. These chromosomal regions, identified by their non-random transmission to affected offspring, are typically several centimorgans in size and may contain hundreds of genes. Whole genomes can be scanned relatively easily and this has been very successful at identifying low frequency alleles with a strong effect (Kerem et al. (1989) for example). The success of this approach reflects the power of linkage analysis when applied to Mendelian phenotypes, which are characterised by a near one-to-one correspondence between genotypes at a single

locus and the observed phenotype.

The detection of genetic factors for common familial disorders with no clear Mendelian inheritance patterns has however proved more challenging. It is postulated that these diseases, such as Crohn's disease, bipolar disorder, and diabetes, involve many genes of smaller effects. There have been numerous reports of genes or loci that may confer an elevated risk of these disorders, but few of these findings have been replicated. The modest nature of the genetic effects for these disorders are likely to explain the contradictory and inconclusive claims about their identification but the relative magnitude of their attributable risk may be large because they are frequent in the population (Risch, 2000).

Risch and Merikangas (1996) argued that for loci of moderate genetic relative risk linkage analysis will never lead to their identification because the number of families required is not practically achievable. However, direct tests of association between genotype and phenotype at a disease locus can still be quite strong. The completion of the Human Genome Project, the development of comprehensive databases of SNPs, substantial catalogues of human haplotype variation and technological advances that allow the cost-effective collection of genotypes have made genome-wide association studies possible. Recently, genome-wide association studies (Wang et al., 2005; Hirschhorn and Daly, 2005) have been extremely successful and have identified and replicated numerous genomic regions associated with complex diseases, for example The Wellcome Trust Case Control Consortium (2007), Zeggini et al. (2007) and Duerr et al. (2006) to name just a few. These studies assay a very dense set of markers (100,000+) across the genome, in thousands of case and control individuals, using one of the commercially available genotyping chips. A direct test of association between genotypes and a given phenotype, is conducted at each marker. The large

sample size confers greater power to detect genes of moderate effects but will also suffer from issues such as genotyping error, population structure and ascertainment bias, which the investigator needs to control for. Another big challenge, which is the focus of this thesis, is to develop sophisticated, but computationally tractable, methodology to extract more power to detect causal variants that traditional single marker tests might miss. The structure of this thesis is as follows.

We will begin in Chapter 2 with a literature review of the current methods used to analyse association studies.

In Chapter 3 we will introduce HAPGEN, a new approach to simulating realistic population genetic data, which we use for simulation studies in later chapters to evaluate and compare the methods that we develop.

In Chapter 4 we describe new approaches to association testing, based on a linkage disequilibrium (LD) model introduced by Li and Stephens (2003). The methods that we describe provide the inspirations for the work presented in the following chapter.

The most important contributions of our work to the field of genetic association studies are described in Chapters 5 to 7. In Chapter 5 we introduce GENECLUSTER, which is a method for detecting association in large scale genome-wide association studies. We combine an established imputation framework with a sophisticated disease model that can allow for multiple disease causing mutations at a single locus. Our simulation study demonstrates that GENECLUSTER can provide a boost in power for detecting rare and untagged causal variants and under allelic heterogeneity. We validate GENECLUSTER in Chapter 6 by applying it to the data from the WTCCC study (The Wellcome Trust Case Control Consortium, 2007), where

we find signals of association that were not found by previous single SNP analyses. Our results also yield useful inferences about the underlying disease model at a disease locus, which can be used for fine-mapping studies and we will discuss this in Chapter 7.

We will end our thesis with a summary of our work and a discussion on how it can be extended in the future.

Chapter 2

Detecting Association: an Overview

Association analysis is used to detect or locate disease causing mutations. The data typically consist of genotypes from unrelated individuals at a subset of common biallelic SNPs in a region, together with phenotype information on the individuals. The investigator aims to use the data to detect or locate the variants that impact upon the trait of interest.

For complex traits, a disease mutation has a relatively modest impact on the total disease risk. The signal of association will be further weakened if the true causal variant is not in the set of typed SNPs nor strongly correlated with any of the typed SNPs (Pritchard and Przeworski, 2001). Environmental factors, dominance, and epistasis mean that some control individuals can carry the disease allele and some case individuals can carry the protective allele, which introduce substantial noise to the relationship between phenotype and genotype (Zollner and Pritchard, 2005).

Finally, allelic heterogeneity, the presence of multiple distinct disease alleles in close proximity, will distort the association between phenotype and genotype at individual SNPs and can cause problems for association mapping (Pritchard and Cox, 2002). It is therefore important to develop methods that extract as much information from the data as possible and detect signals of association in the presence of phenocopies (case individuals carrying the protective allele), allelic heterogeneity and when the true disease allele is only in partial linkage disequilibrium with a typed SNP.

Currently, the literature in this area is large and complex, and there is no consensus on the best method (Balding, 2006a). The relative performance of different methods will depend heavily on the parameters of the simulation studies that are routinely used to assess them. In addition, methods are focused on two different but related goals (Zollner and Pritchard, 2005):

- detection – to detect the presence of a causal variant within a region of interest, and
- localisation – to locate the causal variant given that one exists in the region of interest.

The goal of localising the causal variant is typically reserved for fine-mapping studies, which follow up the regions found in association studies or linkage studies. With the advent of dense databases of SNPs and genotyping chips with dense sets of markers, the traditional meaning of fine-mapping has changed from the kind of scenarios described in papers such as Morris et al. (2000), which try to find the location of a disease mutation. In the modern context, fine-mapping (such as the WTCCC+ study, which we will describe in Chapter 7) means going to the sequence level data with a goal of characterising the disease model. In this thesis we are focused on

detecting regions that contain causal variants from data gathered in genome-wide association studies, where individuals are genotyped at a dense set of SNPs and exhibit a binary (case/control) phenotype. We will also discuss the utility of our proposed methods to fine-mapping but it is not the main focus.

For the rest of this chapter we will outline a number of different approaches relevant to the rest of this thesis and describe their strengths and weaknesses.

2.1 Single marker association tests

The simplest association test is to independently test the data gathered at each typed marker. Let us suppose that in a sample of M control and N case individuals, the phenotype of the ith individual is $\Phi_i \in \{0 = \text{control}, 1 = \text{case}\}$ and the alleles at each SNP are either 0 or 1 so that the genotype of the ith individual is $G_i \in \{0, 1, 2\}$. The data can be summarised as

Phenotype	G=0	G=1	G=2
Control ($\Phi = 0$)	m_0	m_1	m_2
Case ($\Phi = 1$)	n_0	n_1	n_2

Table 2.1: Summary table for genotype data at a SNP.

There are two classes of tests: Frequentist and Bayesian, and Marchini et al. (2007) give a comprehensive account of both, which we will summarise here.

2.1.1 Frequentist tests

Retrospective likelihood

We will first describe testing under a disease model where the relative risk increases multiplicatively in terms of genotype, i.e.

$$\log(\Pr(\Phi = 1|G = g)) = \mu + g\gamma. \tag{2.1}$$

Based on Hardy-Weinberg Equilibrium (HWE) (Thomas, 2004), which assumes that the frequencies of each allele in successive generations remain constant, the allele frequencies for a rare disease under this model are approximated by:

$$\Pr(G = g|\Phi = 0) = \begin{cases} (1-p)^2 & \text{if } g = 0, \\ 2p(1-p) & \text{if } g = 1, \\ p^2 & \text{if } g = 2, \end{cases} \tag{2.2}$$

and

$$\Pr(G = g|\Phi = 1) = \begin{cases} (1-q)^2 & \text{if } g = 0, \\ 2q(1-q) & \text{if } g = 1, \\ q^2 & \text{if } g = 2. \end{cases} \tag{2.3}$$

Deviations away from HWE can be caused by non-random mating, limited population size and selection. The impact of selection means that we can only assume HWE in the control population if the disease is rare. Therefore, any test based on HWE should not be used to test for association with dominant or recessive diseases.

If the case and control individuals are sampled conditional on their phenotype status,

as is normally the case in case-control association studies, then the natural likelihood for the data is the retrospective likelihood:

$$L(\theta) = \Pr(\mathbf{G}|\mathbf{\Phi}, \theta) = \prod_{i=1}^{M+N} \Pr(G_i|\Phi_i), \tag{2.4}$$

where $\mathbf{\Phi} = \{\Phi_1, \ldots, \Phi_{M+N}\}$, $\mathbf{G} = \{G_1, \ldots, G_{M+N}\}$ and $\theta = (p, q)$.

To test for association, Frequentists test the hypotheses

$$H_0 : p = q \qquad \text{vs} \qquad H_1 : p \neq q,$$

where H_0 and H_1 are the null and alternative hypotheses respectively. This can be done using the Maximum Likelihood Ratio Test (MLRT) with the test statistic

$$\lambda = \frac{\text{Max}_{H_0} L(\theta)}{\text{Max}_{H_1} L(\theta)}$$

and under the null $-2\log\lambda \sim \chi^2_1$ as $M, N \to \infty$. The maximum likelihoods are given by

$$\begin{aligned} \text{Max}_{H_0} L(\theta) &= \Big(\frac{N_0}{N_0+N_1}\Big)^{N_0}\Big(\frac{N_1}{N_0+N_1}\Big)^{N_1}\Big(\frac{M_0}{M_0+M_1}\Big)^{M_0}\Big(\frac{M_1}{M_0+M_1}\Big)^{M_1}, \quad (2.5) \\ \text{Max}_{H_1} L(\theta) &= \Big(\frac{M_0+N_0}{M_0+M_1+N_0+N_1}\Big)^{M_0+N_0}\Big(\frac{M_1+N_1}{N_0+N_1+M_0+M_1}\Big)^{M_1+N_1}, \quad (2.6) \end{aligned}$$

where $N_0 = 2n_0 + n_1$, $N_1 = 2n_2 + n_1$, $M_0 = 2m_0 + m_1$ and $M_1 = 2m_2 + m_1$ are the counts of the 0 and 1 alleles in the case and control samples in Table 2.1. The test statistic can therefore be expressed in terms of allele counts and this test is equivalent to testing allele counts against disease status, which is referred to as the Allelic Test.

The test is often summarised using a P value, which is the probability of obtaining a test statistic at least as extreme as the one observed under the null model.

Prospective likelihood

An alternative approach is to consider the prospective likelihood,

$$L(\theta) = \Pr(\mathbf{\Phi}|\mathbf{G}, \theta) = \prod_{i=1}^{N} \Pr(\Phi_i = 1|G_i)^{\Phi_i}(1 - \Pr(\Phi_i = 1|G_i))^{1-\Phi_i}. \tag{2.7}$$

The prospective likelihood is preferred to the retrospective likelihood because it can naturally model genotype and covariant information as determinants of phenotype. However, treating individual phenotype as a random variable is incorrect under a case-control study where individuals are genotyped conditional on their phenotype. Prentice and Pyke (1979) showed that the maximum likelihood estimators and asymptotic covariance matrix obtained from the retrospective likelihood are the same as those obtained from the prospective likelihood. This implies that the significance tests based on both the retrospective and prospective likelihoods will be equivalent asymptotically and very similar for sufficiently large sample sizes (Marchini et al., 2007).

A popular disease model under the prospective likelihood is the log additive disease model where the odds of disease change multiplicatively with genotype and is expressed in a logistic regression framework:

$$\log\big(\frac{\Pr(\Phi = 1|G = g)}{1 - \Pr(\Phi = 1|G = g)}\big) = \mu + g\gamma. \tag{2.8}$$

The odds ratios of disease for individuals with genotypes 1 and 2 (relative to in-

dividuals with the 0 genotype) are e^{γ} and $(e^{\gamma})^2$, so the model is multiplicative on the odds scale and additive on the log odds scale. This model is often referred to as the additive model or the multiplicative model with no real consensus. Further, the same terms are also used to refer to a disease model specified for relative risk, which we defined earlier. It is therefore important to specify the scale on which the effect is measured. Other disease models, for example dominant and recessive, are possible and simply require a different coding of the genotype vector $\mathbf{G}$.

Under this construction, the alternative model of association has the parameters $\theta_1 = (\mu, \gamma)$ and the null model of no association has the parameters $\theta_0 = (\mu, 0)$. The Score Test is widely used to test between the two models and is based on the distribution of the Score, $U(\theta) = \frac{d(l(\theta))}{d\theta}|_{\theta}$ where $l(\theta) = \log L(\theta)$, under the null model. The Score Test is based on the asymptotic result that under the null model

$$U(\hat{\theta}_0) \sim N(0, I(\hat{\theta}_0)),$$

where $I(\hat{\theta}_0) = -\frac{d^2(l(\theta))}{d\theta^2}|_{\theta=\hat{\theta}_0}$, which implies that the Score Statistic,

$$S = U(\hat{\theta}_0)^T I^{-1}(\hat{\theta}_0) U(\hat{\theta}_0),$$

is asymptotically distributed as χ^2_1, where $\hat{\theta}_0 = (\hat{\mu}, 0)$ and $\hat{\mu}$ is the Maximum Likelihood Estimator of μ with γ fixed at 0. The Score Test is equivalent to the Cochran-Armitage Trend Test, which is a popular method used to compare genotype frequencies between cases and controls (Clayton, 2001).

The Score Test is preferred over the MLRT because it only requires the evaluation of the likelihood under the null, as opposed to the maximisation of the log-likelihood,

and so is less computationally intensive. It is also preferred to the Allelic Test because it is valid even if HWE is violated. Guedj et al. (2008) demonstrated that the Allelic Test and the Trend Test are asymptotically equivalent when HWE exists.

Detecting untyped variants

In an association study, the set of SNPs on a genotyping chip is not guaranteed to include the true causal variant. However, the tendency for parental alleles located close together on the same chromosome to be transmitted together to the offspring creates a correlation structure in the genome. This phenomenon is called Linkage Disequilibrium (LD) (Thomas, 2004). A combination of alleles are in LD if they occur more or less frequently than would be expected based on their marginal frequencies. Therefore, the effects of a causal variant can still be detected if it is in LD with a typed SNP. For a single disease SNP under the log additive disease model for relative risk, the Trend test statistic at a typed marker will follow a noncentral χ^2 distribution with the noncentrality parameter η (Chapman et al., 2003) such that:

$$\mathbb{E}[\eta] \propto \frac{NM(p(1-p))^2(\alpha-1)^2r^2}{(N+M)\bar{p}(1-\bar{p})((\alpha-1)p+1)^2}, \tag{2.9}$$

where M and N are the sizes of the case and control samples respectively, α is the relative risk, $\bar{p} = \frac{Mp+Np'}{N+M}$, p and p' are the allele frequencies of the causal variant in the control and case samples respectively and r^2 is the correlation coefficient between the marker and causal SNP. In other words, for a fixed sample size the signal strength of the single marker test at a typed marker is large when the true disease SNP is closely correlated to it and the minor allele frequency at the disease SNP is high.

2.1.2 Bayesian tests

The single marker Bayesian test is the alternative to the frequentist hypothesis test that tests for association at a typed SNP. The Bayes Factor is defined as

$$BF = \frac{\Pr(D|M_1)}{\Pr(D|M_0)} = \frac{\int \Pr(D|\theta_1, M_1)\Pr(\theta_1|M_1)d\theta_1}{\int \Pr(D|\theta_0, M_0)\Pr(\theta_0|M_0)d\theta_0}, \tag{2.10}$$

where D is the observed data (genotype, phenotype, covariates), M_0 and M_1 are the null model of no association and the model of association respectively, and θ_0 and θ_1 are the parameters of the models M_0 and M_1. The Bayes Factor is the Bayesian equivalent of the P value and is the factor by which the prior odds of association are changed in light of the data to produce the posterior odds of association,

$$\begin{aligned} \text{Posterior Odds of Association} &= BF \times \text{Prior Odds of Association} \\ \frac{\Pr(M_1|D)}{\Pr(M_0|D)} &= \frac{\Pr(D|M_1)}{\Pr(D|M_0)}\frac{\Pr(M_1)}{\Pr(M_0)}. \end{aligned}$$

Instead of maximising the likelihood, Bayesians integrate out unknown parameters of each model under a prior distribution, which allows the investigator to incorporate prior knowledge into the result.

The standard approach is to use the prospective likelihood Equation (2.7), and under model M_0

$$\theta_0 = (\mu) \qquad \log(\frac{p_i}{1-p_i}) = \mu \tag{2.11}$$

and under model M_1

$$\theta_1 = (\mu, \gamma) \qquad \log(\frac{p_i}{1-p_i}) = \mu + \gamma G_i,$$

where $p_i = \Pr(\Phi_i = 1|G_i)$. Marchini et al. (2007) use normal distributions for $\Pr(\mu|M_0)$, $\Pr(\mu|M_1)$ and $\Pr(\gamma|M_1)$, and offers a discussion on the suitable values for their mean and variance.

A computationally more efficient approach is to use a haploid model (Balding, 2006b). If $H_{i,1}$ and $H_{i,2}$ are the alleles on each chromosome of individual i so that $G_i = H_{i,1} + H_{i,2}$, then the disease model can be parametrised by $\theta = (p, q)$, where for $j = 1, 2$:

$$\begin{aligned} p &= \Pr(\Phi_i = 1|H_{i,j} = 0), \\ q &= \Pr(\Phi_i = 1|H_{i,j} = 1), \end{aligned}$$

so that under the null model M_0, $p = q$, and under the alternative model M_1, $p \neq q$. Under this framework the phenotypes of each pair of haplotypes are modelled as independent even though they come from the same individual. This allows the likelihood to simplify to a simple binomial likelihood for p and q:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N}\prod_{j=1}^{2} \Pr(\Phi_i|H_{i,j}) \\ &= p^{2n_0+n_1}(1-p)^{2m_0+m_1}q^{2n_2+n_1}(1-q)^{2m_2+m_1}. \end{aligned} \tag{2.12}$$

A conjugate beta prior for these parameters can then be used, which facilitates the exact calculation of the integrals in Equation (2.10). For example, if

$$\Pr(\theta|M_0) = \frac{1}{\beta(a,b)}p^{(a-1)}(1-p)^{(b-1)} \tag{2.13}$$

and

$$\Pr(\theta|M_1) = \frac{1}{\beta(a,b)}p^{(a-1)}(1-p)^{(b-1)}\frac{1}{\beta(a,b)}q^{(a-1)}(1-q)^{(b-1)}, \tag{2.14}$$

where $\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, then

$$BF = \frac{\beta(a+n_1+2n_2, b+m_1+2m_2)\beta(a+n_1+2n_0, b+m_1+2m_0)}{\beta(a,b)\beta(a+2N, b+2M)}, \qquad (2.15)$$

where M and N are the number of control and case individuals respectively, and m_0, m_1, m_2, n_0, n_1 and n_2 are defined in Table 2.1.

Although the haploid model is computationally convenient, to use it one has to assume a log additive disease model for relative risk. For example, the homozygote penetrance for the 0 allele is given by

$$\begin{aligned}
\Pr(\Phi_i = 1|G_i = 0) &= \sum_{H_{i,1}\in\{0,1\}} \sum_{H_{i,2}\in\{0,1\}} \Pr(\Phi_i = 1|H_{i,1}, H_{i,2}, G_i = 0)\Pr(H_{i,1}, H_{i,2}|G_i = 0) \\
&= \Pr(\Phi_i = 1|H_{i,1} = 0, H_{i,2} = 0, G_i = 0) \\
&= \frac{\Pr(H_{i,1} = 0, H_{i,2} = 0, G_i = 0|\Phi_i = 1)\Pr(\Phi_i = 1)}{\Pr(H_{i,1} = 0, H_{i,2} = 0, G_i = 0)} \\
&= \frac{\Pr(H_{i,1} = 0|\Phi_i = 1)\Pr(H_{i,2} = 0|\Phi_i = 1)\Pr(\Phi_i = 1)}{\Pr(H_{i,1} = 0)\Pr(H_{i,2} = 0)} \\
&= \frac{\Pr(\Phi_i = 1|H_{i,1} = 0)\Pr(\Phi_i = 1|H_{i,2} = 0)}{\Pr(\Phi_i = 1)} \\
&= \frac{p^2}{\Pr(\Phi_i = 1)}
\end{aligned}$$

Similarly, $\Pr(\Phi_i = 1|G_i = 1) = \frac{pq}{\Pr(\Phi_i=1)}$ and $\Pr(\Phi_i = 1|G_i = 2) = \frac{q^2}{\Pr(\Phi_i=1)}$. Therefore, we have that the genotype relative risks increase multiplicatively with the number of disease alleles and the heterozygote relative risk is equal to the haplotype relative risk: $\frac{\Pr(\Phi_i=1|G_i=2)}{\Pr(\Phi_i=1|G_i=0)} = (\frac{\Pr(\Phi_i=1|G_i=2)}{\Pr(\Phi_i=1|G_i=1)})^2 = (\frac{q}{p})^2$.

2.1.3 Comparison of *P* values and Bayes Factors

The majority of genome-wide association studies conducted today report their results in terms of P values. A problem with P values is that they summarise the observed data in terms of how likely (or unlikely) it is under the null model. P values under the null have an uniform distribution, so for a given threshold, α, the null model will be rejected, and hence the alternative model favoured, with frequency α without regard to how much evidence there is for the alternative model. Therefore, choosing an appropriate significance threshold becomes difficult. As an illustration, if we apply a common significance threshold to all SNPs in a genome-wide association study, then the type 2 error rate is typically lower for SNPs with high minor allele frequency (MAF) since there is more power to detect those SNPs, and the type 2 error rate for rare SNPs is typically high. The investigator therefore has implicitly assumed that a higher type 2 error rate is acceptable for rare SNPs.

Bayes Factors, on the other hand, have the natural interpretation as the factor by which the prior odds of association are changed in light of the data, so the point at which a Bayes Factor becomes "significant" depends upon the prior odds of association. The use of Bayes Factors has, however, been limited by the need to integrate out the unknown parameters in the models, which can be computationally expensive, and investigators have traditionally been sceptical about the subjectivity involved in choosing a prior. For example, if the same test is performed with two different priors and one gives a "significant" result but the other does not, how should the investigator interpret them? However, Bayesians argue that one will always have prior information on the parameters within each model and the incorporation of the prior allows the investigator flexibility to incorporate this information, which can

lead to increased power (Marchini et al., 2007). For example, to detect association with a complex disease one might expect odds ratios of between 0.5 and 2.0 to be likely, based on previous studies (e.g. The Wellcome Trust Case Control Consortium (2007)) and should specify a prior that places most probability mass on those values.

Servin and Stephens (2007) proposed an approach based on Bayesian regression to model phenotype in terms of the data at a set of SNPs. They used a carefully specified set of priors on the effect size at a single SNP and the number of causal SNPs in their set. In the special case where only data at a single SNP is considered, the test is equivalent to a single SNP Bayesian association test. Through simulations of a number of disease models and study designs they showed that the single SNP Bayes Factor has a number of advantages over the single SNP P value; they include (i) the Bayes Factor takes into account the informativeness of each SNP, in particular, SNPs with small MAF are typically less informative and therefore the Bayes Factor at those SNPs tend to be approximately 1 to reflect a lack of evidence under both the null and the alternative models, (ii) the Bayes Factor provides a principled way to take into account any prior information on each SNP, for example, the prior odds of association can be adjusted to be greater if the SNP is located in a candidate gene, and (iii) averaging single SNP Bayes Factors provides an approach to combine information at multiple SNPs, which can not be done with P values. Servin and Stephens (2007) also point out, though, that Bayes Factors are only ever informative if the models and priors under the alternative and null are correctly specified with respect to the real underlying models. However, the hope is that if the priors and models are sufficiently accurate then the resulting Bayes Factors will lead to accurate inferences.

In general, both the ranking and significance of SNPs will differ when assessed using

Bayes Factors and P values. Wakefield (2008) formulated a Bayes Factor based on the asymptotic properties of maximum likelihood estimators. Importantly, he was able to show that the rankings of SNPs using this "asymptotic" Bayes Factor will be identical to the rankings in terms of the P values from a Frequentist test if a prior of a specific form is used. This prior assumes that the effect sizes are larger for rare SNPs and therefore the Frequentist approach is equivalent to the Bayesian approach assuming this prior. Whilst it is sensible to assume the trend that SNPs with large effect sizes are rare due to selection pressure, it is unclear whether the assumed relationship between MAF and effect size is sensible for complex diseases.

2.2 Haplotype tests

It is widely accepted that single marker tests are not the most powerful for association studies (Zollner and Pritchard, 2005). LD commonly exists between sites in small genetic regions. Consequently, phenotypic associations with neighbouring sites are not statistically independent, so by testing each marker independently information about the population history is discarded (in particular, information about the co-inheritance of markers) that if exploited can yield substantial increases in power.

Studies of the human genome reveal that variation can be characterised by the diversity of a few distinct haplotypes within "blocks" of high LD (The International HapMap Consortium, 2005). This "block" like LD structure is caused by the recombination rate variation across the genome, which is characterised by the occurrence of recombination hotspots that cause the breakdown of LD at focused locations (McVean et al., 2004). In addition, functional properties of proteins are determined by a linear sequence of amino acids corresponding to DNA variation

on haplotypes. Thus, biologically it makes sense to analyse haplotypes (Balding, 2006a). Haplotype analysis holds greater potential than single SNP approaches for two reasons:

1. haplotypes can act as a surrogate for an unobserved causal SNP (de Bakker et al., 2005; Marchini et al., 2007);

2. the causal variant may be combinations of alleles that define a haplotype, as is the case with allelic heterogeneity for example, so a haplotype can be thought of as a causal allele (Pritchard, 2001).

The simplest haplotype test is analogous to the single SNP test where each distinct haplotype, comprised of SNPs within a given window, is defined as a different allele. The data is in the form of a contingency table, similar to Table 2.1, with counts of each haplotype in the case and control samples. A simple contingency test, such as the MLRT, can be used to test the alternative model, where each distinct haplotype confers a different risk of disease, against the null model, where all haplotypes confer the same risk.

There are, however, problems with analysing haplotypes in this way. As the number of markers considered increases, the number of observed haplotypes can become very large leading to the problems of sparse data, multiple comparisons and computational intractability (Thomas, 2004). One common but unsatisfactory solution is to combine all haplotypes that are rare into a "dustbin" category.

In addition, defining the haplotypes blocks is not straightforward since they vary according to the population sampled, the sample size and the SNP density. Often there will be some recombination within a block, and conversely there can be

between-block LD that will not be exploited by a block-based analysis (Balding, 2006a).

We will now provide a broad survey of the multilocus association tests that have been designed to improve on the simple contingency test.

2.2.1 Haplotype sharing tests

Haplotypes carrying a single disease allele are descended from a common ancestor, who was the first to carry the disease allele, assuming the disease mutation has only occurred once. This ancestor transmitted the disease allele along with an entire haplotype background, surrounding the disease locus, to its offspring, and segments of this haplotype background were transmitted with the disease allele to the following generations. Recombination will have reduced the length of this haplotype background in present day haplotypes but it will be preserved within the close vicinity of the causal variant since recombination events between nearby SNPs are rare. Similarly, mutation can change this background but again, typically for short segments in the human genome, mutation events are rare. Therefore, within the vicinity of a causal variant, haplotypes carrying the same disease allele will share the same haplotype background and we would expect to see elevated levels of similarity (Thomas, 2004).

One way to reduce the number of parameters in haplotype tests is to detect an elevated level of similarity amongst case haplotypes based on some similarity measure. A natural and very general test statistic is the mean of all pairwise comparisons of

sharing between sampled haplotypes (Tzeng et al., 2003):

$$U = \frac{2}{n(n-1)} \sum_{i<j}^{n} K(H_i, H_j) \tag{2.16}$$

where $K(H_i, H_j)$ is a symmetric kernel function of some feature in the comparison of the ith and jth haplotypes. Tzeng et al. (2003) described 3 different measures (i) the"matching measure", which defines $K(H_i, H_j)$ to be 1 if the haplotypes match and 0 otherwise, (ii) the"length measure", which defines $K(H_i, H_j)$ to be the longest continuous interval of matching alleles, and (iii) the "counting measure", which defines $K(H_i, H_j)$ to be the number of alleles in common between haplotypes i and j. These similarity measures are applicable to haplotypes comprising of SNPs within a given window. They can be applied in regions of LD but the metrics do not explicitly try to account for recombination. This can be unsatisfactory since, as explained before, LD blocks are not well defined and a sub-optimal window size can lead to a loss of power.

Tzeng et al. (2003) showed that their test statistic can be expressed in terms of a quadratic form and derived a general expression for its asymptotic variance, which allows the computation of a P value. Under simulation, Tzeng et al. (2003) showed that when the disease mutation arises from a common haplotype, haplotype sharing tests are more powerful than the standard contingency test because of the overall increase in similarity within the case sample. However, this increase is reduced when the disease mutation arises from a rare haplotype, and when this occurs the contingency test is more powerful.

A feature of the method described here is that it does not assume a disease model and is therefore robust to any mis-specification. However, our understanding of the

etiologies behind complex diseases are improving and a parametric method based on a well-specified disease model should be more powerful.

2.2.2 Clustering haplotypes

Based on the belief that similar haplotypes are likely to contain the same disease mutations, we can reduce the number of parameters in a haplotype test using a model in which similar haplotypes are given similar risks.

Molitor et al. (2003b) used a Bayesian smoothing of the risk parameters, β_h, for each haplotype. Here, the β_h's are multivariate normally distributed and similar haplotypes are induced to have similar risks:

$$\beta_h \backsim N(\bar{\beta}_{-h}, \sigma_s^2 / \sum_{w \neq h} w_{hk}), \tag{2.17}$$

where β_{-h} is the vector of β_h's not including β_h and w_{hk} is the "closeness" of haplotype h to haplotype k. In order to estimate the location of a mutation, w_{hk} is taken to be some similarity measure around a putative causal locus, x. Molitor et al. (2003b) specified a Bayesian framework and the posterior distribution of x is sampled using Markov Chain Monte Carlo (MCMC) (Gilks et al., 1996).

A related approach involves clustering haplotypes into groups and detecting a signal of association in the form of nonrandom groupings of case and control haplotypes into separate clusters. For example, Molitor et al. (2003a) stochastically assigned a "centre" to each haplotype cluster, which may be thought of as the ancestral haplotype from which the members of each cluster are derived. The other haplotypes are assigned to the closest cluster according to a similarity metric, which compares

the length of the shared segment around the location of a putative mutation, x_c for cluster c. Haplotypes in the same cluster are given the same risk parameter and a MCMC fitting process finds the posterior distribution for the assignment of each haplotype. The summary statistic, η, weighs the probability of a putative causal locus according to the risk and the number of disease individuals associated with that locus:

$$\Pr(\eta = x) \propto \sum_{i:y_i=1} \Pr(\phi_i = 1|\gamma_{c_{H_i}}) I(x_{c_{H_i}} = x), \tag{2.18}$$

where $\phi_i = 1$ if haplotype H_i is comes from a case individual, c_{H_i} is the cluster that contains H_i and $\gamma_{c_{H_i}}$ is the risk of cluster c_{H_i}. Molitor et al. (2003a) computed an empirical null distribution for η by analysing the data without the phenotype information and an empirical "alternative" distribution by analysing the data complete with the phenotype information. Localisation is based on a histogram of Bayes Factors for η, computed by taking the ratio of alternative and null distributions at each possible value for η.

Unlike the haplotype sharing methods described by Tzeng et al. (2003), the methods introduced by Molitor et al. (2003a) and Molitor et al. (2003b) do not assume that all case haplotypes are, on average, more similar at a disease locus. Thus, they are a more powerful approach when there is incomplete penetrance, since not all case haplotypes will carry the disease allele, and when there are multiple causal variants, since there will be several distinct haplotype backgrounds in the case sample each harbouring a different combination of disease alleles. However, neither method can simultaneously model additive and dominant diseases. Morris (2006) introduced GENEBPM, which is a similar clustering based method, that uses a disease model that can account for both additive and dominant effects.

The three methods introduced in this section so far all use MCMC to sample from a posterior distribution and therefore can suffer from chain convergence issues. The convergence of MCMC algorithms is not one to a point, but that of the distribution of a sequence of generated values to the posterior distribution, and hence is not easy to assess; there is no guaranteed diagnostic tool to determine convergence of a MCMC algorithm in general.

An alternative clustering approach is to construct a haplotype tree that relates the case and control sample (Templeton et al., 1987, 2005), which define a nested hierarchy of clades (clusters) of haplotypes. It is hoped that any functional mutation will impact the topology of the tree and lead to whole clades of haplotypes with similar phenotypic attributes. Along these lines, Durrant et al. (2004) proposed CLADH, which uses a logistic regression model to analyse clades. Large genomic regions are broken up into sliding windows of haplotypes and within each window a cladogram is constructed using an agglomerative hierarchical clustering approach, with average linkage and based on a pairwise haplotype similarity measure similar to the counting measure. Statistical significance of the distribution of cases and controls among the clades is computed using the MLRT.

A major weakness of the CLADH, and many other cladistic approaches, is that only one tree is constructed and analysed for each locus or region of interest. Since the true genealogical tree is unknown, this assumes that the constructed genealogy is accurate and ignores the variance in its estimation.

As with the haplotype sharing methods described in the previous section, the methods described here all rely on a pairwise similarity measure. None of the methods have proposed a measure that accounts for recombination and they therefore have

to revert to a sliding window approach to restrict the analysis to within regions of LD. This causes the same problems associated with choosing an optimal window size as described earlier.

2.2.3 The Multimarker Predictor Test

According to Equation (2.9), for a fixed significance threshold, testing association at single SNPs only has power to detect SNPs that are correlated, above some r^2 threshold, with a typed SNP. The Multimarker Predictor (MMP) Test (de Bakker et al., 2005) increases the power to detect a causal variant by testing, in addition to typed SNPs, SNPs that are untyped but correlated with a combination of typed SNPs.

The MMP test is based on rules that predict the allele type at an untyped SNP according to the allelic combinations at multiple nearby SNPs. These rules are found by searching the observed correlation in a reference panel, such as HapMap (The International HapMap Consortium, 2005), for combinations of typed SNPs that are correlated, above a r^2 threshold, with an untyped SNP. The predicted alleles at untyped SNPs can then be utilised for association testing as if they were directly genotyped. de Bakker et al. (2005) showed in a simulation study that extending testing to predicted SNPs overcomes the increased multiple testing penalty and is therefore more powerful than only testing at typed SNPs.

Unless the correlation between the "tag SNPs", in a MMP rule, and the predicted SNP is exactly 1, there is an uncertainty in the prediction for the untyped allele and by ignoring this uncertainty power is lost. Further, since the reference panel is limited in size, measures need to be taken, such as limiting the size of the MMP rule

(number of tag SNPs) and the maximum genetic distance between the tag SNPs and predicted SNP, to ensure that a MMP genuinely predicts an untyped SNP and is not a result of correlation by chance.

2.2.4 Imputation

A more satisfactory solution, to the MMP, for predicting values at untyped SNPs would be to utilise information at all SNPs since they are all potentially informative. Recently, a number of approaches, based on the idea of imputing or predicting untyped SNPs, have been proposed in the literature (Browning and Browning, 2007; Scheet and Stephens, 2006; Marchini et al., 2007; Li et al., 2006) and are being widely used (The Wellcome Trust Case Control Consortium, 2007; Zeggini et al., 2007).

One of the most widely used imputation methods is IMPUTE (Marchini et al., 2007), which is based on a model of LD proposed by Li and Stephens (2003). Since we will be referring to this model extensively in this thesis we will devote a section here to give further details before continuing on the subject of imputation.

The Li and Stephens model

Li and Stephens (2003) (henceforth LS) introduced a new statistical model for patterns of LD among multiple SNPs in a population sample. Their model relates the distribution of sampled haplotypes to the underlying recombination rate, by exploiting the identity

$$\Pr(H_1, \ldots H_N|\rho) = \Pr(H_1|\rho)\Pr(H_2|H_1;\rho)\ldots Pr(H_N|H_1,\ldots,H_{N-1};\rho) \qquad (2.19)$$

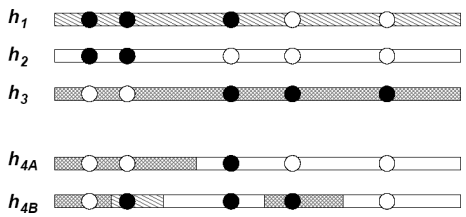


Figure 2.1: The figure (taken directly from Li and Stephens (2003)) illustrates the case $k = 3$, and shows two possible values (h_{4A} and h_{4B}) for h_4, given h_1, h_2, h_3. Each of the possible h_4s can be thought of as having been created by "copying" (imperfectly) parts of h_1, h_2 and h_3. The shading in each case shows which haplotype was "copied" at each position along the chromosome. Intuitively LS think of h_4 as having recent shared ancestry with the haplotype that it copied in each segment. They assume that the copying process is Markov along the chromosome, with jumps (i.e. changes in the shading) occurring at rate ρ/k per physical distance. Thus the more frequent jumps in h_{4B} suggest a higher value of ρ than the less frequent jumps in h_{4A}. Note that for very large values of ρ the loci become independent, as they should. Each column of circles represents a SNP locus, with black and white representing the two alleles. The imperfect nature of the copying process is exemplified at the third locus, where each of h_{4A} and h_{4B} have the black allele, although they "copied" h_2, which has the white allele (Li and Stephens, 2003).

where $H_1, \ldots, H_N$ denote the N sampled haplotypes typed at S biallelic sites, and ρ denotes the recombination parameter (which may be a vector of parameters if the recombination rate is variable along the region). This identity expresses the unknown probability distribution on the left as a product of conditional distributions on the right. LS substitute an approximation for these conditional distributions ($\hat{\pi}$) into the right hand side of Equation (2.19), to obtain an approximation to the distribution of the haplotypes conditional on ρ:

$$\Pr(H_1, \ldots, H_N|\rho) \approx \hat{\pi}(H_1|\rho)\hat{\pi}(H_2|H_1;\rho)\ldots\hat{\pi}(H_N|H_1,\ldots H_N;\rho), \tag{2.20}$$

known as the "Product of Approximate Conditionals" (PAC).

According to the LS model, the distribution of the first haplotype is independent of ρ, i.e. all 2^S possible haplotypes are equally likely, so $\pi(H_1) = (\frac{1}{2})^S$. For the conditional distribution of H_{k+1} given $H_1, \ldots H_k$, LS modelled H_{k+1} as an imperfect mosaic of

$H_1, \ldots, H_k$. That is, at each SNP, H_{k+1} is a (possibly imperfect) copy of one of $H_1, \ldots, H_k$ at that position (see Figure 2.1 taken from Li and Stephens (2003)). Let Z_j denote the "copying state" of haplotype H_{k+1} at locus j (i.e. $Z_j \in \{1, 2, \ldots, k\}$), so for example for haplotype h_{4A} in Figure 2.1, $(Z_1, Z_2, Z_3, Z_4, Z_5) = (3, 3, 2, 2, 2)$. To mimic effects of recombination, Z_j is modelled as a Markov chain on $\{1, \ldots, k\}$, with $\Pr(Z_1 = x) = 1/k$ ($x \in \{1, \ldots, k\}$), and the transition probabilities

$$\Pr(Z_{j+1} = x' | Z_j = x) = \begin{cases} \exp(-\rho_j d_j/k) + (1 - \exp(-\rho_j d_j/k))/k & \text{if } x' = x \\ (1 - \exp(-\rho_j d_j/k))/k & \text{otherwise,} \end{cases} \tag{2.21}$$

where d_j is the physical distance between markers j and $j+1$ (assumed known); and $\rho_j = 4Nc_j$, where N is the effective (diploid) population size, and c_j is the average rate of crossover per unit physical distance, per meiosis, between sites j and $j+1$ (so that $c_j d_j$ is the genetic distance between sites j and $j+1$). This transition matrix captures the idea that, if sites j and $j+1$ are a small genetic distance apart (i.e. $c_j d_j$ is small) then they are highly likely to copy the same chromosome (i.e. $Z_{j+1} = Z_j$).

To mimic the effects of mutation the copying process may be imperfect: with probability $k/(k+\tilde{\theta})$ the copy is exact, while with probability $\tilde{\theta}/(k+\tilde{\theta})$ a "mutation" will be applied to the copied haplotype. Specifically, given the copying states $Z_1, \ldots, Z_L$, the alleles $H_{k,1}, H_{k,2}, \ldots, H_{k,L}$ are independent with

$$\Pr(H_{k+1,j} = a | Z_j = i, H_1, \ldots, H_k) = \begin{cases} k/(k+\tilde{\theta}) + (1/2) \times \tilde{\theta}/(k+\tilde{\theta}) & \text{if } H_{i,j} = a \\ (1/2) \times \tilde{\theta}/(k+\tilde{\theta}) & \text{if } H_{i,j} \neq a \end{cases} \tag{2.22}$$

(The factor of (1/2) appears in both cases, so that as $\tilde{\theta} \to \infty$ both alleles become

equally likely.)

$\tilde{\theta}$ is fixed to be the inverse of the expected number of mutation events at a single site on the genealogical tree relating a random sample of N chromosomes:

$$\tilde{\theta} = \left(\sum_{m=1}^{N-1} \frac{1}{m} \right)^{-1}. \tag{2.23}$$

The expected number of mutation events at a single site on a genealogical tree relating a random sample of n chromosomes is given by $\theta \sum_{m=1}^{n-1} \frac{1}{m}$, where θ is the scaled per site mutation rate; therefore, Equation (2.23) gives a priori the expected number of mutation events at each site as 1.

Li and Stephens (2003) provided some justifications for Equation (2.21) by relating it to the coalescent model (described later), which is an established model for population genetics and is based on the biological processes that shape human evolution. The probability that the genealogies of a sample of k chromosomes and another single chromosome is not separated by recombination is $\frac{k}{\rho+k}$ under the coalescent model, assuming constant population size and no selection. Equation (2.22) can be justified in the same way: the corresponding probability that the chromosomes are not separated by a mutation is $\frac{k}{\theta+k}$.

The LS model satisfies the following properties that we would expect to hold in a population sample of haplotypes:

1. given a sample of haplotypes, the next haplotype is likely to match one that is already in the sample and this probability increases with the sample size;

2. the probability of the next haplotype being novel increases with mutation rate, but any novel haplotype is only likely to differ at a small number of sites

from an existing haplotype, rather than be completely different to all existing haplotypes; and

3. the next haplotype will be similar to existing haplotypes over contiguous regions, and the lengths of these regions decrease with increasing recombination rate.

To assess their model, Li and Stephens (2003) applied it to estimate fine-scale recombination rates in simulated and population data sets. They found that their model provides an accurate estimation of the real underlying recombination rates but produced a small systematic bias. They were able to remove this bias by modifying Equation (2.21) by replacing ρ_j with $\delta_j \rho_j$, where $\delta_j = \exp(a + b \log_{10} \rho_j)$ is a rescaling factor and the values a and b were estimated through simulation.

The LS model is now an established model for LD and has been used to estimate the haplotype phase of genotype data (Scheet and Stephens, 2006), and impute missing data (Scheet and Stephens, 2006; Marchini et al., 2007).

IMPUTE

IMPUTE (Marchini et al., 2007) extended the LS model to impute genotype data at untyped loci. It models each individual genotype as a pair of unknown haplotypes, which are imperfect mosaics of a set of reference haplotypes, which we denote as $\mathbf{H} = \{H_1, \ldots, H_P\}$. IMPUTE uses a Hidden Markov Model (HMM) (Koski, 2001) to calculate the posterior probability of the copying states of the haplotype pair of each individual. The transition states and emission probabilities of the HMM are derived from Equations (2.21) and (2.22).

Imputing the genotype, $G_{i,j}$, of the ith individual at a SNP j, which is typed in the reference panel but not in the case-control sample, is represented by its posterior probability:

$$\Pr(G_{i,j} = g|\mathbf{H}, \rho, \tilde{\theta}) = \sum_{z^{(1)}=1}^{P} \sum_{z^{(2)}=1}^{P} \Pr(G_{i,j} = g|z^{(1)}, z^{(2)}) \Pr(z^{(1)}, z^{(2)}|\mathbf{H}, \rho, \tilde{\theta}), \quad (2.24)$$

where $z^{(1)}$ and $z^{(2)}$ are the copying states of the haplotype pair of individual i at j.

IMPUTE considers information at all loci simultaneously but in a way that decreases with genetic distance from the SNP being imputed, and remains computationally tractable for large genomic regions (up to whole chromosomes). This avoids the decisions faced by other multilocus methods of how many markers to use, or how to use them, or over what physical distance to define haplotypes for haplotype analyses. The posterior probabilities of each allele at the imputed SNP allow the investigator to integrate over the uncertainty in the predicted results when testing for association.

Simulations show that IMPUTE can lead to a boost in power compared to the traditional SNP based approaches, especially for rare variants that are not strongly correlated with any typed SNPs. These variants are in general harder to tag using a few surrogate SNPs but can be better predicted by the extended haplotypes upon which they reside.

A related approach to that of IMPUTE is FASTPHASE (Scheet and Stephens, 2006). In this approach each haplotype, carried by a case or control individual, is modelled to be a member of one of a set of clusters, which are derived from the local genetic variation observed in the reference panel. Cluster membership determines the allele distribution at an untyped SNP and evolves in a Markov manner similar

to the copying states of IMPUTE for each haplotype.

2.2.5 Estimating haplotype phase

Genetic data for association tests are typically unphased and ascertaining haplotype phase can be computationally intensive and lead to a loss of power or elevated type-1 error rates (Balding, 2006a). IMPUTE and FASTPHASE have the advantage of employing a model that can handle genotype data and integrate over the uncertainty in haplotype phase to minimise any potential reduction in power.

2.3 The coalescent model

A potentially powerful approach to detecting association is to model directly the evolutionary processes that lead to genetic variations in the human population (Larribe and Lessard, 2002; Zollner and Pritchard, 2005; Minichiello and Durbin, 2006).

The genealogy of a sample of haplotypes can be described in full by its Ancestral Recombination Graph (ARG) (Griffiths and Marjoram, 1996b). For a population of chromosomes, the ARG describes how they are related to each other, through mutation, recombination and coalescence (reviewed by Hudson (1990)), back to a most recent common ancestor (MRCA). For each position on the chromosome there is a genealogical tree, called the marginal tree, embedded in the ARG, and as one moves along the chromosome, the topologies of consecutive marginal trees shift according to the impact of historical recombination events.

A pair of chromosomes carrying the same recent mutation are expected to share

a more recent common ancestor at the disease gene than a pair of chromosomes carrying different mutations. The disease mutation would have occurred on some internal branch of the marginal tree at the disease locus with the observed case chromosomes clustered underneath it. So one way to find disease associations is to scan across the marginal trees looking for significant clusterings of case haplotypes.

The ARG for a sample of chromosomes is unknown and so must be modelled in order to facilitate inference on the parameter of interest. One such model is the Wright Fisher model with recombination. Using this model, the genealogical history can be described by a stochastic process called the coalescent with recombination (Kingman, 1982; Hudson, 1983). While methods relying on this full model are prohibitively complex, progress has been made on performing computationally tractable approximations.

The approach adopted by Zollner and Pritchard (2005), LATAG, is one of the closest implementations to the full coalescent model with recombination for association mapping. Inference is performed in two stages. First, at a given locus of interest, x, they draw trees from the posterior distribution of coalescent genealogies of all the case and control haplotypes without regard to the phenotype. The unknown tree topology, node times, and ancestral sequence are treated as missing data and sampled using MCMC. Then, averaging across the genealogies, they estimate the likelihood of the phenotype data:

$$L(\mathbf{\Phi}; x, \theta, \mathbf{G}) \approx \frac{1}{M} \sum_{m=1}^{M} \Pr(\mathbf{\Phi} | x, T_x^{(m)}, \theta), \tag{2.25}$$

where $\mathbf{\Phi}$ is the phenotype information, θ denotes a vector of penetrance parameters, $T_x^{(i)}$ is the ith tree drawn from the posterior distribution conditional on $\mathbf{G}$ and

G is the genotype information. Zollner and Pritchard (2005) assume that disease mutations occur as a Poisson process on the tree and haplotypes that carry at least one mutation have a different risk of disease to haplotypes that carry no mutations. The signal of association is the nonrandom clusterings of the haplotypes, at the tips of the trees, with respect to phenotype. The significance of a given clustering is captured in Equation (2.25), by $\Pr(\mathbf{\Phi}|x, T_x^{(m)}, \theta)$, and is used to construct a MLRT or a posterior distribution for the location of the causal variant.

Assuming that the coalescent model is the correct genealogical model, Zollner and Pritchard draw the maximum amount of information that is available from the genetic data. However, even by modern computing standards it is extremely computationally intensive and cannot be applied to large scale studies. In addition, the use of MCMC can lead to the chain convergence problems, as discussed earlier.

MARGARITA (Minichiello and Durbin, 2006) is a method loosely based on the coalescent model but with a computational efficiency near that of haplotype clustering methods. The method is based on a heuristic algorithm to infer a set of "plausible" ARGs for a sample of individuals. The algorithm is initialised at time $T = 1$ (T is incremented backwards in time) by setting S_1 to be the set of contemporary, typed sequences. The algorithm proceeds by finding which coalescences, mutations and recombinations can be performed based on the local genotype information. Applying one of these operations defines an ancestral population S_{T+1}, which is constructed from S_T using a set of heuristic transition rules, and an ARG is completed when the MRCA is found.

Given the ARG, each chromosomal position has a marginal tree associated with it. Minichiello and Durbin (2006) tested each position for disease association by as-

sessing whether its marginal tree contains a branch on which a hypothetical causal mutation can suitably explain the observed disease states of the genotyped individuals. For each marginal tree, hypothetical disease-predisposing mutations are put on each of the branches in turn. These cause the case-control individuals (the leaves of the tree) to be bipartitioned into those with a mutant allele and those with the ancestral allele. A χ^2 test is used to detect any non-independence between the inferred allelic state and the disease status.

Minichiello and Durbin (2006) point out that by taking a heuristic approach MARGARITA has discarded the understood probabilistic framework that the coalescent model with recombination offers. Power is potentially lost because they do not attach a probability to an inferred ARG and treat all of them as equally likely. Further they ignore evolutionary parameters such as the local recombination rate that can lead to more accurate ARGs. MARGARITA is however computationally tractable for large scale analyses and simulations have shown it to compare favourably to the single SNP χ^2 test and CLADH.

An additional feature of LATAG and MARGARITA is that analysis of the marginal tree can yield useful information on the underlying disease model. For example the branch that creates the strongest clustering of cases beneath the marginal tree can lead to an estimation of the frequency of the causal allele and identify possible allelic heterogeneity. Minichiello and Durbin (2006) demonstrated this feature when they applied MARGARITA to an association study of Graves disease at the *CTLA4* gene. Previous analysis of the data set identified an association at the SNP CT60. MARGARITA found 2 branches on the marginal tree at CT60 that significantly segregated the case and control individuals, and lead to the identification of a second associated SNP to CT60.

2.4 Our approach

In the following chapters we will describe new approaches to association testing and will refer to the ideas presented by Durrant et al. (2004), Marchini et al. (2007), Minichiello and Durbin (2006), Molitor et al. (2003a), Tzeng et al. (2003) and Zollner and Pritchard (2005). We will utilise the same framework used by IMPUTE and thus enjoy the same benefits of increased power.

Importantly, we will tackle the problem of detecting association in the presence of allelic heterogeneity in Chapters 5 to 7, which very few methods currently address. Pritchard (2001) and Pritchard and Cox (2002) predicted that at many of the loci responsible for the bulk of genetic variance underlying diseases there will be extensive allelic heterogeneity, which will be a non-trivial impediment to association mapping. Pritchard (2001) goes on to say that "It is critical to develop statistical methods for testing association that are powerful in the presence of allelic heterogeneity". We hope that this thesis will be a step towards meeting this challenge.

Chapter 3

HAPGEN: A Method for Simulating Realistic Association Data

3.1 Introduction

We will begin by introducing a novel approach to simulating genetic data. As described in the previous chapter, the field of association testing is flooded with an array of methodologies for finding causal variants that all claim to be the most powerful under certain situations. In order to make informed choices about the method to use, it is important to evaluate and compare methods under a variety of scenarios. Ideally this would involve using real data sets of complex diseases with known etiologies but few of these are currently available. It is therefore necessary to use simulated data sets and ensure that they are as realistic as possible.

A realistic data set should satisfy two criteria. Firstly, it should contain the same correlation (LD) structure that exists in real data sets (Thomas, 2004). This is shaped by various evolutionary forces, such as recombination and demographic events (e.g. population expansion, bottlenecks and migration events). Secondly, the method should involve a suitable underlying disease model that creates realistic patterns of LD around the disease locus (or loci).

We will focus on the simulation of biallelic data. Since most association tests focus on differences in the base compositions at typed loci between individuals rather than the actual base composition, the task of simulating genetic data reduces to simulating 0's and 1's for haplotype data and 0's, 1's and 2's for genotype data. Currently there are two main approaches:

1. simulating the evolution of a sample population *forwards* in time, and
2. simulating a sample population *backwards* in time using the coalescent model.

An example of (1) is Lam et al. (2000), whose program mimics features of natural populations by direct simulation. Diploid individuals are paired at random in their generation, mated, and then produce a random number of children. The expected number of progeny per couple is determined by an exponential growth rate, and the variance in progeny number is binomial (i.e., Fisher-Wright model, Kingman (1982)). Each population is founded by 1,000 individuals and remains at that size for 50 generations. This initialisation, together with small population growth in early generations, generates random linkage disequilibrium among alleles on normal chromosomes. After 50 generations, a disease mutation is introduced on one chromosome and the population is then expanded exponentially for 100 or 200 generations, to a final size of 50,000 individuals. If the disease mutation is lost at any

generation, or if its relative frequency becomes too common, then the simulation is re-initiated. Simulating using this approach is computationally intensive, which prohibits the production of large data sets using this method, for example, Lam et al. (2000) only simulated 16 markers in a 2Mb region.

An example of (2) is used in Zollner and Pritchard (2005). They simulated an Ancestral Recombination Graph (ARG) (Griffiths and Marjoram, 1996a,b), which describes the history of mutations and recombinations giving rise to a sample of chromosomes. The type of the most recent ancestor is assigned and mutations are added to the graph according to a Poisson process. One of the mutations, chosen uniformly at random, is set to be the disease mutation. Phenotypes are determined by a penetrance function. If the frequency of case haplotypes is not within the desired range, the process can be repeated again. Case and control samples are created by sampling without replacement from the simulated population.

The main disadvantage with the aforementioned methods is that there is no scope for the simulated data to be based on real population genetic data. Instead, simulation is performed according to parameters that control for the evolutionary processes but there is, as yet, no known definitive values for them. Here, we address this problem by proposing a novel fast approach to simulating case-control haplotype (or genotype) data directly from a sample of observed population haplotypes and using estimated population fine-scale recombination rates, for example the haplotypes and estimated recombination rates from the International HapMap project (The International HapMap Consortium, 2005). By conditioning on real population haplotype data, which contain the natural signals of the evolutionary processes that have shaped the correlation structure of the human genome, we hope to reduce the need to make, potentially unrealistic, assumptions concerning the evolutionary

processes.

In this chapter we will first describe our approach to simulating case-control data sets with a single causal SNP and then a slightly modified approach to simulating instances of allelic heterogeneity. Our approach is packaged in a C++ software program, HAPGEN, and is used for simulation studies described in the subsequent chapters.

3.2 Single mutation model

Let $\mathbf{H} = \{H_1, \ldots, H_P\}$ be the set of P observed population haplotypes in our reference panel, where $H_i = \{H_{i,1}, \ldots, H_{i,L}\}$ is a single haplotype, $H_{i,j} \in \{0, 1\}$ and L is the number of SNP loci. We want to simulate M haplotypes with one of the SNPs as the disease-causing SNP. The simulation process can be summarised by the following steps.

Step 1. Specify a SNP as a disease SNP.

Step 2. Simulate the alleles at the disease SNP of the new haplotype.

Step 3. Simulate the alleles left and right flanking to the disease SNP according to the LS model (Li and Stephens, 2003).

Step 4. Return to step 2 to simulate another haplotype or terminate.

We will now describe these steps in more detail.

3.2.1 Step 1: Selecting a disease SNP

The disease SNP can be any one of the L SNPs in the panel and either the minor or the major allele can be the disease allele. In the following steps let us assume that $d \in \{1, \ldots, L\}$ is the disease SNP and 1 (as opposed 0) is the disease allele.

3.2.2 Step 2: Simulating alleles at the disease SNP

Given the panel haplotypes $H_1, \ldots, H_P$, and the simulated haplotypes $H_{P+1}, \ldots, H_k$, let H_{k+1} be the next haplotype to be simulated with phenotype $\Phi_{k+1} \in \{0 = \text{control}, 1 = \text{case}\}$. The allele at the disease SNP, $H_{k+1,d}$, is chosen stochastically according to the probabilities

$$p_0 \quad = \quad \Pr(H_{k+1,d} = 1 | \Phi_{k+1} = 0) \tag{3.1}$$

and

$$p_1 \quad = \quad \Pr(H_{k+1,d} = 1 | \Phi_{k+1} = 1). \tag{3.2}$$

The allele frequency at the disease SNP is given by

$$\Pr(H_{k+1,d} = 1) \quad = \quad p_0 \Pr(\Phi_{k+1} = 0) + p_1 \Pr(\Phi_{k+1} = 1), \tag{3.3}$$

and if we assume that the disease is rare, i.e. $\Pr(\Phi_k = 0) \approx 1$ and $\Pr(\Phi_k = 1) \approx 0$, then we have

$$p_0 \approx \Pr(H_{k+1,d} = 1). \tag{3.4}$$

Thus, we set p_0 as the population frequency of disease allele, which we estimate from the panel haplotypes. This approach effectively simulates a population sample as controls, such as the control samples used in the WTCCC study (The Wellcome Trust Case Control Consortium, 2007), so some haplotypes in our control sample can exhibit the case phenotype.

We parametrise p_1 by the relative risk parameter α,

$$\alpha = \frac{\Pr(\Phi_{k+1} = 1|H_{k+1,d} = 1)}{\Pr(\Phi_{k+1} = 1|H_{k+1,d} = 0)}. \tag{3.5}$$

By Bayes theorem and Equation (3.4) we obtain p_1

$$\begin{aligned} p_1 &= \frac{\Pr(\Phi_{k+1} = 1|H_{k+1,d} = 1)\Pr(H_{k+1,d} = 1)}{\Pr(\Phi_{k+1} = 1|H_{k+1,d} = 1)\Pr(H_{k+1,d} = 1) + \Pr(\Phi_{k+1} = 1|H_{k+1,d} = 0)\Pr(H_{k+1,d} = 0)} \\ &\approx \frac{\alpha p_0}{\alpha p_0 + (1 - p_0)}. \end{aligned} \tag{3.6}$$

If we pair up the haplotypes of the same phenotype then we have genotype data under the log additive model, i.e. heterozygote relative risk α and homozygote relative risk α^2, as demonstrated in Section 2.1.2. Alternatively, by simulating pairs of haplotypes at a time, we can allow a totally general diploid disease model through

the direct specifications of the heterozygote and homozygote relative risks:

$$\begin{aligned}
\Pr(H_{k+1,d}=0, H_{k+2,d}=0|\Phi=0) &= (1-p_0)^2, \\
\Pr(H_{k+1,d}=1, H_{k+2,d}=0|\Phi=0) &= \Pr(H_{k+1,d}=0, H_{k+2,d}=1) = p_0(1-p_0), \\
\Pr(H_{k+1,d}=1, H_{k+2,d}=1|\Phi=0) &= p_0^2, \\
\Pr(H_{k+1,d}=0, H_{k+2,d}=0|\Phi=1) &\propto (1-p_0)^2, \\
\Pr(H_{k+1,d}=1, H_{k+2,d}=0|\Phi=1) &= \Pr(H_{k+1,d}=0, H_{k+2,d}=1|\Phi=1) \propto \alpha p_0(1-p_0), \\
\Pr(H_{k+1,d}=1, H_{k+2,d}=1|\Phi=1) &\propto \beta p_0^2, \qquad (3.7)
\end{aligned}$$

where α and β are the heterozygous and homozygous relative risks, respectively. Note that we have simulated the allele frequencies in the control sample according to Hardy-Weinberg Equilibrium. Therefore, the simulated data is only appropriate for rare diseases in a large population with random mating or random population controls.

3.2.3 Step 3: Simulating alleles at other SNPs

The rest of H_{k+1}, i.e. $H_{k+1,1}, \ldots, H_{k+1,d-1}, H_{k+1,d+1}, \ldots, H_{k+1,L}$, is simulated according to the LS model. This is where we condition on the panel haplotypes and fine-scale recombination rates, and we stochastically simulate the new haplotype as imperfect mosaics of the haplotypes that are in the panel or have already been simulated.

Simulation of alleles to the left of d mirrors the simulation of alleles to the right of d. We therefore describe the simulation of $H_{k+1,d+1}, \ldots, H_{k+1,L}$ only.

Let Z_j be the copying state that denotes which haplotype H_{k+1} copies at site j (so

$Z_j = 1, 2, \ldots, k$). This state is initialised at the disease SNP as follows

$$\Pr(Z_d = z) \propto \begin{cases} \left(1 - (1/2) \times \tilde{\theta}/(k + \tilde{\theta})\right) & \text{if } H_{z,d} = H_{k+1,d} \\ \left((1/2) \times \tilde{\theta}/(k + \tilde{\theta})\right) & \text{otherwise} \end{cases} \tag{3.8}$$

where $\tilde{\theta}$ is a mutation rate parameter explained below. Equation (3.8) captures the idea that when the mutation rate is low, or a large number of haplotypes have already been observed, the initial copying state is likely to be at a haplotype that has the same allele type as H_{k+1} at d.

The hidden states of the HMM proceed according to the following transition rule

$$\Pr(Z_{j+1} = z' | Z_j = z) = \begin{cases} \exp(-\rho_j d_j/k) + (1 - \exp(-\rho_j d_j/k))/k & \text{if } z' = z \\ (1 - \exp(-\rho_j d_j/k))/k & \text{otherwise,} \end{cases} \tag{3.9}$$

which is the same as Equation (2.21); d_j is the physical distance between markers j and $j + 1$ (assumed known); and $\rho_j = \delta_j \rho'_j$, where δ_j is the correction term to adjust for biases in the recombination rates suggested by Li and Stephens (2003) (see Section 2.2.4), $\rho'_j = 4N_e c_j$, N_e is the effective (diploid) population size, and c_j is the average rate of crossover per unit physical distance, per meiosis, between sites j and $j + 1$ (so that $c_j d_j$ is the genetic distance between sites j and $j + 1$).

To mimic the effects of mutation the copying process may be imperfect: with probability $k/(k+\tilde{\theta})$ the copy is exact, while with probability $\tilde{\theta}/(k+\tilde{\theta})$ a "mutation" will be applied to the copied allele. Specifically, given the copying states $Z_{d+1}, \ldots, Z_L$,

the alleles $H_{k+1,d+1}, \ldots, H_{k+1,L}$ are independent with

$$\Pr(H_{k+1,j} = a | Z_j = z, H_1, \ldots, H_k) = \begin{cases} k/(k+\tilde{\theta}) + (1/2) \times \tilde{\theta}/(k+\tilde{\theta}) & \text{if } H_{z,j} = a \\ (1/2) \times \tilde{\theta}/(k+\tilde{\theta}) & \text{if } H_{z,j} \neq a \end{cases} \tag{3.10}$$

for $a \in \{0, 1\}$, which is the same as Equation (2.22).

The value of $\tilde{\theta}$ is the same value as that chosen by Li and Stephens (2003):

$$\tilde{\theta} = \left(\sum_{i=1}^{P-1} \frac{1}{i} \right)^{-1}, \tag{3.11}$$

where P is the number of haplotypes in the panel, which sets the expected number of mutations at each site to be 1.

3.3 Two mutation model

To simulate instances of allelic heterogeneity, where multiple mutations contribute to the risk of disease, we have modified the above algorithm to incorporate two disease SNPs.

Step 1. Specify two disease SNPs, $\mathbf{d} = (d_1, d_2) \in \big\{\{1, \ldots, L\} \times \{1, \ldots, L\}\big\}, d_1 < d_2$.

Step 2. Simulate the alleles at the disease SNPs of the new haplotype.

Step 3. Simulate the alleles between the disease SNPs.

Step 4. Simulate the alleles left flanking to d_1 and right flanking to d_2 according to the LS model.

Step 5. Return to step 2 to simulate another haplotype or terminate.

Steps 1, 4 and 5 are the same as above so we will only spend time explaining steps 2 and 3.

3.3.1 Step 2: Simulating alleles at the disease SNPs

Let 1 be the disease allele at both d_1 and d_2 and f_{ab} be the frequency of the haplotype (a, b), at SNPs d_1 and d_2, in the population, which we estimate from the panel. We set the alleles at the disease SNPs stochastically and for control haplotypes this is set according to

$$\Pr(H_{k+1,\mathbf{d}} = (a, b)|\Phi_{k+1} = 0) = f_{ab}. \tag{3.12}$$

As before, we have made the assumption that the disease is rare in the population:

$$\Pr(\Phi_{k+1} = 1) \approx 0.$$

For case haplotypes we write $\Pr(H_{k+1,\mathbf{d}}|\Phi_{k+1} = 1)$ in terms of α (the relative risk of the disease allele at d_1), β (the relative risk of the disease allele at d_2) and γ (which describes the model of interaction between the two disease SNPs so that $\gamma = 1.0$ describes a log additive model for relative risk with no epistasis):

$$\begin{aligned}
\Pr(H_{k+1,\mathbf{d}} = (0, 0)|\Phi_{k+1} = 1) &\propto f_{00} \\
\Pr(H_{k+1,\mathbf{d}} = (1, 0)|\Phi_{k+1} = 1) &\propto \alpha f_{10} \\
\Pr(H_{k+1,\mathbf{d}} = (0, 1)|\Phi_{k+1} = 1) &\propto \beta f_{01} \\
\Pr(H_{k+1,\mathbf{d}} = (1, 1)|\Phi_{k+1} = 1) &\propto \alpha\beta\gamma f_{11}.
\end{aligned} \tag{3.13}$$

3.3.2 Step 3: Simulating alleles between disease SNPs

In this step we simulate the alleles between d_1 and d_2, i.e. $H_{k+1,j}, j \in \{d_1+1, \ldots, d_2-1\}$ in a similar fashion to Step 3 in the single mutation approach. The main difference is in the assignment of copying states, which we shall describe in detail.

As before, let Z_j be the copying state that denotes which haplotype H_{k+1} copies at site j ($j \in \{d_1, d_1+1, \ldots, d_2\}$). We set the copying states between the disease SNPs to be either constant or to change once, and we make this choice stochastically according to

$$\Pr(\text{change}) = (1 - \exp(-\rho d/k)), \tag{3.14}$$

where d is the physical distance between markers d_1 and d_2, and ρ is the the rescaled average recombination rate between d_1 and d_2. The restriction of at most one change in the copying states simplifies the assignment process described later but it is an approximation since it is possible to have more than one recombination event between the two disease SNPs. This approximation is only accurate if the genetic distance between d_1 and d_2 is small and we suggest that they be selected accordingly. For example, when $k = 120$, $d = 10$kb, $N_e = 11418$ (the estimated effective population size for the CEU population in the HapMap) and $c = 1$cM/Mb (approximately the genome-wide average in humans), then the probability of a change of state under the LS model is 0.0373. By simulating recombination in this way we have also ignored the data in $H_1, \ldots, H_k$ and the mutation rate. For example, if we want to simulate the disease alleles $(H_{k+1,d_1}, H_{k+1,d_2}) = (0, 1)$ and the alleles at d_1 and d_2 of the panel and the simulated haplotypes are $(0, 0)$ and $(1, 1)$, then the only way to simulate a haplotype with a 0 allele at d_1 and a 1 allele at the d_2 is either to simulate a

recombination event between the disease SNPs or a mutation event at one of the disease SNPs. The probability of either event depends on the recombination and mutation rates, so for example the probability for a recombination event decreases as the mutation rate increases, and vice versa. A more realistic approach would be to simulate recombination conditional on all the available data, $\{H_1, \ldots, H_k\}$, $\tilde{\theta}$, and ρ but this would make simulations significantly more complex and we have left this improvement for the future.

If the copying state is constant between d_1 and d_2 then it is drawn from a multinomial distribution with probability distribution

$$\Pr(Z_j = z) \propto \begin{cases} (1-p)^2 & \text{if } H_{z,d_1} = H_{k+1,d_1} \text{ and } H_{z,d_2} = H_{k+1,d_2} \\ p(1-p) & \text{if either } H_{z,d_1} = H_{k+1,d_1} \text{ or } H_{z,d_2} = H_{k+1,d_2} \\ p^2 & \text{otherwise,} \end{cases} \tag{3.15}$$

where $p = (1/2) \times \tilde{\theta}/(k + \tilde{\theta})$ and $\tilde{\theta}$ is given by Equation (3.11).

If the copying state does change, then the copying states at d_1 and d_2 are drawn from multinomial distributions with densities

$$\Pr(Z_{d_1} = z) \propto \begin{cases} \big(1 - (1/2) \times \tilde{\theta}/(k + \tilde{\theta})\big) & \text{if } H_{z,d_1} = H_{k+1,d_1} \\ \big((1/2) \times \tilde{\theta}/(k + \tilde{\theta})\big) & \text{otherwise} \end{cases} \tag{3.16}$$

and

$$\Pr(Z_{d_2} = z) \propto \begin{cases} \big(1 - (1/2) \times \tilde{\theta}/(k + \tilde{\theta})\big) & \text{if } H_{z,d_2} = H_{k+1,d_2} \\ \big((1/2) \times \tilde{\theta}/(k + \tilde{\theta})\big) & \text{otherwise.} \end{cases} \tag{3.17}$$

A switch point, s, is drawn uniformly from integers between $d_1 + 1$ and $d_2 - 1$ so

that

$$Z_j = \begin{cases} Z_{d_1} & \text{for } j = d_1 + 1, \ldots, s \\ Z_{d_2} & \text{for } j = s + 1, \ldots, d_2 - 1. \end{cases} \tag{3.18}$$

Once the copying states $Z_j, j \in \{d_1+1, \ldots, d_2-1\}$ have been determined the alleles, $H_{k+1,j}$ for $j \in \{d_1 + 1, \ldots, d_2 - 1\}$, are determined according to Equation (3.10), as before.

3.4 Illustration of HAPGEN

In this section we illustrate the utility of HAPGEN by simulating data sets conditional on the HapMap haplotypes (The International HapMap Consortium, 2005). The HapMap project compiled a public database of common variation in the human genome, now comprising of more than 3 million SNPs for 269 DNA samples. The 269 HapMap samples came from four populations: 30 parent-offspring trios from the Yoruba people in Ibadan, Nigeria (YRI); 30 parent-offspring trios from Utah, USA, with northern and western European ancestry (from the Centre d'Etude du Polymorphisme Humain, CEU); 45 unrelated Han Chinese people from Beijing, China (CHB); and 44 unrelated Japanese people from Tokyo, Japan (JPT).

The HapMap data include ten ENCODE regions (The ENCODE Project Consortium, 2004), of 500kb, with a marker density of approximately 1 per 300bp and an allele frequency distribution that is almost complete for common alleles. We simulated data sets using the 120 CEU parental haplotypes, phased by PHASE v2.1 (Stephens and Donnelly, 2003), using the corresponding PHASE recombination map and an effective population size of $N_e = 11418$, as estimated by the HapMap for the

CEU population.

On a Intel Pentium 4 3 GHz desktop with 512 MB of RAM, simulating data for 2000 control and 2000 case individuals (8000 haplotypes) at 857 SNPs with a single disease SNP, took approximately 6 seconds; simulating data sets with two disease SNPs took approximately 13 seconds.

It is difficult to determine whether the HAPGEN simulated data sets are correct since there are no large real data sets with which to compare them. We can however illustrate that our data sets have realistic LD structures.

Figure 3.1 displays the r^2 LD structure of the ENCODE region ENr321 and recombination rates at the bottom shows how LD is broken down at recombination hot spots.

Figure 3.2 displays the LD structures of 3 HAPGEN data sets, all with a single disease SNP but different relative risks. The P values, of a single SNP association test, are displayed (in grey) with the recombination rates at the bottom and as expected the signal at the disease SNP (blue line) increases with relative risk.

Similarly, Figure 3.3 displays the LD structure of a HAPGEN data set with two disease SNPs (blue lines).

The LD structure in all 3 figures are very similar, which suggests that the LD structure in the HAPGEN data sets is consistent with the LD structure in the reference panel.

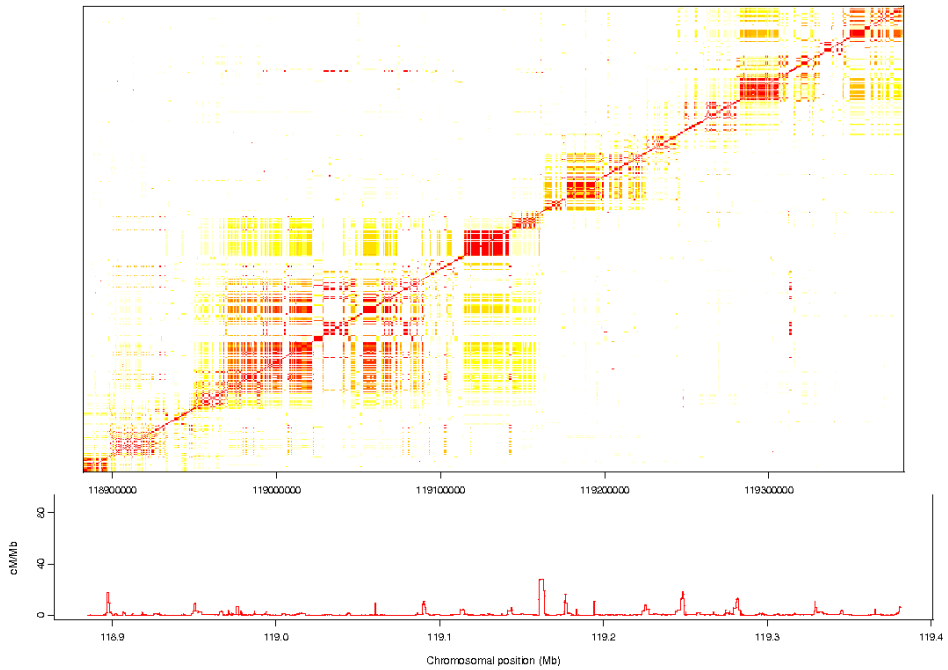


Figure 3.1: The top panel displays the r^2 structure across the ENCODE region ENr321. Dark red indicate regions of strong LD. Recombination rates are displayed in the bottom panel.

3.5 Comparing HAPGEN to coalescent simulations

The simulations in the section above illustrate graphically that HAPGEN is able to produce case-control data sets with realistic levels of LD. In order to assess in a more quantitative fashion how well HAPGEN does at creating appropriate levels of LD we compared it to the widely-used coalescent simulator called MS (Hudson, 2002). Our simulations involved the following steps.

1. We used MS to simulate 620 haplotypes over a 1Mb region. The recombination rate was set to 1cM/Mb and the mutation rate was set to 200, 400 and 600. In a SNP discovery phase, we used the segregating sites on the first 4

simulated haplotypes to ascertain a set of non-monomorphic SNPs. We found approximately 400, 700 and 1000 SNPs per simulation when the mutation rate was set to 200, 400 and 600, respectively. To proceed we thinned the data for all 620 haplotypes to the non-monomorphic SNPs only.

2. We then used the first 120 haplotypes as a panel of genetic variation. We determined the set of SNPs (denoted T) that tagged all common SNPs (minor allele frequency $\geq 5\%$), with $r^2 > 0.8$, in the panel and the set of SNPs that are common minus the tag set (denoted S).

3. We then measured the fraction of SNPs in the set S that are tagged by the set T in the remaining set of 500 haplotypes simulated by MS.

4. We then simulated a new set of 500 haplotypes using HAPGEN conditional on the panel of 120 haplotypes and calculated the fraction of common SNPs captured by the tag set.

We repeated steps 1-4 300 times and Table 3.1 shows the average fraction of tagged common variation for both MS and HAPGEN. Our results show that HAPGEN creates less variation than MS but the close agreement of these fractions shows that HAPGEN is simulating a realistic amount of genetic variation.

Mutation Rate	HAPGEN	MS
200	0.871	0.852
400	0.871	0.851
600	0.868	0.854

Table 3.1: The average fraction of tagged common variation for both MS and HAPGEN at different mutation rates.

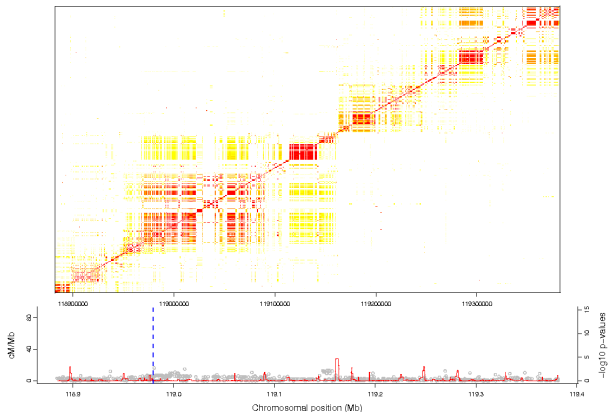
3.6 Discussion

We have described a new approach, based on the model presented by Li and Stephens (2003), that can simulate haplotype and genotype data under a general disease model. Unlike currently available methods, our method uses population haplotypes and fine-scale recombination rates for simulation. This allows the simulated data sets to naturally inherit features of population genetic data, such as their LD structure, without having to make genealogical assumptions.

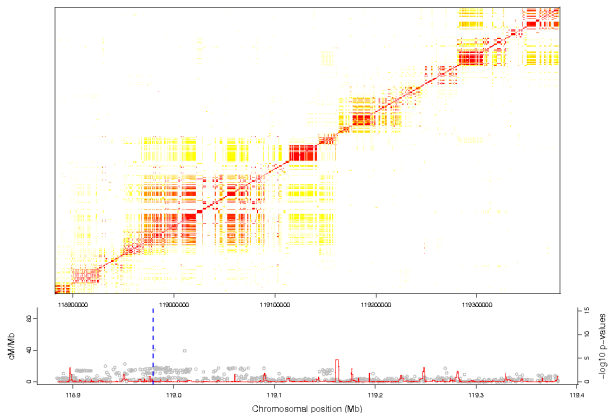
One final important feature of HAPGEN is that it is computationally fast, and thus can perform large-scale simulations.

HAPGEN has already been used in the design stage of the WTCCC study (The Wellcome Trust Case Control Consortium, 2007) and for the simulation study conducted by Marchini et al. (2007). We will be using HAPGEN for our own simulation studies in the later chapters.

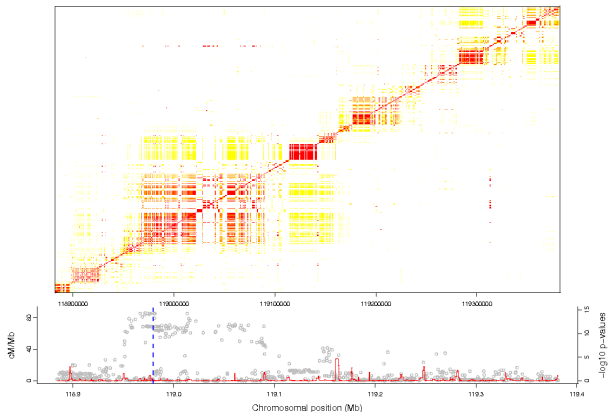
HAPGEN is currently available as a C++ programme from the Marchini group software website at `http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html`.



(a)



(b)



(c)

Figure 3.2: The top panel displays the r^2 structure of 3 HAPGEN data sets simulated from the ENCODE region ENr321. Dark red indicate regions of strong LD. The bottom panel indicates the recombination rates (red lines) and the $-\log_{10} P$ values (grey circles) of a single SNP association test performed at each SNP in the region. There is a single causal SNP at 118979782, indicated by the blue vertical line, with relative risks **a** 1.5, **b** 2.0 and **c** 2.5.

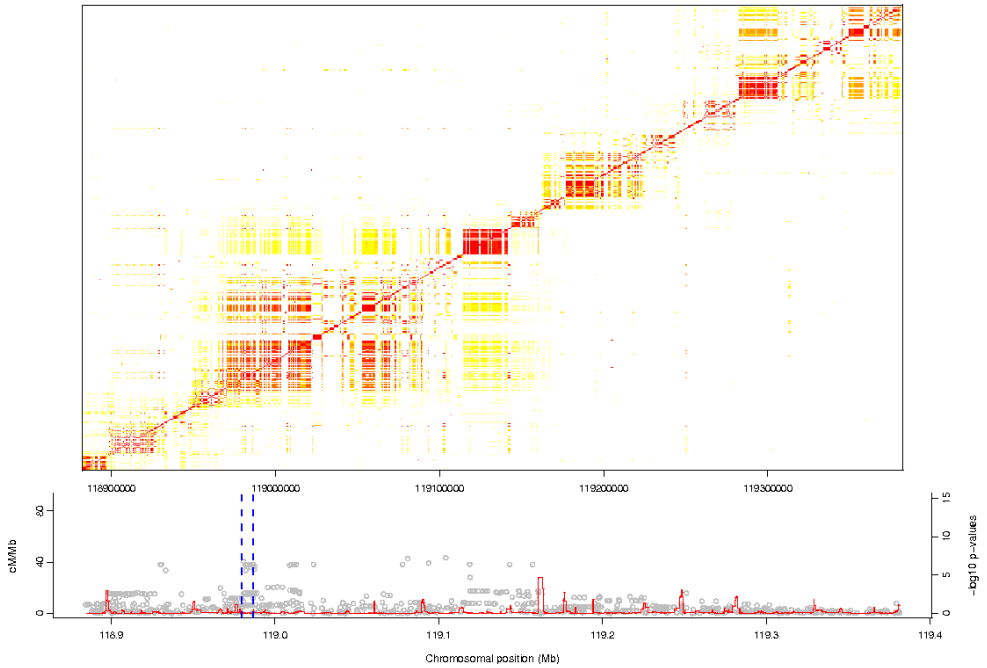


Figure 3.3: The top panel displays the r^2 structure of a HAPGEN data set simulated from the ENCODE region ENr321. Dark red indicate regions of strong LD. The bottom panel indicates the recombination rates (red lines) and the $-\log_{10} P$ values (grey circles) of a single SNP association test performed at each SNP in the region. There are two causal SNPs at 118986868 and 118979782, indicated by the blue vertical lines. The relative risks are 2.5 for 118986868 and 1.3 for 118979782 under a log additive model (i.e. $\gamma = 1.0$).

Chapter 4

Methods for Detecting Association in Haplotype Data

4.1 Introduction

In this chapter we will introduce two methods for association testing, a non-parametric method called HAPSEARCH and a parametric method called HAPCLUSTER. We will assume that our data are in the form of case-control haplotypes with known phase and no missing data, but in the next chapter we will be extending our approach to deal with genotype data and missing data. Both of our methods detect signals of association in the form of elevated similarity amongst case haplotypes. They perform a test of association at each SNP of interest and output a test statistic that summarises whether a causal variant is likely to lie in the vicinity. The output of HAPSEARCH is a non-parametric test statistic and the output of HAPCLUSTER is a Bayes Factor (Denison and Holmes, 2005).

As discussed in Section 2.2.1 haplotypes carrying a disease allele will share a common haplotype background and we expect to see elevated similarity amongst case haplotypes. Methods for detecting association based on pair-wise haplotype similarity currently exist (Durrant et al., 2004; Molitor et al., 2003a; Tzeng et al., 2003). These methods rely on a sliding window approach and similarity is measured within each window, which is unsatisfactory since the optimal window size is usually unclear. Our solution is to use the LS model (Li and Stephens, 2003), which allows us to consider information at all typed SNPs simultaneously in a computationally tractable way. The derived similarity measure conditions on the fine-scale recombination rates to vary the amount of local genetic data used to measure similarity at each position.

We will begin by describing our haplotype similarity measure and then describe the two approaches to detect a causal variant. Finally, we conduct a simulation study to evaluate our methods, and show that our methods are more powerful than the single SNP test and the Multimarker Predictor (MMP) test (de Bakker et al., 2005).

4.2 Notations

We will use the following notations throughout this chapter; let us denote

- $\mathbf{H} = \{H_1, \ldots, H_{m+n}\}$ as the haplotypes in our sample, which contains m control and n case haplotypes;
- $\{1, \ldots, L\}$ as the set of typed SNPs, so that $H_{i,j}$ is the allele of haplotype H_i at the SNP $j \in \{1, \ldots, L\}$;

- $\mathbf{\Phi} = \{\Phi_1, \ldots, \Phi_{m+n}\}$ be the phenotypes of $\mathbf{H}$, so that $\Phi_i = 1$ if H_i is a case and $\Phi_i = 0$ if it is a control, thus $m = \sum_i \mathrm{I}(\Phi_i = 0)$ and $n = \sum_i \mathrm{I}(\Phi_i = 1)$;

- $\mathbf{H}_{-i} = \{H_1, \ldots, H_{i-1}, H_{i+1}, \ldots, H_{m+n}\}$ as the set of all case and control haplotypes, except H_i;

- d_j be the physical distance between markers j and $j+1$;

- $\rho = (\rho_1, \ldots, \rho_{m+n-1})$ be a vector of rescaled fine-scale recombination rates, so that $\rho_j = \delta_j \rho'_j$, where δ_j is the correction term to adjust for biases in the recombination rates suggested by Li and Stephens (2003) (see Section 2.2.4), $\rho'_j = 4N_e c_j$, N_e is the effective (diploid) population size and c_j is the average rate of crossover per unit physical distance, per meiosis, between sites j and $j+1$;

- $\tilde{\theta}$ is the mutation parameter in the LS model, which we set as $\tilde{\theta} = (\sum_{i=1}^{m+n-1} \frac{1}{i})^{-1}$, as recommended by Li and Stephens (2003);

- $\Theta = (\rho, \tilde{\theta})$, be the vector of rescaled fine-scale recombination rates and the mutation parameter.

4.3 A haplotype similarity measure

We assume the LS model (Li and Stephens, 2003), namely that each haplotype $H_i \in \mathbf{H}$ is an imperfect mosaic of the other haplotypes, $\mathbf{H}_{-i}$. We set up a HMM for each haplotype H_i, with the hidden states $Z_{i,j} \in \{1, \ldots, i-1, i+1, \ldots, n\}$, which are the copying states of haplotype H_i at SNPs $j \in \{1, \ldots, L\}$, and the observed

variables $\mathbf{H}$. Our model conditions on the parameters in Θ, which we assume are known a priori.

If haplotypes H_{k_1} and H_{k_2} share a similar background near SNP j, then the posterior probability $\Pr(Z_{k_1,j} = k_2|\mathbf{H}, \Theta)$ will be high, and vice versa. Thus, the posterior probability distribution of $Z_{i,j}$ might be a good pair-wise measure of local haplotype similarity between H_i and the other haplotypes in $\mathbf{H}$.

4.3.1 Similarity at typed SNPs

The copying states $Z_{i,1}, \ldots, Z_{i,L}$ form a "path" from SNPs 1 to L. The transition and emission probabilities are modified forms of Equations (2.21) and (2.22) of the LS model:

$$\Pr(Z_{i,(j+1)} = k'|Z_{i,j} = k, \mathbf{H}_{-i}, \Theta) = \begin{cases} \exp(-\frac{\rho_j d_j}{(m+n-1)}) + \frac{1-\exp(-\frac{\rho_j d_j}{m+n-1})}{m+n-1} & \text{if } k' = k \\ \frac{1-\exp(-\frac{\rho_j d_j}{m+n-1})}{m+n-1} & \text{otherwise,} \end{cases} \tag{4.1}$$

and

$$\Pr(H_{i,j}|Z_{i,j} = k, \mathbf{H}_{-i}, \Theta) = \begin{cases} \frac{m+n-1}{m+n-1+\tilde{\theta}} + \frac{\tilde{\theta}}{2(m+n-1+\tilde{\theta})} & \text{if } H_{i,j} = H_{k,j} \\ \frac{\tilde{\theta}}{2(m+n-1+\tilde{\theta})} & \text{if } H_{i,j} \neq H_{k,j}, \end{cases} \tag{4.2}$$

for $i \in \{1, \ldots, m+n\}$ and $k, k' \in \{1, \ldots, i-1, i+1, \ldots, m+n\}$.

We implement the forward-backward algorithm to calculate the posterior probability distribution of $Z_{i,j}$ for each SNP j on each haplotype H_i. Let $f_i^j(k)$ and $b_i^j(k)$ be

the associated forward and backward probabilities:

$$f_i^j(k) = \Pr(H_{i,1}, \ldots, H_{i,j}, Z_{i,j} = k | \mathbf{H}_{-i}, \Theta) \qquad \text{and} \tag{4.3}$$

$$b_i^j(k) = \Pr(H_{i,(j+1)}, \ldots, H_{i,L} | Z_{i,j} = k, \mathbf{H}_{-i}, \Theta). \tag{4.4}$$

The posterior probability is given by

$$\Pr(Z_{i,j} = k | \mathbf{H}, \Theta) \propto \Pr(Z_{i,j} = k, H_i | \mathbf{H}_{-i}, \Theta) = f_i^j(k) b_i^j(k). \tag{4.5}$$

For each haplotype H_i, we calculate the values of b_i^j and f_i^j at each SNP j using dynamic programming on Equations (4.3) and (4.4).

For the forward path

$$\begin{aligned} f_i^{j+1}(k) = {} & \Pr(H_{i,(j+1)} | Z_{i,(j+1)} = k, \mathbf{H}_{-i}, \Theta) \Big(\Pr(Z_{i,(j+1)} = k | Z_{i,j} = k, \mathbf{H}_{-i}, \Theta) f_i^j(k) + \\ & \Pr(Z_{i,(j+1)} = k | Z_{i,j} \neq k, \mathbf{H}_{-i}, \Theta) \sum_{l \neq k} f_i^j(l) \Big), \end{aligned} \tag{4.6}$$

with initial values $f_i^0(k) = \frac{1}{m+n-1}$.

Similarly, for the backward path

$$b_i^{j-1}(k) = \sum_l \Pr(Z_{i,j} = k | Z_{i,(j-1)} = l, \mathbf{H}_{-i}, \Theta) \Pr(H_{i,j} | Z_{i,j} = l, \mathbf{H}_{-i}, \Theta) b_i^j(l), \tag{4.7}$$

with initial values $b_i^L(k) = 1.0$.

4.3.2 Similarity at untyped SNPs

We can extend our method to untyped SNPs. Suppose $H_{i,j'}$ is not a typed allele but is flanked by $H_{i,j}$ and $H_{i,(j+1)}$, which are typed, then the same HMM described in 4.3.1 still applies:

$$\Pr(Z_{i,j'} = k|\mathbf{H}, \Theta) \propto f_i^{j'}(k) b_i^{j'}(k). \tag{4.8}$$

We use Equations (4.6) and (4.7) to calculate $f_i^{j'}$ and $b_i^{j'}$ but assume the standard missing data approach that $\Pr(H_{i,j'}|Z_{i,j'} = k, \Theta) = 1$:

$$\begin{aligned} f_i^{j'}(k) &= \Pr(Z_{i,j'} = k|Z_{i,j} = k, \mathbf{H}_{-i}, \Theta) f_i^j(k) + \\ &\quad \Pr(Z_{i,j'} = k|Z_{i,j} \neq k, \mathbf{H}_{-i}, \Theta) \sum_{l \neq k} f_i^j(l), \end{aligned} \tag{4.9}$$

$$b_i^{j'}(k) = \sum_l \Pr(Z_{i,(j+1)} = k|Z_{i,j'} = l, \mathbf{H}_{-i}, \Theta) b_i^j(l), \tag{4.10}$$

4.3.3 The similarity matrix

Given f_i^j and b_i^j for each haplotype H_i and a SNP of interest, j, we can produce a $(m+n) \times (m+n)$ symmetric matrix, M, with entries

$$M_{k_1 k_2}^j = \begin{cases} \frac{\Pr(Z_{k_1,j}=k_2|\mathbf{H}_{-k_1},\Theta)+\Pr(Z_{k_2,j}=k_1|\mathbf{H}_{-k_2},\Theta)}{2} & \text{if } k_1 \neq k_2 \\ 0 & \text{otherwise.} \end{cases} \tag{4.11}$$

M^j can be used as a similarity matrix at SNP j so that if haplotypes H_{k_1} and H_{k_2} are closely matched in a region around SNP j then $M_{k_1 k_2}^j$ and $M_{k_2 k_1}^j$ will be large. This can be seen through Equation (4.2) – if $H_{k_1,j'} = H_{k_2,j'}$ for SNPs j' near j, then $\Pr(H_{k_1,j'}|Z_{k_1,j'} = k_2, \mathbf{H}_{-k_1}, \Theta)$ is large, which results in a large value for $\Pr(Z_{k_1,j} =$

$k_2|\mathbf{H}_{-k_1}, \Theta)$. We have taken the average of $\Pr(Z_{k_2,j} = k_1|\mathbf{H}_{-k_2}, \Theta)$ and $\Pr(Z_{k_1,j} = k_2|\mathbf{H}_{-k_1}, \Theta)$ to define the similarity between H_{k_1} and H_{k_2} at SNP j. This ensures that our definition of the similarity between H_{k_1} and H_{k_2}, and the similarity between H_{k_2} and H_{k_1}, at each SNP is well defined. Taking the average assumes that the copying states of both H_{k_1} and H_{k_2} are equally informative about the similarity between them, which is heuristically sensible, but other, perhaps equally sensible, choices include the maximum or minimum. We do not expect there to be much difference in the 3 choices, mainly because of the symmetry in Equation (4.2) between the terms i and k, which implies that $\Pr(Z_{i,j} = k|\mathbf{H}_{-i}, \Theta) \approx \Pr(Z_{k,j} = i|\mathbf{H}_{-k}, \Theta)$. Henceforth, we will refer to $M^j_{k_1k_2}$ as the "LS similarity" between haplotypes k_1 and k_2 at SNP j.

A feature of our similarity measure is that we consider data at all SNPs simultaneously, modulated by their genetic distance from the position of interest. For example, as ρ increases, $Z_{i,j}$ becomes increasingly independent of nearby copying states (see Equation (4.1)) and thus increasingly independent of nearby alleles. This is heuristically sensible since surrounding genetic data are less informative about the genealogy of the locus of interest when the recombination rate is high. Our similarity measure therefore naturally accounts for recombination and avoids the need to define a window size, which is a drawback for many of the measures proposed in the literature.

4.4 HAPSEARCH: a non-parametric association test

As discussed before, we expect elevated levels of haplotype similarity amongst the case haplotypes in the vicinity of a disease locus. Similarly, the control haplotypes will preferentially carry the non-disease allele at the disease SNP and will also show increased levels of similarity in this region. Our challenge is to find this signal in the matrix M^j to detect the presence of a disease mutation.

Our non-parametric method, HAPSEARCH, is comparable to approaches used in non-parametric linkage analysis in which genetic regions are assessed for elevated sharing between related case individuals (Thomas, 2004). The difference is that in non-parametric linkage analysis similarity is defined in terms of identity by descent, calculated from known familial relationships, whereas in case-control studies familial information is not available and so similarity is based on identity by state (at a SNP).

We investigate 5 non-parametric test statistics $(T_1, \ldots, T_5)$ to detect elevated probabilities of control haplotypes copying control haplotypes and case haplotypes copying case haplotypes.

Consider the quantities α_j and β_j defined by

$$\alpha_j^U = \frac{1}{m(m-1)} \sum_{k_1:\Phi_{k_1}=0} \sum_{k_2:\Phi_{k_2}=0} M^j_{k_1 k_2}, \tag{4.12}$$

$$\alpha_j^A = \frac{1}{n(n-1)} \sum_{k_1:\Phi_{k_1}=1} \sum_{k_2:\Phi_{k_2}=1} M^j_{k_1 k_2}, \tag{4.13}$$

$$\beta_j = \frac{1}{mn} \sum_{k_1:\Phi_{k_1}=1} \sum_{k_2:\Phi_{k_2}=0} M^j_{k_1 k_2}. \tag{4.14}$$

α_j^U and α_j^A are the mean similarities within the control and case samples, respectively, which are similar to the test statistics, also based on pair-wise haplotype comparisons, discussed by Tzeng et al. (2003). We would expect the following test statistics to increase in the vicinity of a disease locus:

1. $T_1 = \alpha_j^U$
2. $T_2 = \alpha_j^A$

T_1 and T_2 consider only the similarity within the case and control samples. A better test statistic might be to also consider the similarities between the case and control samples, which we would expect to decrease in the vicinity of a disease locus. β_j is the mean similarity between the control and case samples, which we use for T_3.

3. $T_3 = (\alpha_j^U + \alpha_j^A)/(2\beta_j)$

Now consider the quantities γ_j^U, γ_j^A and λ_j

$$\begin{aligned}
\gamma_j^U &= \frac{1}{m(m-1)} \sum_{k_1:\Phi_{k_1}=0} \sum_{\substack{k_2:\Phi_{k_2}=0 \\ k_2 \neq k_1}} \log(M^j_{k_1k_2}) \\
&= \log\Big(\prod_{k_1:\Phi_{k_1}=0} \prod_{\substack{k_2:\Phi_{k_2}=0 \\ k_2 \neq k_1}} (M^j_{k_1k_2})^{\frac{1}{m(m-1)}} \Big) \qquad (4.15) \\
\gamma_j^A &= \frac{1}{n(n-1)} \sum_{k_1:\Phi_{k_1}=1} \sum_{\substack{k_2:\Phi_{k_2}=1 \\ k_2 \neq k_1}} \log(M^j_{k_1k_2}) \\
&= \log\Big(\prod_{k_1:\Phi_{k_1}=1} \prod_{\substack{k_2:\Phi_{k_2}=1 \\ k_2 \neq k_1}} (M^j_{k_1k_2})^{\frac{1}{n(n-1)}} \Big) \qquad (4.16) \\
\lambda_j &= \frac{1}{mn} \sum_{k_1:\Phi_{k_1}=1} \sum_{k_2:\Phi_{k_2}=0} \log(M^j_{k_1k_2}) \\
&= \log\Big(\prod_{k_1:\Phi_{k_1}=1} \prod_{k_2:\Phi_{k_2}=0} (M^j_{k_1k_2})^{\frac{1}{mn}} \Big) \qquad (4.17)
\end{aligned}$$

γ_j^U, γ_j^A and λ_j are the logarithmic geometric means of the similarity within the control and case samples and between the control and case samples, and thus changes to small values in M^j will make a bigger impact to those statistics compared to α_j^U, α_j^A and β_j. Therefore we might expect γ_j^U, γ_j^A and λ_j to be more sensitive to differences in similarity if the average similarity between each pair of haplotypes is low. For the same reasons as discussed earlier we would expect the following test statistics to increase in the vicinity of a disease locus.

4. $T_4 = (\gamma_j^U + \gamma_j^A)/(2\lambda_j)$

5. $T_5 = (\gamma_j^U + \gamma_j^A) - 2\lambda_j$

The test statistics we have suggested in this section are quite arbitrary and are not derived from any formal theory. Although they are unlikely to be optimal test

statistics, they do not assume a disease model and will detect a causal variant if, as expected, there is increased similarity within the case and control samples in the region of a disease locus.

In Section 4.6 we will assess the power to detect a causal variant using each test statistic. Before that, however, we will describe a related parametric method. By specifying our model of association accurately to match what we believe about complex diseases, we hope to gain more power with this parametric approach.

4.5 HAPCLUSTER: a parametric association test

In the previous section we introduced 5 non-parametric test statistics, which are based on summarising the pair-wise haplotype similarities for the case and control samples. However, Tzeng et al. (2003) showed under simulation, that methods based on the mean pair-wise similarities in the case and control samples can be underpowered in certain scenarios. For complex diseases we would expect incomplete penetrance (control individuals carrying the deleterious allele), allelic heterogeneity (caused by the presence of multiple causal variants) and phenocopies (case individuals carrying the protective allele), all of which will attenuate the signal near the disease locus. We introduce in this section a more sophisticated parametric clustering based association test, called HAPCLUSTER. HAPCLUSTER explicitly partitions the haplotypes into clusters of varying penetrance and will detect structures of haplotype sharing within the case and control samples, which might not be detected through their mean pair-wise haplotype similarities.

HAPCLUSTER works in 3 stages:

1. at each point of the genome construct a tree relating the control and case haplotypes;

2. partition the haplotypes into groups according to the estimated tree;

3. calculate a Bayes Factor that compares a model of association, where the penetrance parameter of each haplotype is defined by its group membership and the penetrance parameters are independently distributed for each group, with a null model of no association, where all haplotypes share the same penetrance parameter.

This is similar to the cladistic method presented by Durrant et al. (2004). However, Durrant et al. (2004) only analysed a single tree at each SNP, which ignores the uncertainty of the estimated tree. We hope to gain more power by integrating over a space of tree topologies and therefore account for the uncertainty in estimating our tree. In contrast to HAPSEARCH, HAPCLUSTER is based on a parametric Bayesian framework and we assign a penetrance parameter to each haplotype, which is defined by its group membership. This is similar to the approach presented by Molitor et al. (2003a), where haplotypes are assigned risk parameters according to their group membership and MCMC is used to sample the posterior probability distribution for the group membership of each haplotype. HAPCLUSTER performs the integration process directly and therefore has the advantage of avoiding MCMC and its convergence issues.

4.5.1 Constructing haplotype trees

In Section 4.3.3 we defined a matrix of haplotype similarities, which we now use to construct a tree by hierarchical clustering. We use a deterministic agglomerative hierarchical clustering algorithm and average linkage (for more details see Hastie et al. (2001), who also give details of alternative clustering algorithms). We build the tree from the bottom to the top. Initially, each haplotype is in its own cluster and clusters with the highest similarity measure are merged iteratively until we are left with a single cluster. If G and H represent two clusters, then the similarity $s(G, H)$ between G and H is the average pair-wise LS similarities, $M^j_{ii'}$, where one member of the haplotype pair i is in G and the other i' is in H:

$$s(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} M^j_{ii'}, \tag{4.18}$$

where N_G and N_H are the respective number of haplotypes in each cluster.

If there is increased similarity within the case sample, then we would expect case haplotypes to merge with other case haplotypes before they merge with other control haplotypes. This would result in clusters of case haplotypes forming at the leaves of the tree. As a proof of principle, we simulated 20 case and 20 control HAPGEN haplotypes from a pseudo HapMap Phase 1 panel. The relative risks were set to 10 for heterozygotes and 20 for homozygotes. Figure 4.1 illustrates the tree constructed using LS similarities at the disease SNP. We observe a large cluster of 7 case haplotypes ({A2, A14, A4, A8, A12, A16, A20}), all carrying the disease allele, under a single branch of our tree, which we have coloured red. We do not observe any significant clusters of haplotypes carrying the disease allele when the tree is either constructed away from the disease SNP in Figure 4.2(a) or when the effects of the

disease allele are removed (by setting the heterozygotes and homozygotes relative risks to 1) in Figure 4.2(b). These observations suggest that partitioning haplotypes using a tree constructed from LS similarities can be potentially powerful for detecting disease mutations. The challenge is to devise a method powerful enough to detect non-random clusters of case haplotypes in the leaves of a tree for large data sets and disease alleles of moderate relative risks.

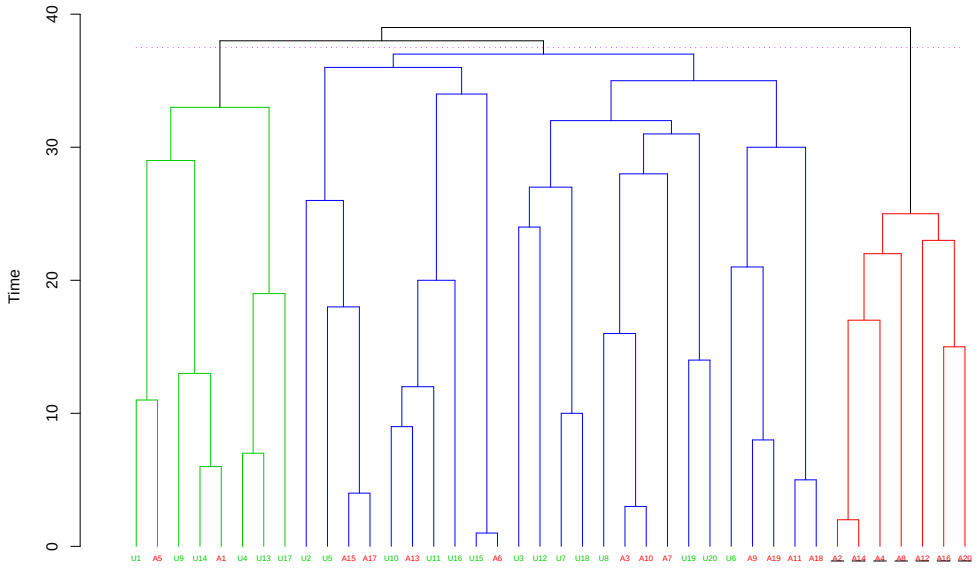
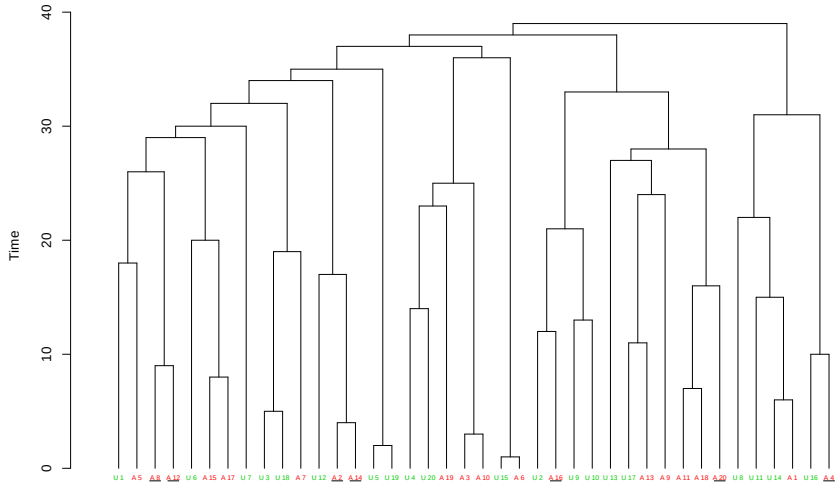
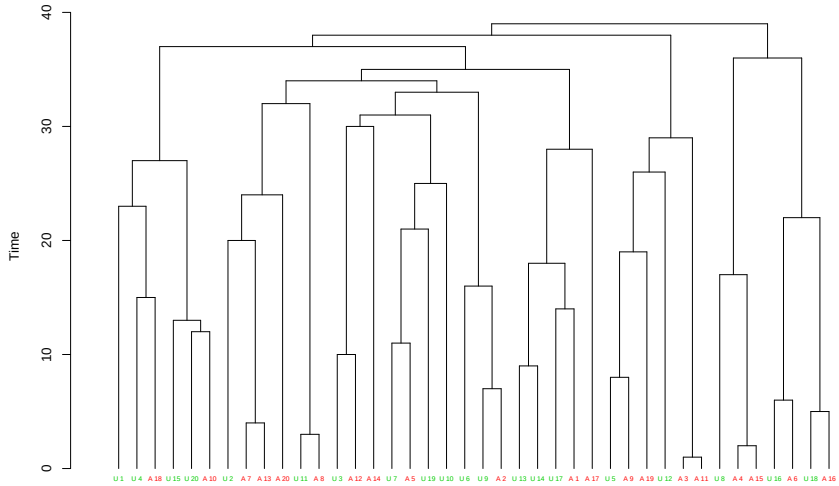


Figure 4.1: Hierarchical haplotype tree at the disease SNP; cases (Affected) are labelled A, controls (Unaffected) are labelled U and those carrying the disease allele are underlined. The colour of the branches (green, blue and red) define the partitioning of the haplotypes into 3 basis clusters, with each cluster containing the haplotypes at the leaves of the same coloured branches. These basis clusters summarise the topology of the tree below the dotted horizontal line, which we call "level 3" and is the level where there are only 3 distinct branches in the tree. Merging these clusters lead to groupings of haplotypes that represent different configurations of the tree above the dotted line.



(a)



(b)

Figure 4.2: Hierarchical haplotype trees constructed for a set of 20 case and control haplotypes. The trees are constructed at **(a)** 200Kb from a disease SNP, which has a haplotype relative risk of 10, and **(b)** at a disease SNP, which has a haplotype relative risk of 1.0, i.e. no effect. Cases (Affected) are labelled A, controls (Unaffected) are labelled U and those carrying the disease allele are underlined.

4.5.2 Constructing basis clusters and groupings

Durrant et al. (2004) noted that a haplotype tree defines successive partitions of a set of haplotypes, $\mathbf{T}[h]$, $\mathbf{T}[h-1]$, ..., $\mathbf{T}[1]$. The first partition, $\mathbf{T}[h]$ is derived from the tree at the level where there are h distinct branches. $\mathbf{T}[h]$ consists of the h clusters that correspond to the haplotypes under each distinct branch. Subsequent partitions consist of clusters derived one level higher on the tree and contain increasingly diverse clusters of haplotypes. The final partition, $\mathbf{T}[1]$, at the top of the tree, combines all haplotypes into a single cluster.

Our approach is to chose a level, which we call "level k", of the tree and use the partition $\mathbf{T}[k]$ as "basis clusters", which we denote as $\{B_1, B_2, \ldots, B_k\}$. These basis clusters approximate the tree at the point where there are only k distinct branches and haplotypes falling under the same branch are placed in the same basis cluster. By making this approximation we discard the tree topology below level k and we therefore rely on the topology above level k to contain enough information to detect significant clusterings of cases and controls in the leaves. For example, for $k = 3$ and the tree in Figure 4.1, $\mathbf{T}[3]$ defines the basis clusters $B_1 = \{$U1, A5, ... U17$\}$ (haplotypes under the green branches), $B_2 = \{$U2, U5, ... A18 $\}$ (haplotypes under the blue branches) and $B_3 = \{$A2 ,A14, ..., A20 $\}$ (haplotypes under the red branches), which summarise the tree at level 3 indicated by the purple dotted line.

We perform inference by merging the basis clusters to construct further partitions of the haplotypes, which we call groupings, and we perform a test of association on each grouping. Our test statistic is a Bayes Factor that integrates over the set of groupings that can be constructed from the basis clusters. For the example above, the groupings that we can construct from the 3 basis clusters are (B_1, B_2, B_3),

$(B_1 \cup B_2, B_3)$, $(B_1 \cup B_3, B_2)$, $(B_2 \cup B_3, B_1)$ and $(B_1 \cup B_2 \cup B_3)$ (we refer to $B_1 \cup B_2$ and B_3 as the 2 "groups" in the grouping $(B_1 \cup B_2, B_3)$). The set of groupings generated from B_1, B_2 and B_3 are precisely the set of non-unary haplotype partitions $\mathbf{T}[3]$ and $\mathbf{T}[2]$ that can be constructed from trees with haplotype clusters B_1, B_2 and B_3 as leaves. Similarly, all groupings that can be formed from k basis clusters are precisely the set of haplotype partitions defined by the set of trees with the k basis clusters as leaves. Therefore, by integrating over all the possible groupings, we allow for the uncertainty in the topologies of the estimated tree in levels above level k.

HAPCLUSTER uses a Bayesian framework and incorporates priors for haplotype grouping size (i.e. the number of groups in a grouping) and penetrance. Evidence of association comes in the form of groupings that partition the control and case haplotypes into distinct groups.

4.5.3 Analysing groupings

Let

- x be the position of interest;
- T_x be the haplotype tree that we estimate using the LS similarities at x, using the method described in Section 4.5.1;
- $\mathbf{B}_x = \{B_1, \ldots, B_k\}$ be the set of basis clusters constructed from T_x at level k, and $\mathbf{b} = \{b_1, \ldots, b_{m+n}\}$ be the membership vector so that $b_i = j$ if $H_i \in B_j$;
- D_x^s be the space of partitions (groupings) of $\mathbf{B}_x$ into s non-empty subsets (groups);

- $C_x^s \in D_x^s$ be a grouping, with $C_{x,i}^s = j$ if the H_i is in the jth group of C_x^s.

To detect the presence of a disease SNP in a specified region we assess the evidence for association at a set of positions in the region and compare it to the null model of no association. At each position, x, this can be summarised as a Bayes factor (Denison and Holmes, 2005):

$$BF(x) = \frac{\Pr(\mathbf{H}, \mathbf{\Phi}|\text{association at } x)}{\Pr(\mathbf{H}, \mathbf{\Phi}|\text{no association at } x)}. \tag{4.19}$$

Note that we model phenotype as a random variable and will therefore use a prospective likelihood.

To calculate Equation (4.19), we require $\Pr(\mathbf{H}, \mathbf{\Phi}|x)$ for each model. We can do this by using the same approach as Zollner and Pritchard (2005), but instead of integrating over marginal trees, we integrate over the groupings of haplotypes.

$$\begin{aligned}
\Pr(\mathbf{H}, \mathbf{\Phi}|x) &= \sum_{s=2}^{k} \Pr(\mathbf{H}, \mathbf{\Phi}|x, s) \\
&= \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(\mathbf{H}, \mathbf{\Phi}|x, C_x^s, s) \Pr(C_x^s|x, s) \Pr(s) \\
&= \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(\mathbf{\Phi}|x, C_x^s) \Pr(\mathbf{H}|\mathbf{\Phi}, x, C_x^s) \Pr(C_x^s|x, s) \Pr(s) \\
&= \sum_{s=2}^{k} \sum_{C_x \in D_x^s} \Pr(\mathbf{\Phi}|C_x^s) \Pr(\mathbf{H}|\mathbf{\Phi}, x, C_x^s) \Pr(C_x^s|x) \Pr(s). \qquad (4.20)
\end{aligned}$$

We make two simplifications on the final line based on our model assumptions:

- $\Pr(\mathbf{\Phi}|x, C_x) = \Pr(\mathbf{\Phi}|C_x)$. We model penetrance by group membership and haplotypes within the same group have the same penetrance parameter, thus

the phenotypes are conditionally independent of x given a haplotype grouping;

- $\Pr(C_x^s|x,s) = \Pr(C_x^s|x)$. C_x^s contains s groups by definition.

Zollner and Pritchard (2005) made the assumption that $\Pr(\mathbf{H}|\mathbf{\Phi},x,T_x) \approx \Pr(\mathbf{H}|T_x)$. They claim that this approximation is good if the disease SNP is not actually in the marker set and if mutations at different positions occur independently. We make a similar assumption $\Pr(\mathbf{H}|\mathbf{\Phi},x,C_x^s) \approx \Pr(\mathbf{H}|C_x^s,x)$ to obtain

$$\Pr(\mathbf{H},\mathbf{\Phi}|x) \approx \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(\mathbf{\Phi}|C_x^s)\Pr(\mathbf{H}|C_x^s,x)\Pr(C_x^s|x)\Pr(s). \tag{4.21}$$

Since $\Pr(\mathbf{H}|C_x^s,x)\Pr(C_x^s|x) = \Pr(C_x^s|\mathbf{H},x)\Pr(\mathbf{H}|x)$,

$$\begin{aligned} \Pr(\mathbf{H},\mathbf{\Phi}|x) &= \Pr(\mathbf{H}|x) \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(\mathbf{\Phi}|C_x^s)\Pr(C_x^s|\mathbf{H},x)\Pr(s) \\ &\approx \Pr(\mathbf{H}) \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(\mathbf{\Phi}|C_x^s)\Pr(C_x^s|\mathbf{H},x)\Pr(s). \end{aligned} \tag{4.22}$$

Here, we make the approximation that $\Pr(\mathbf{H}|x) \approx \Pr(\mathbf{H})$. This approximation is poor for high penetrance diseases since we would expect to see distinct haplotype backgrounds (for cases and controls) around the disease SNP. However, for complex diseases, where the effect of the causal variant is moderate, our approximation is more likely to be accurate.

Model of association

We use a haploid model and assign a haplotype penetrance parameter, $\gamma_{C_{x,i}^s}$, to each haplotype H_i placed in group $C_{x,i}^s$ of the grouping C_x^s; we also assume that these

penetrance parameters are independent for each group:

$$\begin{aligned}
\Pr(\mathbf{\Phi}|C_x^s) &= \int \Pr(\mathbf{\Phi}|C_x^s,\gamma)\Pr(\gamma)\mathrm{d}\gamma \\
&= \int \prod_{i=1}^{n} \gamma_{C_{x,i}^s}^{\Phi_i}(1-\gamma_{C_{x,i}^s})^{1-\Phi_i}\Pr(\gamma)\mathrm{d}\gamma \\
&= \int \prod_{i=1}^{s} \gamma_i^{n_i}(1-\gamma_i)^{m_i}\Pr(\gamma)\mathrm{d}\gamma \\
&= \prod_{i=1}^{s}\int_0^1 \gamma_i^{n_i}(1-\gamma_i)^{m_i}\Pr(\gamma_i)\mathrm{d}\gamma_i, \qquad (4.23)
\end{aligned}$$

where $(\gamma) = \{\gamma_i, \ldots, \gamma_s\}$ is a vector of penetrance parameters for the s groups, and n_i and m_i are the number of cases and controls in the ith group, which are given by

$$n_i = \sum_{j:\Phi_j=1} \mathrm{I}(C_{x,j}^s = i), \qquad (4.24)$$

$$m_i = \sum_{j:\Phi_j=0} \mathrm{I}(C_{x,j}^s = i). \qquad (4.25)$$

We use a conjugate $\beta(p_1, p_2)$ prior on γ_i, which makes Equation (4.23) easier to compute:

$$\begin{aligned}
\Pr(\mathbf{\Phi}|C_x^s) &= \prod_{i=1}^{s}\int_0^1 \gamma_i^{n_i}(1-\gamma_i)^{m_i}\frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)}\gamma^{p_1-1}(1-\gamma)^{p_2-1}\mathrm{d}\gamma_i \\
&= \prod_{i=1}^{s}\int_0^1 \frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)}\gamma^{p_1+n_i-1}(1-\gamma)^{p_2+m_i-1}\mathrm{d}\gamma_i \\
&= (\beta(p_1,p_2))^{-s}\prod_{i=1}^{s}\beta(p_1+n_i, p_2+m_i). \qquad (4.26)
\end{aligned}$$

Finally, we substitute Equation (4.26) into Equation (4.22) to get

$$\Pr(\mathbf{H}, \mathbf{\Phi}|x) = \Pr(\mathbf{H}) \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(C_x^s|\mathbf{H}, x) \Pr(s) (\beta(p_1, p_2))^{-s} \prod_{i=1}^{s} \beta(p_1 + n_i, p_2 + m_i). \tag{4.27}$$

Model of no association

Under the model of no association all haplotypes share the same penetrance parameter, which is equivalent to placing all haplotypes into a single group, C_0.

$$\begin{aligned} \Pr(\mathbf{H}, \mathbf{\Phi}|x) &= \Pr(\mathbf{H}|x) \Pr(\mathbf{\Phi}|C_0) \\ &= \Pr(\mathbf{H}|x) (\beta(p_1, p_2))^{-1} \beta(p_1 + n, p_2 + m) \\ &\approx \Pr(\mathbf{H}) (\beta(p_1, p_2))^{-1} \beta(p_1 + n, p_2 + m). \end{aligned} \tag{4.28}$$

In the absence of association at x, we would expect $\mathbf{H}$ to be distributed as they would be in the population, which allows us to make the approximation $\Pr(\mathbf{H}|x) \approx \Pr(\mathbf{H})$. This ignores any ascertainment bias, which may lead to certain haplotypes being oversampled.

4.5.4 Model choices

To calculate $BF(x)$ we must specify a choice for the number of basis clusters, k, and the probability distribution on groupings, $\Pr(C_x^s|\mathbf{H}, x)$. In this section we propose some sensible choices and assess the sensitivity of our method to each choice, by simulation, in Section 4.6.

Number of basis clusters

Choosing the number of basis clusters to use is equivalent to choosing the level to cut or summarise T_x. Cutting the tree too high, i.e. too few basis clusters, could result in a loss of power since too much information on the topology of T_x below level k is discarded. Conversely, the number of groupings grows exponentially with the number of basis clusters, which lead to computation problems for large k. In addition, the set of haplotype groupings generated from k basis clusters are a subset of the haplotype groupings generated from $k+1$ basis clusters. Therefore, the model space considered by using $k+1$ basis clusters is bigger. The Bayesian framework contains a natural penalty against over-complex models because its priors will be distributed over a larger parameter space (Denison and Holmes, 2005). So, if the data can be reasonable supported by a model involving only a few basis clusters then its posterior probability will be larger, since its priors are more concentrated around the observed data, than the posterior probability based on a model involving many basis clusters. Therefore, choosing a k too large can also lead to a loss of power.

We will compare the power of using 2-10 basis clusters in Section 4.6, which allows HAPCLUSTER to remain computationally tractable for large data sets.

Probability distribution on groupings

$\Pr(C_x^s|\mathbf{H}, x)$ specifies the probability of a haplotype grouping, C_x^s with s distinct groups, under our model. It is difficult to know how best to specify this distribution and we consider the following two possibilities:

A $\Pr(C_x^s|\mathbf{H}, x) \propto 1$, for all $C_x^s \in D_x^s$;

B $\Pr(C_x^s|\mathbf{H}, x) \propto \text{mean}(D_i)$, where $D_i = \text{mean}(M_{ab}^x : C_{x,a}^s = C_{x,b}^s = i, b_a \neq b_b)$ and $i \in \{1, \ldots, s\}$. In words this is the mean pair-wise LS similarity between haplotypes of different basis clusters placed in the same group of C_x^s. This distribution places more mass on groupings that place similar haplotypes in the same group.

We will compare how the choice of distribution affects power in our simulation study.

4.5.5 Model priors

Our models contain priors for the grouping size and the penetrance of each haplotype group. In this section we propose a set of priors and assess the sensitivity of our method to each of them in Section 4.6.

Prior for Grouping Size

$\Pr(s)$ is the prior on the size of a given grouping C_x^s in Equation (4.27). We propose the following two priors.

1 $\Pr(s) \propto 1$, for $s = 1, \ldots, k$, where k is the number of basis clusters;

2 $\Pr(s) \propto n(s)$, where $n(s)$ is the number of possible groupings of size s that can be constructed from k basis clusters, $s = 1, \ldots, k$.

Prior for haplotype penetrance

For each haplotype group we have specified a beta distribution, $\beta(p_1, p_2)$, as the prior for the penetrance parameter. We selected the beta distribution to facilitate the analytic form of the integral in Equation (4.23). However, the parameters p_1 and p_2 also provide us with flexibility to incorporate our prior knowledge on haplotype penetrance in our model and we will now briefly discuss their interpretation.

Let $\gamma = \{\gamma_1, \ldots, \gamma_s\}$ be the set of penetrance parameters for each group in the grouping $C^s_{x,i}$. The penetrance parameter, γ_j, specifies the probability that a haplotype, H_i, in group j is a case:

$$\gamma_j = \Pr(\Phi_i = 1 | C^s_{x,i} = j).$$

The range of γ_j is from 0 to 1, where values close to 0 describe a protective group (which contain haplotypes carrying a protective allele) and close to 1 describe a deleterious group. In the absence of any prior information it might be sensible to use a $\beta(1.0, 1.0)$ (uniform) prior. However, we know the following information a priori:

- the number of case (n) and control (m) haplotypes in our study and therefore the probability that a haplotype, picked at random from our sample, is a case is $\frac{n}{m+n}$, and
- the range of genetic effect sizes that we are likely to observe in our study based on previous genome-wide association studies.

The first piece of information tells us that it may be sensible to set p_1 and p_2 such that $\mathbb{E}(\gamma_j) = \mathbb{E}(\beta(p_1, p_2)) = \frac{n}{m+n}$, i.e. $\frac{p_1}{p_2} \propto \frac{n}{m}$. For the simulation study in this

chapter and in Chapter 5, we will be analysing data sets containing equal number of case and control haplotypes. Therefore, we will use priors of the form $\beta(p_1, p_1)$.

For the second piece of information if we assume that we are studying a complex disease of moderate relative risk then we should adjust the variance, $V(\beta(p_1, p_2)) = \frac{p_1 p_2}{(p_1+p_2)^2(p_1+p_2+1)}$, according to this belief. A large variance assumes a large difference in penetrance between each group and therefore high relative risk. In this chapter, we will assess the power of our method using the penetrance priors $\beta(0.6, 0.6)$, $\beta(1, 1)$ and $\beta(2.0, 2.0)$ and in Chapter 5, we will assess the power of the method introduced there using the penetrance priors $\beta(1.0, 1.0)$, $\beta(5.0, 5.0)$, $\beta(10.0, 10.0)$ and $\beta(15.0, 15.0)$. Figure 4.3 illustrates, from a simulation of 10,000,000 samples, the haplotype relative risk between each haplotype group (or penetrance group) assumed by each prior. The distributions of the relative risks assumed by $\beta(0.6, 0.6)$ and $\beta(1.0, 1.0)$ are too diverse to be plotted on the same figure; they have a mean of 20868 and 7.53, respectively, and a variance of 1.41×10^{15} and 3.5×10^6, respectively.

The disease loci detected by the WTCCC study (The Wellcome Trust Case Control Consortium, 2007) mainly have risk allele heterozygote relative risks in the range of 1-2, with most between 1 and 1.5. This corresponds to a protective allele heterozygote relative risk between 0.5 and 1, with most between 0.66 and 1. Therefore the priors, $\beta(0.6, 0.6)$ and $\beta(1.0, 1.0)$, which we have selected for our simulation study in this chapter appear to place too much mass on large effect sizes. However, it is important to assess how our method performs when we use a potentially mis-specified penetrance prior distribution and we have therefore chosen those two priors to compare to the $\beta(2.0, 2.0)$ prior, which is fairly well calibrated to the effect sizes we expect for complex diseases. In the next chapter, we will perform another set of analyses by comparing the prior distributions $\beta(5.0, 5.0)$, $\beta(10.0, 10.0)$ and

$\beta(15.0, 15.0)$, which assume more realistic effect sizes for complex diseases.

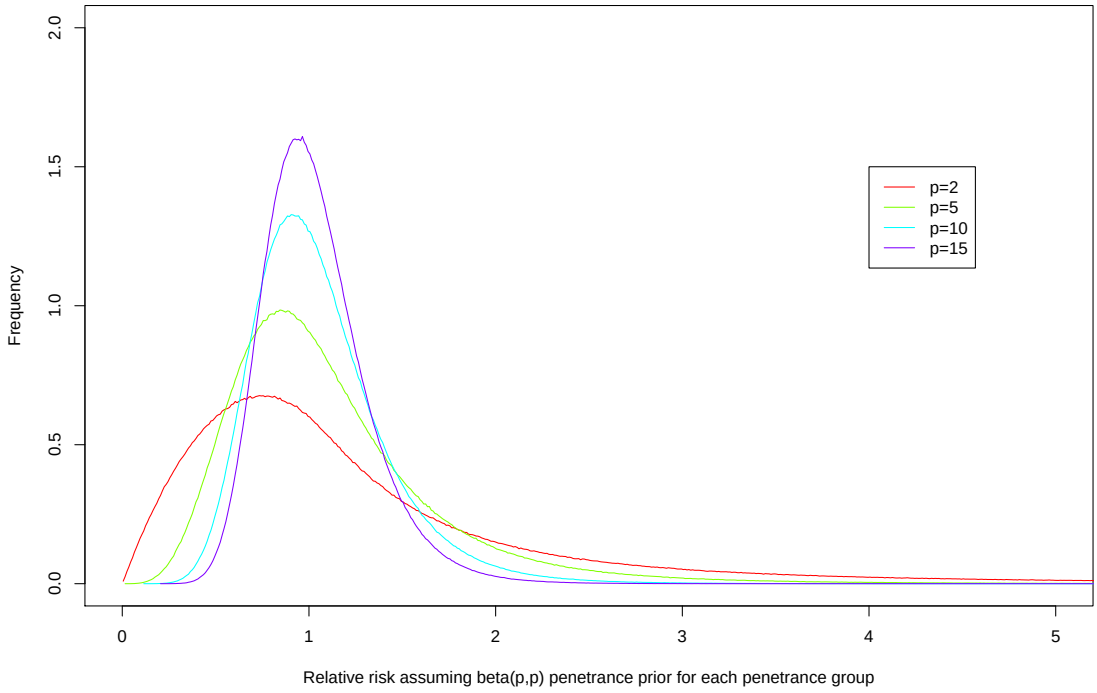


Figure 4.3: The distribution of relative risk between haplotype penetrance groups assumed by a beta penetrance prior. The distribution was obtained through simulation of 10^7 samples of $\frac{a}{b}$, where a and b are sampled from the relevant $\beta(p, p)$ distribution.

4.5.6 The test statistics

Using Equations (4.27) and (4.28), Equation (4.19) becomes

$$\begin{aligned} BF(x) &= \frac{\sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(C_x^s|\mathbf{H}, x)\Pr(s)(\beta(p_1, p_2))^{-s} \prod_{i=1}^{s} \beta(p_1 + n_i, p_2 + m_i)}{(\beta(p_1, p_2))^{-1}\beta(p_1 + n, p_2 + m)} \\ &= \frac{\sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(C_x^s|\mathbf{H}, x)\Pr(s) \prod_{i=1}^{s} \beta(p_1 + n_i, p_2 + m_i)}{(\beta(p_1, p_2))^{s-1}\beta(p_1 + n, p_2 + m)}. \end{aligned} \tag{4.29}$$

Note that $BF(x) = \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(C_x^s|\mathbf{H}, x)\Pr(s)BF(x, C_x^s)$, where $BF(x, C_x^s)$ is the Bayes Factor assessing the model of association conditional on the grouping C_x^s against the null model of no association:

$$\begin{aligned} BF(x, C_x^s) &= \frac{\Pr(\mathbf{\Phi}|C_x^s)}{\Pr(\mathbf{\Phi}|C_0)} \\ &= \frac{\prod_{i=1}^{s} \beta(p_1 + n_i, p_2 + m_i)}{(\beta(p_1, p_2))^{s-1}\beta(p_1 + n, p_2 + m)}. \end{aligned} \tag{4.30}$$

To detect a disease locus we calculate $BF(x)$ at a set of SNPs, $\{x_1, x_2, \ldots, x_N\}$, within the region, R, of interest. As a summary test statistic we can use the mode $BF(x)$ or the Bayes Factor for the region, $BF(R)$:

$$\begin{aligned} BF(R) &= \frac{\int_{x \in R} \Pr(\mathbf{H}, \mathbf{\Phi}|\text{association at } x)\Pr(x)\mathrm{d}x}{\int_{x \in R} \Pr(\mathbf{H}, \mathbf{\Phi}|\text{no association at } x)\Pr(x)\mathrm{d}x} \\ &= \frac{\sum_{i=1}^{N} \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(C_x^s|\mathbf{H}, x_i)\Pr(s)(\beta(p_1, p_2))^{-s} \prod_{i=1}^{s} \beta(p_1 + n_i, p_2 + m_i)\Pr(x_i)}{\sum_{i=1}^{N} (\beta(p_1, p_2))^{-1}\beta(p_1 + n, p_2 + m)\Pr(x_i)} \\ &= \frac{\sum_{i=1}^{N} \sum_{s=2}^{k} \sum_{C_x^s \in D_x^s} \Pr(C_x^s|\mathbf{H}, x_i)\Pr(s)(\beta(p_1, p_2))^{-s} \prod_{i=1}^{s} \beta(p_1 + n_i, p_2 + m_i)\Pr(x_i)}{\beta(p_1, p_2))^{-1}\beta(p_1 + n, p_2 + m)} \\ &= \sum_{i=1}^{N} \Pr(x_i)BF(x_i). \end{aligned} \tag{4.31}$$

Given no prior knowledge and assuming there is only one causal locus, we use the uniform prior on its location, i.e. $\Pr(x) = 1/N$, so $BF(R)$ is the mean $BF(x)$ in R:

$$BF(R) = \frac{1}{N} \sum_{i=1}^{N} BF(x_i). \tag{4.32}$$

4.6 Simulation study

4.6.1 Simulation of case-control data

To systematically evaluate the relative performance of our methods we carried out a simulation study.

We used the same data sets as Marchini et al. (2007) did in their simulation study to evaluate IMPUTE. Each data set contains 200 control and 200 case haplotypes generated by HAPGEN, conditional on a panel of 120 CEU parental haplotypes in the ENCODE regions (The International HapMap Consortium, 2005; The ENCODE Project Consortium, 2004), which were phased by PHASE v2.1 (Stephens and Donnelly, 2003), and using the corresponding PHASE recombination map.

For every SNP in the ENCODE marker set a "disease" data set was simulated with the minor allele at that SNP as the causal allele. The effect size was such that if this SNP is directly tested, then the χ^2 test would achieve a power of 95% at the 1% significance level. This requires the minor allele frequency (MAF) to be inversely correlated with risk, so rare alleles were assigned a stronger effect than common alleles. This is the same scheme used by de Bakker et al. (2005). For our simulation study we took data sets generated from 3 regions: ENr123, ENr213 and ENr232, which were chosen to reflect the levels of recombination that is found in the human genome. There are 2607 disease data sets in total – 1154 from ENr123, 660 from ENr232 and 793 from ENr213. For every disease data set a null data set was also generated where the "causal" allele had no effect (relative risks set to 1.0). Data for genome-wide association analyses are collected from genotyping chips, which have a lower marker density than the ENCODE marker sets. In order

to realistically simulate genome-wide association studies we thinned the haplotype data to the markers on the Affymetrix 500K chip (approximately 100 markers per data set).

Our approach is to scan sequentially across the region of each data set, considering a set of possible positions for the disease SNP. At each position we performed a test for association and calculated a test statistic. For our HMM, we used the PHASE recombination map and an effective population size of $N_e = 11418$, as estimated by the HapMap for the CEU panel.

We measure performance in terms of the power to detect the presence of a causal SNP in each "disease" data set. We calculate power, in the same way as de Bakker et al. (2005) and Marchini et al. (2007), by deriving a summary test statistic from the set of test statistics for each data set. We consider two possible summary test statistics for a data set:

- the mode test statistic, across the region of the data set, and
- the mean test statistic, across the region of the data set.

The sample of summary test statistics from the null data sets provides an empirical null distribution and we obtain power by comparing it to the empirical alternative distribution derived from the summary test statistics of the disease data sets. Given a significance level of α, the significance threshold for declaring association is the minimum null summary test statistic exceeding the top $(1 - \alpha)\%$ of null summary test statistics and the power is the frequency of alternative summary test statistics that exceeds this significance threshold. This approach, also used by Servin and Stephens (2007) and Marchini et al. (2007), allows us to compare Bayesian methods,

which have Bayes Factors as a test statistic, with Frequentist methods, which have P values as a test statistic.

4.6.2 Assessing HAPSEARCH

We calculated HAPSEARCH test statistics (T1-T5) at positions every 5Kb apart. We find that the empirical distribution of the summary test statistics under the null model (derived from the null data sets) and the alternative model (derived from the alternative data sets) are noticeably different. These two distributions are shown in Figure A.1 in Appendix A. The test statistics are larger under the alternative, which suggests that HAPSEARCH has detected the presence of a causal SNP for a number of disease data sets.

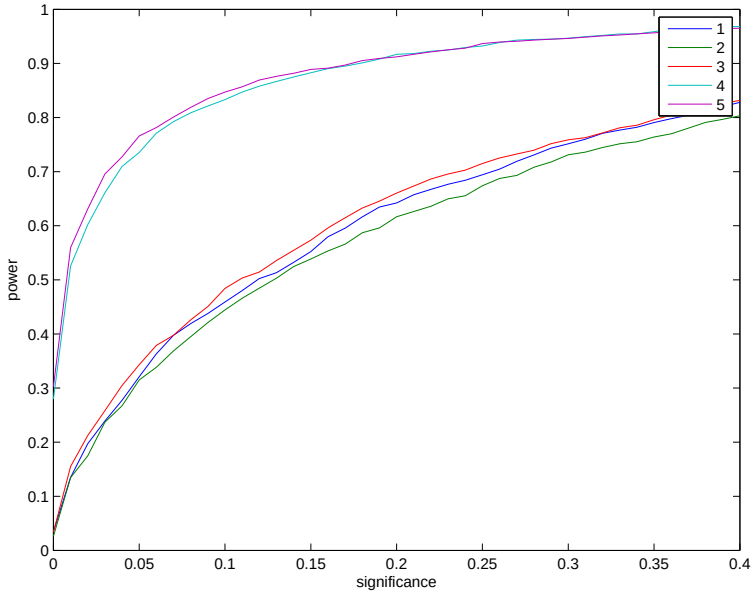
Figure 4.4 displays the Receiver Operating Characteristic (ROC) curve, which shows the power of each summary test statistic averaged over the 3 regions that the data was generated from. Test statistics T_4 and T_5, based on the geometric means of the similarity matrix, have similar power and are uniformly more powerful than the other test statistics. They offer over 40% more power at the 1% and 5% significance levels than T_1, T_2 and T_3. Tables 4.1 and 4.2 indicate that there is little difference in power between using the mean and the mode as summary test statistics.

Test statistic	Significance (%)		
	1	5	10
T_1	0.14	0.32	0.46
T_2	0.13	0.32	0.44
T_3	0.16	0.34	0.48
T_4	0.53	0.74	0.83
T_5	0.56	0.77	0.85

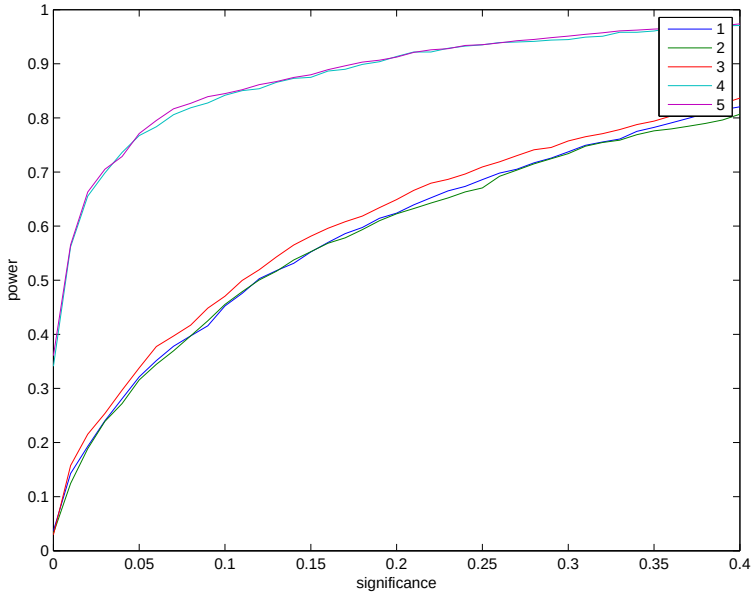
Table 4.1: The power of HAPSEARCH mode summary test statistics. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

Test statistic	Significance (%)		
	1	5	10
T_1	0.14	0.32	0.45
T_2	0.12	0.32	0.46
T_3	0.16	0.34	0.47
T_4	0.56	0.77	0.84
T_5	0.57	0.77	0.85

Table 4.2: The power of HAPSEARCH mean summary test statistics. The results are averaged over the 3 ENCODE regions that the data sets were generated from.



(a)



(b)

Figure 4.4: The power of HAPSEARCH test statistics, using **a** the mode and **b** the mean test statistic as the summary test statistic, averaged over the 3 ENCODE regions – ENr123, ENr213 and ENr232 – that the data sets were generated from. The power of T_1 is represented by blue, T_2 by green, T_3 by red, T_4 by cyan and T_5 by magenta.

4.6.3 Assessing HAPCLUSTER

We ran HAPCLUSTER at positions every 5Kb apart and used 2-10 basis clusters. Figure A.2 in Appendix A provides an example of the HAPCLUSTER Bayes Factors across the region of a disease data set and its corresponding null data set, generated from the ENCODE region ENr123. We observe an increase in the Bayes Factors near the disease SNP for the disease data set but not the null data set. This suggests that HAPCLUSTER has detected the presence of a disease SNP in the disease data set.

We also observe that the Bayes Factors for the null data set decrease as the number of basis clusters increases. The reason for this is that, as explained earlier, the complexity of the models of association increases with the number of basis clusters and therefore HAPCLUSTER incurs a greater penalty under the Bayesian framework, if the data can be supported by a simpler model.

Comparing summary test statistics

We ran HAPCLUSTER with model prior B1 (option B for $\Pr(C_x|s, H, x)$ and option 1 for $\Pr(s)$), penetrance prior $\beta(2.0, 2.0)$ and 2-10 basis clusters. Figure A.3 in Appendix A shows the empirical distributions of the summary test statistics for the 1154 disease and null data sets generated from the ENCODE region ENr123. There is a clear difference between the alternative and null distributions, with the Bayes Factors larger for the disease data sets, which is encouraging for detecting association. Again, for reasons stated above, we observe that the Bayes Factors for the null data sets decreases as the number of basis clusters increases.

Table 4.3 shows that we can gain more power, $1-2\%$ at the 5% significance level, by averaging the Bayes Factors across the region of each data set (i.e. using the regional Bayes Factor, $BF(R)$). This result is also supported by Figure 4.5, which displays the ROC curve for the two summary test statistics. Therefore, combining the information at each SNP is more powerful, which has also been reported by Servin and Stephens (2007).

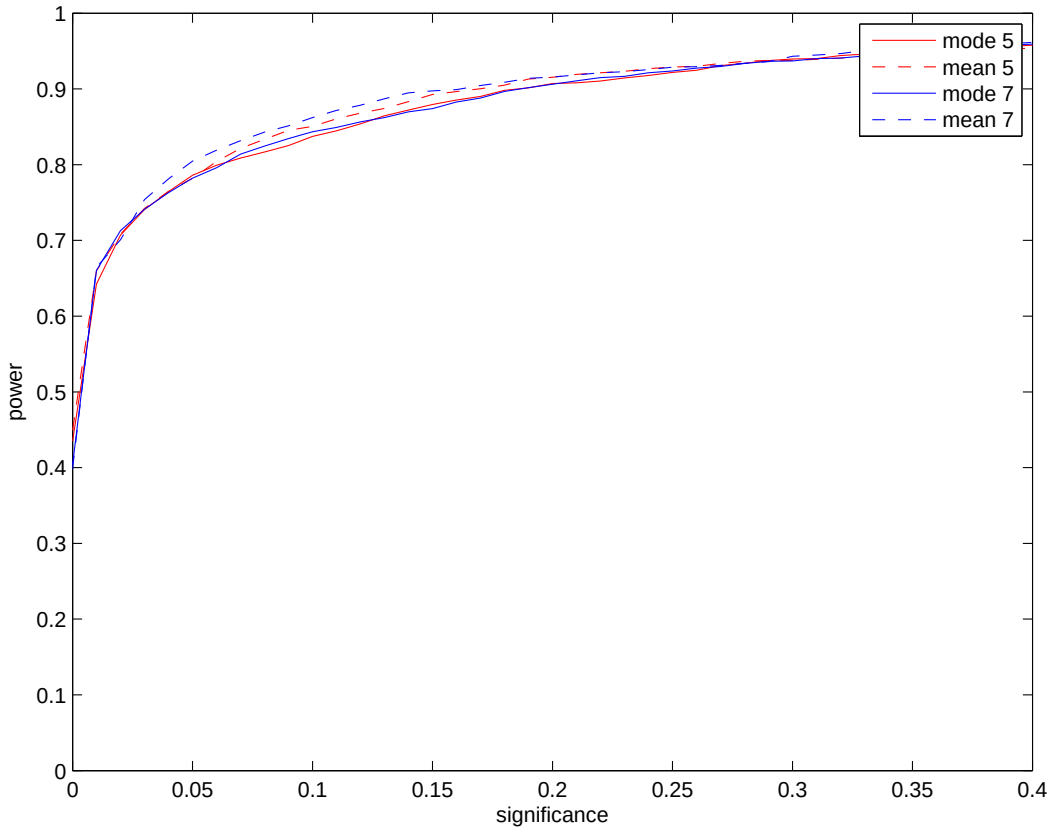


Figure 4.5: The power of HAPCLUSTER using the mode (solid lines) and the mean (dashed lines) Bayes Factor as the summary test statistic. The graph displays the ROC for 5 basis clusters (red lines) and 7 basis clusters (blue lines). The results are averaged over the 3 ENCODE regions that the data sets were generated from.

Number of Basis Clusters	Significance(%)					
	1		5		10	
	mode	mean	mode	mean	mode	mean
2	0.56	0.58	0.72	0.70	0.79	0.79
3	0.63	0.64	0.75	0.75	0.83	0.83
4	0.64	0.65	0.77	0.77	0.84	0.84
5	0.64	0.66	0.79	0.78	0.84	0.85
6	0.66	0.66	0.79	0.80	0.84	0.86
7	0.66	0.66	0.78	0.80	0.84	0.86
8	0.63	0.65	0.78	0.80	0.85	0.86
9	0.62	0.63	0.78	0.80	0.85	0.86
10	0.62	0.63	0.78	0.80	0.85	0.86

Table 4.3: The power of HAPCLUSTER using the mode and the mean Bayes Factor as the summary test statistic. We have provided the power at the 1%, 5% and 10% significance levels. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

Sensitivity to model priors

We compare power between using 2-10 basis clusters and between using the 4 possible choices for model distributions (distributions A or B for the probability distribution on groupings, and distributions 1 or 2 for the prior distribution on grouping size, as defined earlier) – A1, A2, B1 and B2. Our results are summarised in Table 4.4 and Figure 4.6.

It appears that using less than 4 basis clusters is underpowered. The disease alleles in our study have incomplete penetrance and therefore we can expect some control haplotypes to carry the disease allele and some case haplotypes to carry the protective allele. These factors will reduce the power of our constructed tree to bipartition the case and control haplotypes under its top two branches. This was not possible even in our example in Figure 4.1, where the relative risks are extremely high: 10 for heterozygotes and 20 for homozygotes. The increase in power from using 2 basis

clusters to 4 suggests that cases are clustered under internal branches, as observed in Figure 4.1, and by using more basis clusters, i.e. cutting the tree at a lower level, we have more power to detect these clusters.

Using more than 4 basis clusters does not seem to offer any significant extra power (see also Figure 4.6), which would imply that, under our simulations at least, the signal of association can be captured using only 4 basis clusters.

Comparisons between the different combinations of the probability distribution on groupings and the prior on grouping size suggest that HAPCLUSTER is robust to their choices. This is perhaps an indication that the strength of our data can overcome any mis-specifications to their choices.

Number of Basis Clusters	Significance(%)							
	1				5			
	A1	A2	B1	B2	A1	A2	B1	B2
2	0.55	0.55	0.55	0.55	0.69	0.69	0.69	0.69
3	0.61	0.61	0.61	0.60	0.73	0.74	0.73	0.73
4	0.62	0.62	0.62	0.63	0.76	0.76	0.75	0.75
5	0.62	0.63	0.62	0.62	0.77	0.77	0.77	0.77
6	0.61	0.63	0.63	0.64	0.78	0.78	0.78	0.78
7	0.62	0.63	0.62	0.63	0.78	0.78	0.78	0.78
8	0.63	0.63	0.63	0.63	0.78	0.78	0.78	0.79
9	0.61	0.62	0.63	0.62	0.77	0.78	0.78	0.78
10	0.61	0.60	0.61	0.62	0.78	0.78	0.78	0.78

Table 4.4: Comparison of power using the 4 possible choices for model distributions for HAPCLUSTER. The mean Bayes Factor is used as the summary test statistic and we have used a $\beta(1.0, 1.0)$ penetrance prior. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

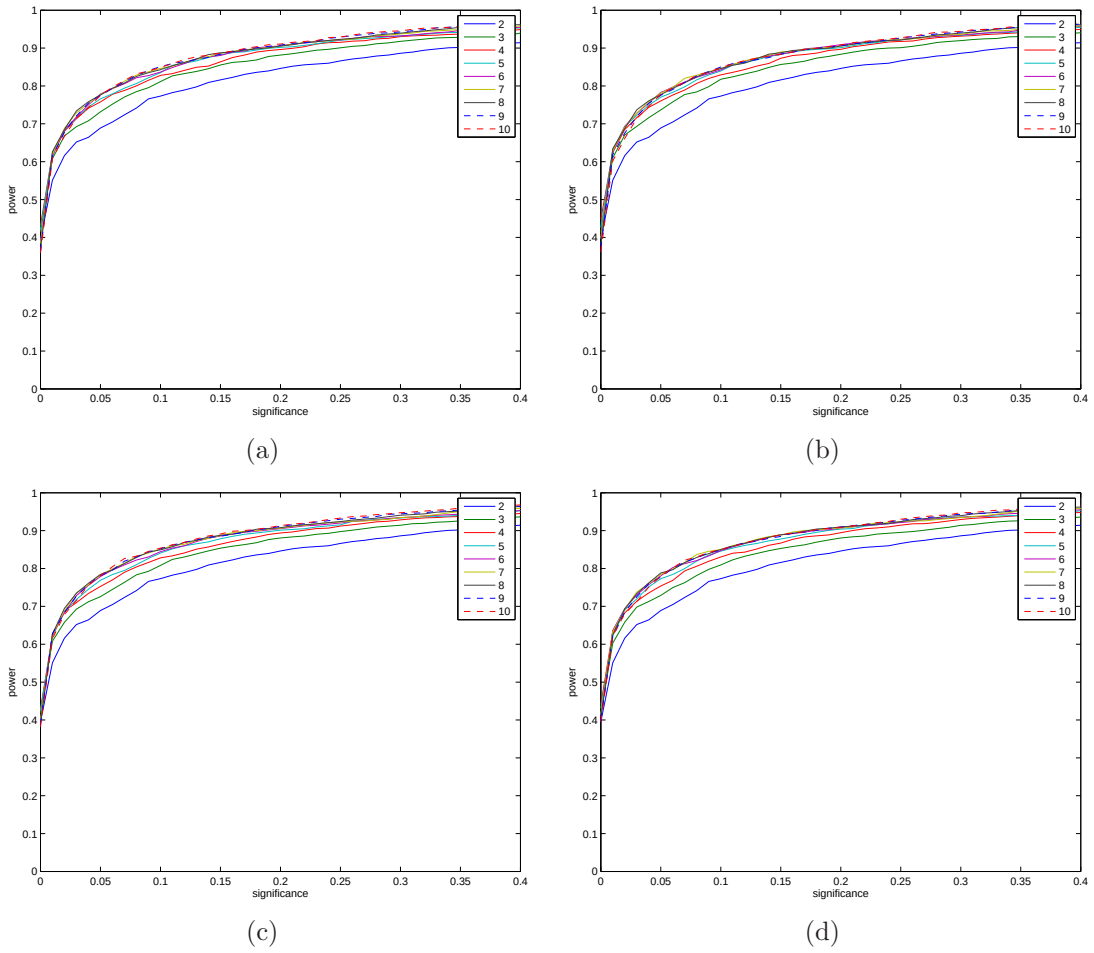


Figure 4.6: The power of HAPCLUSTER using model distributions **a** A1, **b** A2, **c** B1, and **d** B2 and the mean Bayes Factor as the summary test statistic.

Sensitivity to penetrance priors

We set HAPCLUSTER to use model distributions B1 but different haplotype penetrance priors – $\beta(0.6, 0.6)$, $\beta(1.0, 1.0)$ and $\beta(2.0, 2.0)$. Table 4.5 shows that HAPCLUSTER using $\beta(2.0, 2.0)$ is uniformly the most powerful and $2-4\%$ more powerful than the other priors at the 5% significance level. This is illustrated in Figure 4.7, which compares the power of using each penetrance prior with 7 basis clusters, and Figure A.4 in Appendix A, which displays the power of using each penetrance prior

with 2-10 basis clusters.

The causal variant in our data sets have a relatively small effect, so priors that place large probability mass on large penetrance values are inappropriate as the null model would be favoured over the alternative, because the data would be somewhat consistent with zero effects and completely inconsistent with large effects. Therefore, $\beta(2.0, 2.0)$, which places the most mass on small effects (see Section 4.5.5), provides the most power.

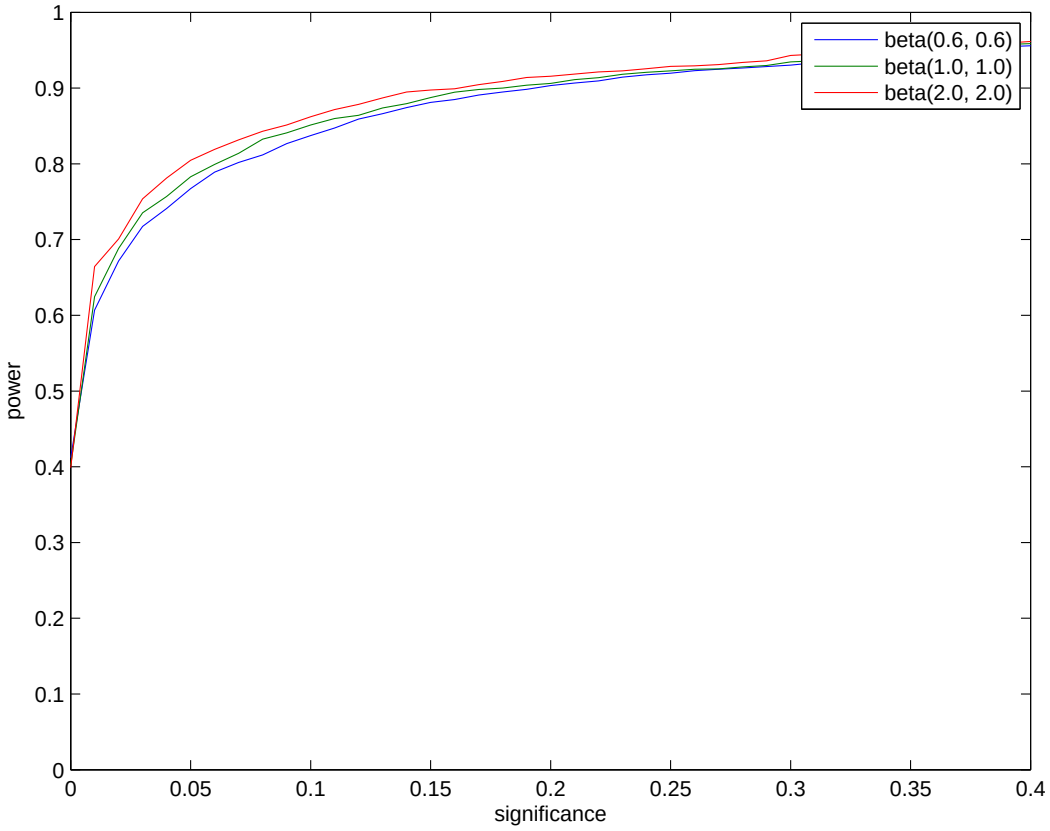


Figure 4.7: The power of HAPCLUSTER with 7 basis clusters and the penetrance priors $\beta(0.6, 0.6)$, $\beta(1.0, 1.0)$ and $\beta(2.0, 2.0)$. The mean Bayes Factor is used as the summary test statistic. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

Number of Basis Clusters	Significance(%)					
	1			5		
	β $(0.6, 0.6)$	β $(1.0, 1.0)$	β $(2.0, 2.0)$	β $(0.6, 0.6)$	β $(1.0, 1.0)$	β $(2.0, 2.0)$
2	0.53	0.55	0.58	0.67	0.69	0.70
3	0.58	0.61	0.64	0.71	0.73	0.75
4	0.61	0.62	0.66	0.74	0.75	0.77
5	0.59	0.62	0.66	0.75	0.77	0.78
6	0.60	0.63	0.66	0.76	0.78	0.80
7	0.61	0.62	0.66	0.77	0.78	0.80
8	0.61	0.63	0.65	0.77	0.78	0.80
9	0.60	0.63	0.63	0.77	0.78	0.80
10	0.61	0.61	0.63	0.77	0.78	0.80

Table 4.5: Comparison of power using different penetrance priors for HAPCLUSTER. The mean Bayes Factor is used as the summary test statistic. We have used B1 for the grouping distributions. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

4.6.4 Comparing to SNP based methods

In this section we compare the power of HAPSEARCH and HAPCLUSTER with the single SNP χ^2 test and the MMP test (de Bakker et al., 2006).

For HAPSEARCH we use test statistic T_5 and for HAPCLUSTER we use 7 basis clusters, model distributions B1 and a $\beta(2.0, 2.0)$ penetrance prior.

We tested for association using the Allelic Test (described in section 2.1.1) at every SNP on the Affymetrix 500K chip within each region and also at SNPs predicted by a MMP. We define the test statistics for the "Single SNP" analyses as the χ^2 statistics at all Affymetrix SNPs and for the MMP analyses as the χ^2 statistics at all Affymetrix and predicted SNPs.

To perform the MMP analysis, we took the set of MMPs used by Marchini et al. (2007) in their simulation study. They created a pseudo HapMap Phase 2 panel (Frazer et al., 2007) by thinning the ENCODE data sets so that each panel has a HapMap Phase 2 allele frequency distribution and marker density, with the added constraint that each panel contains the Affymetrix 500K markers that lie in the region. Multimarker predictors were designed using the Affymetrix marker set in each region based on LD patterns within the pseudo HapMap panels. We used all MMPs of size 2 that predict SNPs in the panel with a sample $r^2 > 0.8$, with the constraint that all pairs of predictor and predicted SNPs in each MMP rule lie within 200Kb of each other.

We find that the relative performances of each method depends greatly on the type of summary test statistic used. The Single SNP and MMP methods are more powerful using the mode as the summary test statistic whereas HAPSEARCH and

HAPCLUSTER are more powerful using the mean as the summary test statistic. We therefore compare the power of each method using their most powerful summary test statistic, which is displayed in Figure 4.8. We observe that HAPCLUSTER is the most powerful at significance levels below 20%. The MMP test is HAPCLUSTER's closest rival and is 7% less powerful at the 1% significance level and 3% less powerful at the 5% significance level (Table 4.6). HAPSEARCH is the most powerful method above the 20% significance level but such levels are less important in association testing.

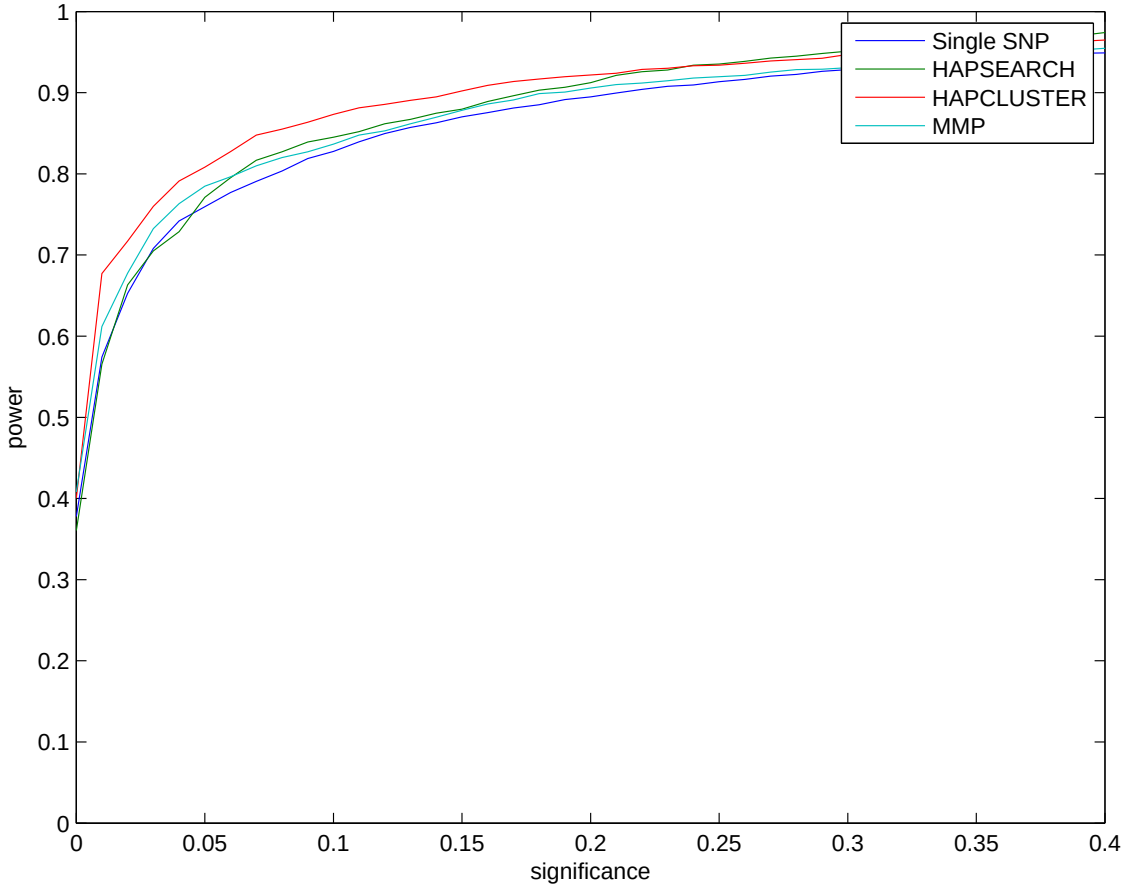


Figure 4.8: A comparison of power between HAPCLUSTER, HAPSEARCH, MMP and Single SNP methods using the most powerful summary test statistic, averaged over the 3 ENCODE regions that the data sets were generated from.

Method	Significance (%)		
	1	5	10
HAPSEARCH	0.57	0.77	0.85
HAPCLUSTER	0.68	0.81	0.87
Single SNP	0.57	0.76	0.83
MMP	0.61	0.78	0.84

Table 4.6: The power of each method using their most powerful summary test statistic, at significance levels 1%, 5% and 10%. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

4.6.5 Stratifying by type of disease SNP

To further analyse the performance of our methods, we stratify the results in Figure 4.8 according to the type of disease allele of each data set. We define 4 categories according to whether the disease allele is rare ($< 5\%$) and whether it is tagged (with $r^2 > 0.8$) by a typed SNP or by a MMP:

a rare not tagged,

b rare tagged,

c common not tagged, and

d common tagged.

Our results are displayed in Figure 4.9 and Table 4.7.

When the causal variant is tagged, the Single SNP and MMP methods are more powerful than HAPSEARCH and HAPCLUSTER, but the difference is small for HAPCLUSTER (less than 3% at the 5% significance level). When the causal variant is untagged, however, our methods provide a clear boost in power (over 10% at the 5% significance level, in the case of rare and untagged variants).

We do not observe much correlation between power and disease allele frequency, probably because the effect size of each variant was set to be inversely proportional to the allele frequency.

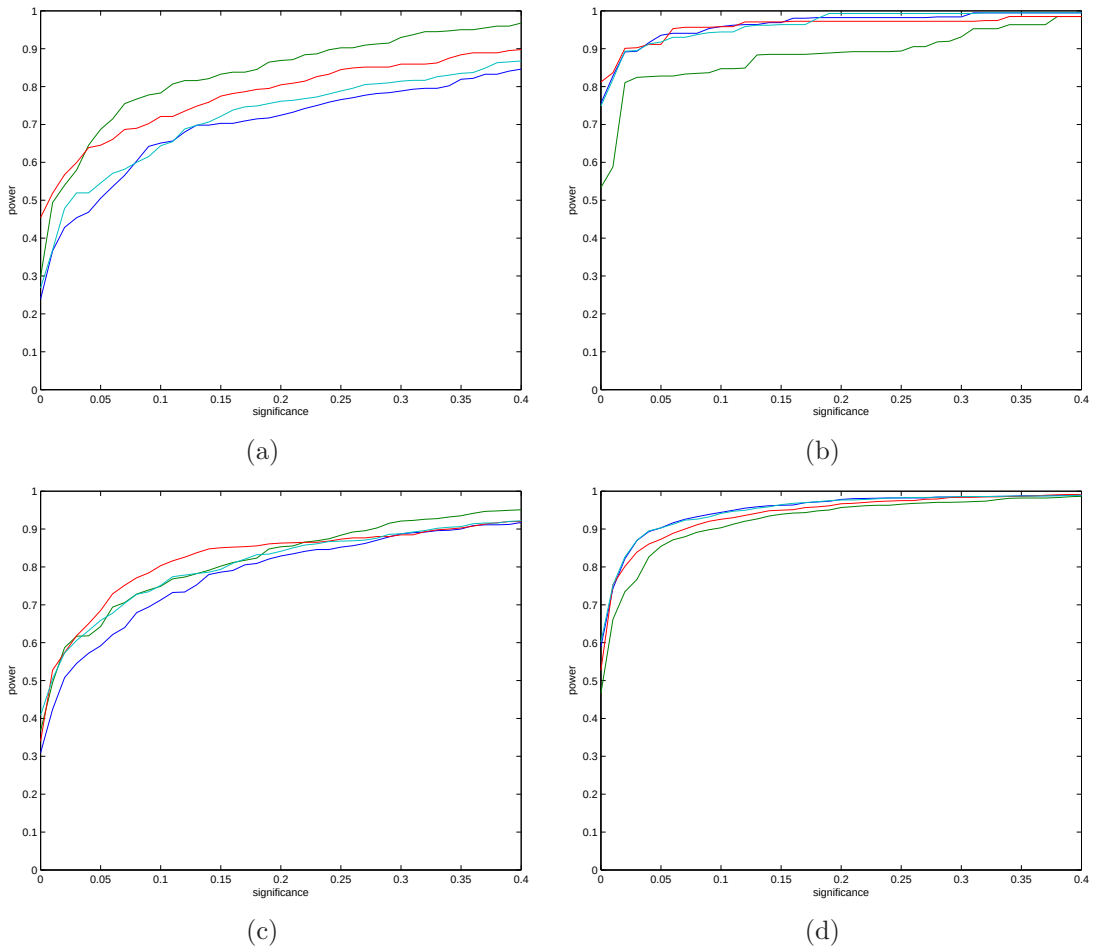


Figure 4.9: Power of HAPCLUSTER, HAPSEARCH, MMP and Single SNP methods to detect causal variants that are **a** rare not tagged, **b** rare tagged, **c** common not tagged, **d** common tagged. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

Method	Disease SNP type			
	rare not tagged	rare tagged	common not tagged	common tagged
HAPSEARCH	0.69	0.83	0.64	0.85
HAPCLUSTER	0.65	0.91	0.69	0.87
Single SNP	0.51	0.94	0.59	0.90
MMP	0.55	0.92	0.66	0.90

Table 4.7: Power of each method at the 5% significance level, stratified by the type of causal variant. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

4.7 Discussion

In this chapter we have presented two methods for detecting association in case-control haplotype data. Our methods rely on the LS model (Li and Stephens, 2003), which naturally incorporates fine-scale recombination rates to model the extent of LD around a given region. With the use of a HMM, we can find the pair-wise haplotype similarity at any position including at untyped SNPs. This allows us to test for association at any position whereas SNP based methods are restricted to typed markers and a limited number of untyped markers tagged by MMPs.

In Section 4.4 we introduced HAPSEARCH, which detects elevated similarity in haplotype background within the case and control samples at a position of interest. Our approach is similar to those presented by Tzeng et al. (2003) but with the added advantage that our similarity measure can consider data at all SNPs and does not require the specification of a window size. We attempted to find the best test statistics that can detect a signal of association, in the form of elevated sharing within the case and control samples, and tried 5 non-parametric test statistics. We found two test statistics, both based on the logarithmic geometric means of haplotype similarity, to be the most powerful for detection. One possible explanation

for this is that with large sample sizes, the average pairwise haplotype similarity is low. This is because the number of off-diagonal elements in our similarity matrix, M^j, at SNP j is $(m+n)(m+n-1)$ but those elements only sum up to $m+n$. Transformation on a logarithmic scale will therefore be more sensitive to small differences in the elements of the similarity matrix.

In complex disease etiologies we would expect phenocopies, incomplete penetrance and allelic heterogeneity, which might leave HAPSEARCH underpowered. We therefore proposed HAPCLUSTER in Section 4.5, which is a more sophisticated parametric haplotype association test. Our approach is similar to that presented by Durrant et al. (2004) and Molitor et al. (2003b). A haplotype tree is constructed using hierarchical clustering, which partitions the haplotypes into basis clusters at a given level of the tree and we form groupings of haplotypes by merging the basis clusters. This addresses the the problem of high haplotype diversity leading to large multiple comparisons and data sparsity. Each group in a grouping is given an independent penetrance parameter and we perform a Bayesian test for association to detect an excess of case haplotypes in a group. We average over all groupings that can be generated from the basis clusters to produce a Bayes Factor, which in effect integrates over a space of tree topologies to account for some of the uncertainty in the estimated haplotype tree.

We have adopted a Bayesian approach and our results are summarised by Bayes Factors. Bayes Factors are the alternative to the Frequentist hypothesis tests commonly used to test for association and their use is beginning to emerge in the literature as a more powerful alternative to classical tests of association (Balding, 2006b; Servin and Stephens, 2007; Marchini et al., 2007). The Bayesian approach offers us a coherent probabilistic framework to work from, in particular it allows us perform

inference over a set of disease models and combine the evidence from each by averaging over the corresponding set of Bayes Factors. For example, the evidence for association conditional on a grouping is a Bayes Factor, so the evidence for association can summarised by their average, which is also a well-defined Bayes Factor. Bayes Factors also allow for the incorporation of useful prior knowledge about the parameters in our disease model that we might expect to observe, which can lead to more accurate estimates and increased power. Our approach includes a prior on disease penetrance and our simulations show that we can increase power if it is well-calibrated with respect to the underlying disease model.

In Section 4.6.4 we compared our methods with the single SNP χ^2 test applied at typed SNPs and at SNPs predicted by MMPs. Overall we found HAPCLUSTER to be the most powerful method under a single disease SNP model. Further analyses showed that HAPSEARCH and HAPCLUSTER provide a significant boost in power when the disease allele is untagged. This is expected since our methods consider data at all SNPs simultaneously, and is therefore able to detect associations with untagged alleles from the haplotypes upon which they reside. However, when the disease allele is tagged, we found the SNP based methods to be more powerful than our methods. This again is expected, since previous results predict that the SNP based methods are powerful when the disease SNP is highly correlated with a typed SNP (Chapman et al., 2003; Pritchard and Przeworski, 2001). We therefore recommend that our methods should be complementary to, rather than a replacement for, the SNP based methods that current association studies use.

A major weakness of our methods is that they require phased haplotype data as input whereas usually in association studies only unphased genotype data are available. Although accurate phasing methods are available (Stephens et al., 2001), we

have discounted the uncertainty in the estimated haplotype phase. Marchini et al. (2007) applied the LS model to imputing missing SNPs in genotype data. A natural extension to our methods would therefore be to incorporate phasing into our model and extend their applications to genotype data.

In the next chapter we will take some of the ideas presented in this chapter to formulate a more sophisticated approach that is applicable to genotype data whilst remaining computationally tractable. In addition, we will account for multiple disease mutations in our disease model so that we can retain power in the presence of allelic heterogeneity.

Chapter 5

GENECLUSTER: A Powerful Method for Analysing Association Data

In this chapter we will describe a new approach, which we call GENECLUSTER, and extends some of the ideas from the last chapter to handle genotype data and missing data. Our approach is to condition on a reference panel of known haplotypes and model each genotype as a pair of unknown haplotypes, which are imperfect mosaics of the reference haplotypes. The main feature of our method is the use of a sophisticated disease model, which can account for multiple disease mutations and allow us to retain power and identify allelic heterogeneity when it exists.

Our approach proceeds in 3 stages.

1 At each position across the genome we use a panel of known haplotype vari-

ation such as HapMap (The International HapMap Consortium, 2005; Frazer et al., 2007) and an estimate of the fine-scale recombination rate to construct an approximation to the genealogy of the sample of haplotypes.

2 We then consider the genotype data for each case and control individual and effectively map the individual's pair of haplotypes to the tips of the genealogical tree from step 1 at each position. This step involves calculating the probability distribution that the individual's pair of haplotypes copy the panel haplotypes. This probability distribution is calculated based on a model similar to that used by IMPUTE (Marchini et al., 2007). In carrying out this step we average over the phase uncertainty of the haplotypes of each individual.

3 We treat the genealogical tree constructed in step 1, with the haplotype mapping in step 2, as a genealogical tree relating the haplotypes of the case and control individuals. Based on this we assess the evidence that the locus is associated with disease status by fitting a model in a Bayesian framework. At each position, we calculate a Bayes Factor that compares a model of association with a null model of no association. We propose two sets of models of association. The first is similar to the HAPCLUSTER model, where we summarise the lower topology of the tree in the form of basis clusters and construct groupings by merging them; haplotype penetrance is modelled by group membership. The second model is similar to the one proposed by Minichiello and Durbin (2006), where we place putative, disease causing, mutations on the tree and penetrance is modelled by the number of disease alleles carried by each haplotype.

We will first provide details on each of the above steps. Later in the chapter we will

describe two simulation studies, where first we simulate a disease model involving a single causal mutation and then under a disease model involving allelic heterogeneity. Our results show that GENECLUSTER is applicable to genome-wide association studies and is a powerful method for detecting association with single and multiple causal variants. In the following chapter we will evaluate GENECLUSTER further by applying it to the large scale association data obtained from the WTCCC study (The Wellcome Trust Case Control Consortium, 2007).

5.1 Method

5.1.1 Notations

We will describe in this section how we test for association at a single position of interest. We will use the following notations throughout the chapter but we will also add notations as we proceed. We let

- x be the position of interest;
- $\mathbf{G} = \{G_1, \ldots, G_M\}$ be the genotype data of M individuals, where $G_i = \{G_{i,x_1}, \ldots, G_{i,x_L}\}$ is the genotype data of the ith individual typed at the L positions $\{x_1, \ldots, x_L\}$ with $G_{i,x_j} \in \{0, 1, 2, \text{missing}\}$;
- $\mathbf{\Phi} = \{\Phi_1, \ldots, \Phi_M\}$ be the phenotype information, where $\Phi_i \in \{0 = \text{control}, 1 = \text{case}\}$ is the phenotype of the ith individual;
- $\mathbf{H} = \{H_1, \ldots, H_P\}$ be the set of P haplotypes in a reference panel, where $H_i = \{H_{i,x_1}, \ldots, H_{i,x_L}\}$ is the ith haplotype and $H_{i,x_j} \in \{0, 1\}$.

5.1.2 Step 1: Constructing genealogy for the panel haplotypes

In this step we construct a genealogical tree, relating the reference panel haplotypes at x. For genome-wide association studies we would therefore be required to construct a set of trees across the genome.

We use a tree construction method, TREESIM (Cardin, 2007), based on the coalescent model with recombination, which at each locus constructs the posterior modal tree given the haplotypes. We use the recombination rates estimated from the HapMap and an infinite sites mutation model to construct our tree. An added feature of TREESIM is that it can be adapted to produce a sample of trees along with their probabilities under the coalescent model with recombination. Details of TREESIM are provided in Appendix C.

In Section 4.5.1 we described how a haplotype tree can be constructed from the LS model (Li and Stephens, 2003). This approach, which we will call LSTREE, constructs a tree that we hope is an accurate estimation of the true genealogy. Our construction assumes that the similarity between a pair of haplotypes, given by the posterior probability of their copying states, is a measure of relatedness and we successively merge the closest related haplotype pair. We will evaluate the use of TREESIM and LSTREE trees with GENECLUSTER in the simulation studies later in this chapter.

An advantage of our method is that since the tree at position x is constructed using only the panel haplotypes, it need only be constructed once and can be stored for any future analyses using the same panel.

5.1.3 Step 2: Constructing genealogy for the case-control sample

In this step we use a HMM to map the haplotypes of each case and control individual to the leaves of the tree constructed in step 1. The objective of this procedure is to use the tree to construct a genealogy for the case-control sample, which we will use to perform inference in step 3. We will describe our HMM in this section, which is very similar to the one used by IMPUTE (Marchini et al., 2007).

First, here is a summary of the additional notations in this section; we use

- T_x to be the tree estimated at x for $\mathbf{H}$;
- $(h_i^{(1)}, h_i^{(2)})$ to be the pair of unknown haplotypes of individual i, so that $h_{i,x_j}^{(1)}, h_{i,x_j}^{(2)} \in \{0,1\}$ and $(h_{i,x_j}^{(1)} + h_{i,x_j}^{(2)}) = G_{i,x_j}$ at all SNPs x_j, $j \in \{1,\ldots,L\}$;
- $Z_{i,x_j} = (Z_{i,x_j}^{(1)}, Z_{i,x_j}^{(2)})$ to be the copying states of $(h_i^{(1)}, h_i^{(2)})$ at position x_j;
- $n_{i,x}$ and $m_{i,x}$ to be the expected number of case and control haplotypes mapped to the leaf in T_x corresponding to haplotype $H_i \in \mathbf{H}$.

We will now explain our HMM in detail and derive the expressions for Z_{i,x_j}, $n_{i,x}$ and $m_{i,x}$.

We model $G_i \in \mathbf{G}$ as a pair of unobserved haplotypes, $h_i^{(1)}$ and $h_i^{(2)}$. In accordance to the LS model, we model $h_i^{(1)}$ and $h_i^{(2)}$ to be imperfect mosaics of the panel haplotypes. That is, at each SNP x_j the alleles $h_{i,x_j}^{(1)}$ and $h_{i,x_j}^{(2)}$ are imperfect copies of the alleles H_{k_1,x_j} and H_{k_2,x_j} on haplotypes H_{k_1} and H_{k_2} in $\mathbf{H}$ conditional on the copying states $Z_{i,x_j} = (k_1, k_2)$.

Our HMM has hidden states $Z_{i,x_j} = (Z^{(1)}_{i,x_j}, Z^{(2)}_{i,x_j})$, which evolves according to the observed variables G_i and $\mathbf{H}$. The emission probabilities given in Equation (5.1) and transition probabilities given in Equation (5.3) are extensions of the single copying state version that we used for HAPSEARCH and HAPCLUSTER.

If G_{i,x_j} is not missing, then

$$\Pr(G_{i,x_j}|Z_{i,x_j} = (k_1, k_2), \mathbf{H}, \tilde{\theta}) = \sum_{\substack{h^{(1)}_{i,x_j}, h^{(2)}_{i,x_j}: \\ h^{(1)}_{i,x_j} + h^{(2)}_{i,x_j} = G_{i,x_j}}} \Pr(h^{(1)}_{i,x_j}|H_{k_1,x_j}, \tilde{\theta}) \Pr(h^{(2)}_{i,x_j}|H_{k_2,x_j}, \tilde{\theta}), \tag{5.1}$$

where

$$\Pr(h^{(\cdot)}_{i,x_j}|H_{k,x_j}, \tilde{\theta}) \;\; = \;\; \begin{cases} P/(P+\tilde{\theta}) + (1/2) \times \tilde{\theta}/(P+\tilde{\theta}), & \text{if } h^{(\cdot)}_{i,x_j} = H_{k,x_j} \\ (1/2) \times \tilde{\theta}/(P+\tilde{\theta}) & \text{if } h^{(\cdot)}_{i,x_j} \neq H_{k,x_j}, \end{cases}$$

and $\tilde{\theta}$ is the mutation parameter, which we set $\tilde{\theta} = (\sum_{i=1}^{(P-1)} \frac{1}{i})^{-1}$ as before.

If G_{i,x_j} is missing, then

$$\Pr(G_{i,x_j} = \text{missing}|Z_{i,x_j} = (k_1, k_2), \mathbf{H}, \tilde{\theta}) = 1. \tag{5.2}$$

The transition probabilities assume independence in the two copying states

$$\Pr\big(Z_{i,x_j+1} = (k'_1, k'_2)|Z_{i,x_j} = (k_1, k_2), \rho_j\big) = \Pr(k'_1|k_1, \rho_j) \Pr((k'_2|k_2, \rho_j), \tag{5.3}$$

where

$$\Pr(k'|k,\rho_j) = \begin{cases} \exp(-\rho_j d_j/P) + (1-\exp(-\rho_j d_j/P))(1/P) & \text{if } k'=k \\ (1-\exp(-\rho_j d_j/P))(1/P) & \text{otherwise,} \end{cases}$$

d_j is the physical distance between x_j and x_{j+1}, $\rho_j = \delta_j \rho'_j$, δ_j is the correction term to adjust for biases in the recombination rates suggested by Li and Stephens (2003) (see Section 2.2.4), $\rho'_j = 4N_e c_j$, N_e is the effective (diploid) population size and c_j is the average rate of crossover per unit physical distance, per meiosis, between x_j and x_{j+1}. We use the standard forward and backward algorithms to calculate the marginal posterior probability distribution $\Pr((Z^{(1)}_{i,x_j}, Z^{(2)}_{i,x_j})|\mathbf{H}, G_i, \tilde{\theta}, \rho)$, where $\rho = (\rho_0, \ldots, \rho_{(L-1)})$ is a vector of rescaled fine-scale recombination rates.

For the forward path

$$\begin{aligned} f_i^{x_{(j+1)}}(k_1,k_2) &= \Pr\big(G_{i,x_1},\ldots,G_{i,(x_{(j+1)})}, Z_{i,x_{(j+1)}} = (k_1,k_2)|\mathbf{H},\tilde{\theta},\rho\big) \\ &= \Pr\big(G_{i,x_{(j+1)}}|Z_{i,x_{(j+1)}} = (k_1,k_2),\tilde{\theta},\mathbf{H}\big) \times \\ &\quad \sum_{k'_1,k'_2} \Pr\big(Z_{i,x_{(j+1)}} = (k_1,k_2)|Z_{i,x_j} = (k'_1,k'_2),\rho_j\big) f_i^{x_j}(k'_1,k'_2), \end{aligned} \tag{5.4}$$

with initial values $f_i^{x_0}(k_1,k_2) = \frac{1}{P^2}$.

Similarly, for the backward path

$$\begin{aligned} b_i^{x_{(j-1)}}(k_1,k_2) &= \Pr\big(G_{i,x_j},\ldots,G_{i,x_L}|Z_{i,x_{(j-1)}}=(k_1,k_2),\mathbf{H},\tilde{\theta},\rho\big) \\ &= \sum_{k_1',k_2'}\Big(\Pr\big(G_{i,x_j}|Z_{i,x_j}=(k_1',k_2'),\tilde{\theta},\mathbf{H}\big)\times \\ &\quad \Pr\big(Z_{i,x_j}=(k_1',k_2')|Z_{i,x_{(j-1)}}=(k_1,k_2),\rho_{(j-1)}\big)b_i^{x_j}(k_1',k_2')\Big), \end{aligned} \tag{5.5}$$

with initial values $b_i^{x_L}(k_1',k_2')=1.0$.

The posterior probability of the hidden state Z_{i,x_j} is therefore

$$\Pr(Z_{i,x_j}=(k_1,k_2)|\mathbf{H},G_i,\tilde{\theta},\rho)\propto f_i^{x_j}(k_1,k_2)b_i^{x_j}(k_1,k_2). \tag{5.6}$$

Although computations of the forward and backward probabilities given above are standard, extra care is required to ensure that these computations have complexity $O(P^2)$ and not $O(P^4)$ (Scheet and Stephens, 2006).

We calculate the forward and backward probabilities using

$$\begin{aligned} f_i^{x_{(j+1)}}(k_1,k_2) &= \Pr(G_{i,x_{(j+1)}}|Z_{i,x_{(j+1)}}=(k_1,k_2),\tilde{\theta},\mathbf{H})\Big((\beta-\alpha)^2 f_i^{x_j}(k_1,k_2)+ \\ &\quad \alpha(\beta-\alpha)\big(\sum_{k'}f_i^{x_j}(k_1,k')+\sum_{k'}f_i^{x_j}(k',k_2)\big)+\alpha^2\sum_{k_1',k_2'}f_i^{x_j}(k_1',k_2')\Big) \end{aligned} \tag{5.7}$$

and

$$
\begin{aligned}
b_i^{x_{(j-1)}}(k_1,k_2) &= \Big((\beta-\alpha)^2 b_i^{x_j}(k_1,k_2)\Pr(G_{i,x_j}|Z_{i,x_j}=(k_1,k_2),\tilde{\theta},\mathbf{H}) + \\
&\quad \alpha(\beta-\alpha)\big(\sum_{k'} b_i^{x_j}(k_1,k')\Pr(G_{i,x_j}|Z_{i,x_j}=(k_1,k'),\tilde{\theta},\mathbf{H}) + \\
&\quad \sum_{k'} b_i^{x_j}(k',k_2)\Pr(G_{i,x_j}|Z_{i,x_j}=(k',k_2),\tilde{\theta},\mathbf{H})\big) + \\
&\quad \alpha^2 \sum_{k_1',k_2'} b_i^{x_j}(k_1',k_2')\Pr(G_{i,x_j}|Z_{i,x_j}=(k_1',k_2'),\tilde{\theta},\mathbf{H})\Big),
\end{aligned}
\tag{5.8}
$$

where $\alpha = (1-\exp(-\rho_j d_j/P))(1/P)$ and $\beta = \exp(-\rho_j d_j/P)+(1-\exp(-\rho_j d_j/P))(1/P)$.

As before, our model can be extended to untyped positions. Suppose that position $x_{j'}$ is untyped and is flanked by typed SNPs x_j and x_{j+1}, then we assume that $\Pr(G_{i,x_{j'}}|Z_{i,x_{j'}}=(k_1,k_2),\theta,\mathbf{H})=1$, i.e. we effectively treat $G_{i,x_{j'}}$ as missing:

$$
f_i^{x_{j'}}(k_1,k_2) = \sum_{k_1',k_2'} \Pr(Z_{i,x_{j'}}=(k_1,k_2)|Z_{i,x_j}=(k_1',k_2'),\rho_j) f_i^{x_j}(k_1',k_2'), \tag{5.9}
$$

$$
b_i^{x_{j'}}(k_1,k_2) = \sum_{k_1',k_2'} \Pr(Z_{i,x_{(j+1)}}=(k_1',k_2')|Z_{i,x_{j'}}=(k_1,k_2),\rho_j) b_i^{x_{(j+1)}}(k_1',k_2'), \tag{5.10}
$$

and thus

$$
\Pr(Z_{i,x_{j'}}=(k_1,k_2)|\mathbf{H},G_i,\tilde{\theta},\rho) \propto f_i^{x_{j'}}(k_1,k_2) b_i^{x_{j'}}(k_1,k_2). \tag{5.11}
$$

We are now in a position to map the haplotypes of each case-control individual to the leaves of T_x and thus construct a genealogy for our case-control sample $\mathbf{G}$. We introduce the following notations; let

- $\mathbf{h} = \{h_1, \ldots, h_{2M}\}$ be the set of unknown haplotypes of $\mathbf{G}$ so that $(h_{(2i-1)}, h_{2i}) = (h_i^{(1)}, h_i^{(2)})$;

- $\mathbf{z}_x = \{z_{1,x}, \ldots, z_{2M,x}\}$ be the corresponding copying states for haplotypes in $\mathbf{h}$, i.e. $(z_{(2i-1),x}, z_{2i,x}) = (Z_{i,x}^{(1)}, Z_{i,x}^{(2)})$;

- $\phi = \{\phi_1, \ldots, \phi_{2M}\}$ be the phenotype information for $\mathbf{h}$, i.e. $\phi_{(2i-1)} = \phi_{2i} = \Phi_i$.

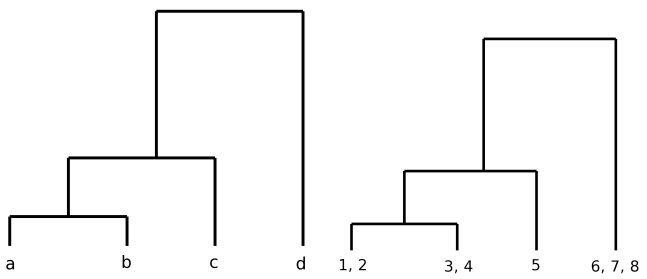
We make the following modelling assumptions.

- $\mathbf{H}$ is a set of ancestral haplotypes of $\mathbf{h}$.

- Given $\mathbf{z}_x$, at time 0 the case-control haplotypes, $\{h_j : z_{j,x} = i\}$, coalesced to the ancestral haplotype H_i, for each $H_i \in \mathbf{H}$. In other words, the copying states map each haplotype of our case-control sample, $\mathbf{G}$, to a haplotype in $\mathbf{H}$. However, since mapping to haplotype $H_i \in \mathbf{H}$ is equivalent to mapping to the leaf that corresponds to H_i in T_x, $(T_x,\ \mathbf{z}_x)$ describes a genealogy for $\mathbf{G}$. For example if $z_{(2j-1),x} = i_1$ and $z_{2j,x} = i_2$, then one of the haplotypes of individual j in $\mathbf{G}$ is mapped to the leaf of T_x, which corresponds to $H_{i_1} \in \mathbf{H}$, and the other to the leaf, which corresponds to $H_{i_2} \in \mathbf{H}$.

As a simple example, let the panel contain 4 haplotypes, labelled $\mathbf{H} = \{a, b, c, d\}$, with tree, T_x, displayed in Figure 5.1(a), and the case-control sample contain 8 haplotypes, labelled $\mathbf{h} = \{1, 2, \ldots, 8\}$. Conditional on the copying states $\mathbf{z}_x = \{z_1 = a, z_2 = a, z_3 = b, z_4 = b, z_5 = c, z_6 = d, z_7 = d, z_8 = d\}$, the constructed genealogy for $\mathbf{G}$, (T_x, z_x), is displayed in Figure 5.1(b).

In the following step we will perform inferences based on $(T_x, \mathbf{z}_x)$ as the genealogical tree for $\mathbf{G}$. By doing so, we make the following approximations.



(a) (b)

Figure 5.1: (a) The genealogical tree for a panel of haplotypes labelled $\{a, b, c, d\}$. (b) The constructed genealogy for the haplotypes $\mathbf{h} = \{1, 2, \ldots, 8\}$ conditional on the tree in (a) and the copying states $\mathbf{z} = (z_1 = a, z_2 = a, z_3 = b, z_4 = b, z_5 = c, z_6 = d, z_7 = d, z_8 = d)$.

.

1 We approximate the time taken for all lineages originating from a case-control haplotype in $\mathbf{h}$ to coalesce with a lineage originating from a panel haplotype in $\mathbf{H}$ as 0. This will be valid if every haplotype in $\mathbf{h}$ is closely related to a haplotype in $\mathbf{H}$. For example, in the best case scenario, the first $2M - P$ coalescence events all involve at most one lineage that originates from a panel haplotype, so that after those coalescence events the only surviving lineages are those that originate from at least one panel haplotype. In the case of $M = 4000$ and $P = 120$, the expected time taken for the first $2M - P$ coalescence events, under the coalescent model (Kingman, 1982), is 0.016 (in $4N_e$ generations), which is small compared to 1.98, the time take taken for the last P coalescence events.

2 Let T_x^G be the true genealogical tree for $\mathbf{G}$, we assume that after the first $2M - P$ coalescence events, T_x^G takes a similar form to T_x. In other words, we approximate the upper levels of T_x^G with T_x. For this approximation to work well the panel, $\mathbf{H}$, and the case-control sample, $\mathbf{G}$, must come from a common ancestral population. $\mathbf{H}$ must also contain enough genetic variation local to x in order to provide good proxies for the true ancestral haplotypes

to the haplotypes of **G**.

3 Finally, we assume that the copying states are a good model for approximating the ancestor, in **H**, of each case-control haplotype of **G**. IMPUTE (Marchini et al., 2007) and FASTPHASE (Scheet and Stephens, 2006) make similar assumptions to impute missing SNPs, and the accuracy of their results lead us to believe that this is a good approximation.

We do not claim that our approach leads to good approximations of the true genealogy of **G** under the coalescent model, however we do hope that it leads to good approximations for the purposes of detecting association. We will be assessing our approach under simulation later in this chapter and applied to population data in the next chapter.

5.1.4 Step 3: Testing for association

In this section we will describe how we test for association based on the constructed genealogy for **G** described in step 2. We use a Bayesian framework and at the position of interest, x, our test statistic is the Bayes Factor

$$\begin{aligned} BF(x) &= \frac{\Pr(\mathbf{G}, \mathbf{\Phi}|\text{association at } x)}{\Pr(\mathbf{G}, \mathbf{\Phi}|\text{no association at } x)} \\ &= \frac{\Pr(\mathbf{G}, \mathbf{\Phi}|M_1, x)}{\Pr(\mathbf{G}, \mathbf{\Phi}|M_0, x)}, \end{aligned} \tag{5.12}$$

where M_1 and M_0 are the alternative model of association and the null model of no association, respectively.

Here are some more notations, let

- $T_x^+ = (T'_x, \mathbf{z}'_x)$ be our constructed genealogy for $\mathbf{G}$ at x, comprised of a genealogical tree for $\mathbf{H}$, T'_x, and a mapping from $\mathbf{G}$ to T'_x, $\mathbf{z}'_x$, at x;

- ${\mathscr{T}^+}_x$ be the set of possible T_x^+, so that ${\mathscr{T}^+}_x = \mathscr{T}_x \times \mathscr{Z}_x$, where $\mathscr{T}_x$ is the set of possible genealogical trees for the panel $\mathbf{H}$ and $\mathscr{Z}_x$ is the set of mappings from $\mathbf{G}$ to $\mathbf{H}$;

- $\Theta = (\mathbf{H}, \tilde{\theta}, \rho)$, which are the resources we need to perform inference;

- $\lambda^A_{i,\mathbf{z}_x}$ be the number of case (Affected) haplotypes of $\mathbf{G}$ that $H_i \in \mathbf{H}$ is ancestral to, conditional on $\mathbf{z}_x \in \mathscr{Z}_x$, i.e. $\lambda^A_{i,\mathbf{z}_x} = \sum_{j:\phi_j=1} \mathrm{I}(z_{j,x} = i)$, similarly $\lambda^U_{i,\mathbf{z}_x} = \sum_{j:\phi_j=0} \mathrm{I}(z_{j,x} = i)$ is the number of control (Unaffected) haplotypes that $H_i \in \mathbf{H}$ is ancestral to;

- $\Lambda_{\mathbf{z}_x} = \{(\lambda^U_{1,\mathbf{z}_x}, \lambda^A_{1,\mathbf{z}_x}), \ldots, (\lambda^U_{P,\mathbf{z}_x}, \lambda^A_{P,\mathbf{z}_x}))$, which summarises $\mathbf{z}_x$ in terms of the number of case and control haplotypes of $\mathbf{G}$ that $\mathbf{z}_x$ maps to each haplotype in $\mathbf{H}$;

- $\mathscr{L}_x = \{\Lambda_{{\mathbf{z}'}_x} : {\mathbf{z}'}_x \in \mathscr{Z}_x\}$, which is the set of possible $\Lambda_{\mathbf{z}_x}$ created by a mapping in $\mathscr{Z}_x$;

- $\bar{\Lambda}_x = \{(\mathbb{E}(\lambda^U_{1,\mathbf{z}_x}), \mathbb{E}(\lambda^A_{1,\mathbf{z}_x})), \ldots, (\mathbb{E}(\lambda^U_{P,\mathbf{z}_x}), \mathbb{E}(\lambda^A_{P,\mathbf{z}_x}))\}$, which are the expected number of case and control haplotypes that $H_i \in \mathbf{H}$ is ancestral to under the

posterior probability distribution on $\mathbf{z}_x$, where

$$\begin{aligned} m_{i,x} = \mathbb{E}(\lambda^U_{i,\mathbf{z}_x}) &= \sum_{j:\phi_j=0} \Pr(z_{j,x} = i|\mathbf{H}, G_j, \tilde{\theta}, \rho) \\ &= \sum_{j:\Phi_j=0} \sum_{k=1}^{P} \Big(\Pr(Z_{j,x} = (i,k)|\mathbf{H}, G_j, \tilde{\theta}, \rho) + \\ &\qquad \Pr(Z_{j,x} = (k,i)|\mathbf{H}, G_j, \tilde{\theta}, \rho) \Big), \qquad (5.13) \\ n_{i,x} = \mathbb{E}(\lambda^A_{i,\mathbf{z}_x}) &= \sum_{j:\phi_j=1} \Pr(z_{j,x} = i|\mathbf{H}, G_j, \tilde{\theta}, \rho) \\ &= \sum_{j:\Phi_j=1} \sum_{k=1}^{P} \Big(\Pr(Z_{j,x} = (i,k)|\mathbf{H}, G_j, \tilde{\theta}, \rho) + \\ &\qquad \Pr(Z_{j,x} = (k,i)|\mathbf{H}, G_j, \tilde{\theta}, \rho) \Big). \qquad (5.14) \end{aligned}$$

Under the model $M' \in \{M_0, M_1\}$, we formulate our likelihood in a similar way to Zollner and Pritchard (2005). There are 5 approximating steps, which we will explain below.

$$
\begin{aligned}
\Pr(\mathbf{G},\mathbf{\Phi}|M',x,\Theta) &= \int_{T_x^+\in\mathscr{T}^+{}_x} \Pr(\mathbf{\Phi}|M',x,T_x^+,\Theta)\Pr(\mathbf{G}|M',x,T_x^+,\mathbf{\Phi},\Theta) \\
&\qquad \Pr(T_x^+|M',x,\Theta)\mathrm{d}T_x^+ \\
&\approx \int_{T_x^+\in\mathscr{T}^+{}_x} \Pr(\mathbf{\Phi}|M',x,T_x^+,\Theta)\Pr(\mathbf{G}|T_x^+,\Theta)\Pr(T_x^+|M',x,\Theta)\mathrm{d}T_x^+ \\
&\approx \int_{T_x^+\in\mathscr{T}^+{}_x} \Pr(\mathbf{\Phi}|M',x,T_x^+,\Theta)\Pr(\mathbf{G}|T_x^+,\Theta)\Pr(T_x^+|\Theta)\mathrm{d}T_x^+ \\
&= \int_{T_x^+\in\mathscr{T}^+{}_x} \Pr(\mathbf{\Phi}|M',x,T_x^+,\Theta)\Pr(T_x^+|\mathbf{G},\Theta)\Pr(\mathbf{G}|\Theta)\mathrm{d}T_x^+ \\
&= \int_{(T'_x,\mathbf{z}'_x)\in\mathscr{T}^+{}_x} \Pr(\mathbf{\Phi}|M',x,(T'_x,\mathbf{z}'_x),\Theta)\Pr((T'_x,\mathbf{z}'_x)|\mathbf{G},\Theta) \\
&\qquad \Pr(\mathbf{G}|\Theta)\mathrm{d}T'_x\mathrm{d}\mathbf{z}'_x \\
&= \int_{(T'_x,\mathbf{z}'_x)\in\mathscr{T}^+{}_x} \Pr(\mathbf{\Phi}|M',x,(T'_x,\mathbf{z}'_x),\Theta)\Pr(T'_x|\mathbf{G},\Theta)\Pr(\mathbf{z}'_x|\mathbf{G},\Theta) \\
&\qquad \Pr(\mathbf{G}|\Theta)\mathrm{d}T'_x\mathrm{d}\mathbf{z}'_x \\
&\approx \int_{\mathbf{z}'_x\in\mathscr{Z}_x} \Pr(\mathbf{\Phi}|M',x,T_x,\mathbf{z}'_x,\Theta)\Pr(\mathbf{z}'_x|\mathbf{G},\Theta)\Pr(\mathbf{G}|\Theta)\mathrm{d}\mathbf{z}'_x \\
&= \int_{\Lambda'\in\mathscr{L}_x}\int_{\mathbf{z}'_x\in\mathscr{Z}_x} \Pr(\mathbf{\Phi}|M',x,T_x,\mathbf{z}'_x,\Theta,\Lambda')\Pr(\Lambda'|M',x,T_x,\mathbf{z}'_x,\Theta) \\
&\qquad \Pr(\mathbf{z}'_x|\mathbf{G},\Theta)\Pr(\mathbf{G}|\Theta)\mathrm{d}\mathbf{z}'_x\mathrm{d}\Lambda' \\
&= \int_{\Lambda'\in\mathscr{L}_x}\int_{\mathbf{z}'_x\in\mathscr{Z}_x} \Pr(\mathbf{\Phi}|M',x,T_x,\mathbf{z}'_x,\Theta,\Lambda')\Pr(\Lambda'|\mathbf{z}'_x) \\
&\qquad \Pr(\mathbf{z}'_x|\mathbf{G},\Theta)\Pr(\mathbf{G}|\Theta)\mathrm{d}\mathbf{z}'_x\mathrm{d}\Lambda' \\
&= \int_{\mathbf{z}'_x\in\mathscr{Z}_x} \Pr(\mathbf{\Phi}|M',x,T_x,\mathbf{z}'_x,\Theta,\Lambda_{\mathbf{z}'_x})\Pr(\Lambda_{\mathbf{z}'_x}|\mathbf{z}'_x) \\
&\qquad \Pr(\mathbf{z}'_x|\mathbf{G},\Theta)\Pr(\mathbf{G}|\Theta)\mathrm{d}\mathbf{z}'_x \\
&\approx \int_{\mathbf{z}'_x\in\mathscr{Z}_x} \Pr(\mathbf{\Phi}|M',x,T_x,\Lambda_{\mathbf{z}'_x})\Pr(\Lambda_{\mathbf{z}'_x}|\mathbf{z}'_x)\Pr(\mathbf{z}'_x|\mathbf{G},\Theta) \\
&\qquad \Pr(\mathbf{G}|\Theta)\mathrm{d}\mathbf{z}'_x \\
&\approx \Pr(\mathbf{\Phi}|M',x,T_x,\bar{\Lambda}_x)\Pr(\mathbf{G}|\Theta) \qquad (5.15)
\end{aligned}
$$

where T_x is the tree estimated in step 1 at x. We make the following approximations.

- $\Pr(\mathbf{G}|M',x,T_x^+,\mathbf{\Phi},\Theta) \approx \Pr(\mathbf{G}|T_x^+,\Theta)$. A similar approximation is made in

Zollner and Pritchard (2005). $\Pr(\mathbf{G}|T_x^+, \Theta)$ is the probability of the set of mutations on T_x^+ that is compatible with the observed variation in $\mathbf{G}$. This probability distribution will change if we condition on the phenotype information and x being causal, since they provide information about the location of the causal mutation on T_x^+. Our approximation will hold if we do not observe the SNP x in $\mathbf{G}$, and mutations on T_x^+ occur independently of the causal mutation.

- $\Pr(T_x^+|M', x, \Theta) \approx \Pr(T_x^+|\Theta)$. A similar assumption is made in Zollner and Pritchard (2005). This approximation implies that in the absence of the phenotype data, the tree topology itself contains no or only a small amount of information about the location of the disease mutation. Thus, we ignore the effects of selection and the ascertainment scheme that might affect the distribution of branch lengths of T_x. Consider, for example, a rare and penetrant disease mutation at x, under the model of association an over-sampling of the case haplotypes might result in a tree that partitions the haplotypes into two distantly related groups (the case and control haplotypes). Zollner and Pritchard (2005) relied on the data being strong enough to overcome this approximation and it is reasonable to use the same justification here since our data will be on much larger scale than those considered in that paper.

- $\Pr((T'_x, \mathbf{z}'_x)|\mathbf{G}, \Theta) = \Pr(T'_x|\mathbf{G}, \Theta)\Pr(\mathbf{z}'_x|\mathbf{G}, \Theta)$. This is a consequence of the way we have constructed the genealogy for $\mathbf{G}$, so that T_x is estimated conditionally independent of $\mathbf{z}'_x$ given $\mathbf{H}$ and the copying states are determined conditionally independent of T'_x given $\mathbf{G}$ and Θ.

- $\Pr(T'_x = T_x|\mathbf{G}, \Theta) \approx 1$. We assume that the true tree for reference panel is T_x and make this approximation to reduce the computation burden. We

underestimate the uncertainty in our estimated tree, which may result in a loss of power. Averaging over a set of sampled trees should be straightforward and we will discuss this in the final chapter as an improvement for our method.

- $\Pr(\Lambda'|M', x, T_x, \mathbf{z}'_x, \Theta) = \Pr(\Lambda'|\mathbf{z}'_x)$ and $\Pr(\Lambda' = \Lambda_{\mathbf{z}'_x}|\mathbf{z}'_x) = 1$. These result from the definition of Λ': conditional on the mapping, $\mathbf{z}'_x$, the number of case and control haplotypes mapped to each leaf in T_x is uniquely determined as $\Lambda_{\mathbf{z}'_x}$.

- $\Pr(\mathbf{\Phi}|M', x, T_x, \mathbf{z}'_x, \Theta, \Lambda_{\mathbf{z}'_x}) \approx \Pr(\mathbf{\Phi}|M', x, T_x, \Lambda_{\mathbf{z}'_x})$. We base our inference on the number of case and control haplotypes mapped to each leaf of T_x, which is a summary of the mapping $\mathbf{z}'_x$. In order to model a general diploid disease, we need to know the number of disease alleles carried by each individual. In our model of association this is equivalent to knowing which leaves the pair of haplotypes of each individual are mapped to. Conditioning on $\Lambda_{\mathbf{z}'_x}$ means that we no longer have this information. Therefore, we can no longer use a diploid disease model and model for example dominance.

- $\Pr(\Lambda_{\mathbf{z}'_x} = \bar{\Lambda}_x|\mathbf{z}'_x) \approx 1$. We make this approximation to reduce the computation burden, which underestimates the variance of the copying states. Averaging over a set of $\mathbf{z}'_x$ sampled from the posterior distribution can be straightforward and we will discuss this in the final chapter as an improvement for our method.

Substituting Equation (5.15) into Equation (5.12) we obtain

$$BF(x) = \frac{\Pr(\mathbf{\Phi}|M_1, x, T_x, \bar{\Lambda}_x)}{\Pr(\mathbf{\Phi}|M_0, x, T_x, \bar{\Lambda}_x)}. \tag{5.16}$$

We model the phenotype data based on the estimated genealogical tree. We hope

to capture the signal, of non-random clusterings of case haplotypes, that one would expect to observe in the leaves of the true genealogical tree for $\mathbf{G}$ at a disease locus, and we propose the following two models.

1 Model A. We place k ($k \geq 1$) mutations on the tree. Each mutation defines a hypothetical disease mutation that is propagated to the leaves. Disease risk is determined by the number of mutations carried by each haplotype.

2 Model B. This is the HAPCLUSTER approach of cutting the tree at a certain level to create k ($k \geq 2$) basis clusters and merging the basis clusters creates a set of haplotype groupings.

We will now give further details of how we formulate the Bayes Factors for each approach.

Model A

This method works by averaging over the possible locations of a disease causing mutation on the branches, $\mathbf{B} = \{b_1, \ldots, b_{(2P-2)}\}$, in T_x, where T_x is the estimated genealogical tree for $\mathbf{H}$. Let us first consider a single mutation model of association where there is a single disease causing mutation on a branch $b \in \mathbf{B}$. This mutation defines a hypothetical untyped disease SNP, d, that can be added into the panel haplotypes. Conditional on this mutation, we model the penetrance of the haplotypes of $\mathbf{G}$, which are mapped to the leaves of T_x, according to their allele at d. Under the null model of no association, we model phenotype as independent of the

allele at d. Our Bayes Factor under this model therefore becomes

$$\begin{aligned} BF(x) &= \frac{\Pr(\mathbf{\Phi}|M_1, x, T_x, \bar{\Lambda}_x)}{\Pr(\mathbf{\Phi}|M_0, x, T_x, \bar{\Lambda}_x)} \\ &= \frac{\sum_{b\in\mathbf{B}} \Pr(\mathbf{\Phi}|M_1, x, T_x, \bar{\Lambda}_x, b)\Pr(b|M_1, x, T_x, \bar{\Lambda}_x)}{\Pr(\mathbf{\Phi}|M_0, x, T_x, \bar{\Lambda}_x)}. \end{aligned} \tag{5.17}$$

We will now derive the expression for $BF(x)$ and use the following additional notations; let

- m, n be the number of control and case haplotypes in $\mathbf{G}$, i.e. $n = 2\sum_{i=1}^{M} \Phi_i$ and $m = 2M - n$,
- $\mathbf{B}$ be the set of branches on T_x;
- $b \in \mathbf{B}$ be the branch on which we place a putative disease mutation;
- d be the hypothetical SNP defined by the mutation on b;
- $\mathbf{H}^{x,b}$ be the set of panel haplotypes augmented with the disease SNP, d, created by the mutation on b, so that $H_{i,d}^{x,b} \in \{A_b, a_b\}$ is the allele carried by the haplotype H_i at d;
- $\gamma_{A_b} = \Pr(\Phi = 1|A_b)$ to be the penetrance parameter for haplotypes carrying the allele A_b at d under our model of association, and similarly $\gamma_{a_b} = \Pr(\Phi = 1|a_b)$;
- $\gamma_{M_0} = \Pr(\Phi = 1)$ be the penetrance parameter under the null model.

The expected numbers of control and case haplotypes carrying the disease allele A_b

are given by

$$m_{A_b} = \sum_{i=1}^{P} m_{i,x} \mathrm{I}(H_{i,d}^{x,b} = A_b), \tag{5.18}$$

$$n_{A_b} = \sum_{i=1}^{P} n_{i,x} \mathrm{I}(H_{i,d}^{x,b} = A_b), \tag{5.19}$$

where $m_{i,x}$ and $n_{i,x}$ are defined in Equations (5.13) and (5.14). We can therefore summarise the data at d in a contingency table (Table 5.1).

	a_b	A_b
$\Phi = 0$	$m_{a_b} = m - m_{A_b}$	m_{A_b}
$\Phi = 1$	$n_{a_b} = n - n_{A_b}$	n_{A_b}
Penetrance	γ_{a_b}	γ_{A_b}

Table 5.1: A summary of the phenotype and expected allele counts at SNP d, created by a mutation on branch b, with the corresponding penetrance parameters.

As an example, consider the tree in Figure 5.2. We have marked two branches, $b1$ and $b2$ with mutations. Below each leaf, we have marked the expected number of case and control haplotypes that are mapped to it, so that, for example, the first branch on the left has 0.5 case haplotypes and 0.8 control haplotypes mapped to it. Suppose that the mutation on $b1$ creates the A_{b1} allele at the putative disease SNP, then the expected number of case and control haplotypes carrying the A_{b1} allele are $n_{A_{b1}} = 0.5 + 1.1 + 0.4 = 2.0$ and $m_{A_{b1}} = 0.8 + 0.2 + 1.1 = 2.1$, respectively. The expected number of case and control haplotypes carrying the other possible allele at the putative disease SNP, a_{b1}, are $n_{a_{b1}} = 2.0$ and $m_{a_{b1}} = 1.9$, respectively. Likewise, if the mutation on $b2$ creates the A_{b2} allele, then the expected number of case and control haplotypes carrying the A_{b2} allele are $n_{A_{b2}} = 0.4$ and $m_{A_{b2}} = 1.1$, respectively, and carrying the a_{b2} allele are $n_{a_{b2}} = 0.5 + 1.1 + 2.0 = 3.6$ and $m_{a_{b2}} = 0.8 + 0.2 + 1.9 = 2.9$, respectively.

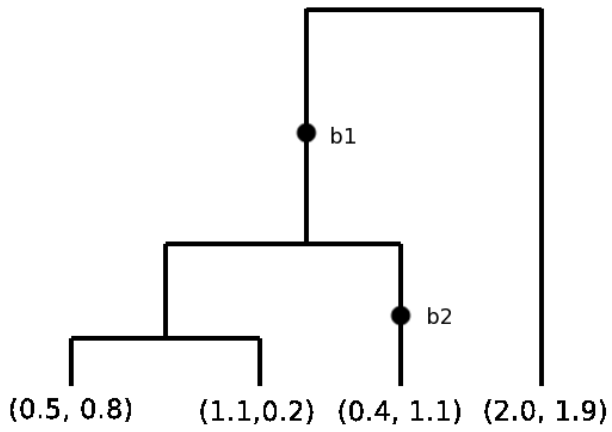


Figure 5.2: An illustration of model A. Suppose the tree above is the estimated genealogical tree for the panel and below each leaf we have marked the expected number of case and control haplotypes that are mapped to it, so that, for example, the first branch to the left has 0.5 case haplotypes and 0.8 control haplotypes mapped to it. Suppose that the mutation on $b1$ creates the A_{b1} allele at the putative disease locus, then the expected number of case and control haplotypes carrying the A_{b1} allele are $n_{A_{b1}} = 0.5 + 1.1 + 0.4 = 2.0$ and $m_{A_{b1}} = 0.8 + 0.2 + 1.1 = 2.1$, respectively. Likewise, if the mutation on $b2$ creates the A_{b2} allele, then the expected number of case and control haplotypes carrying the A_{b2} allele are $n_{A_{b2}} = 0.4$ and $m_{A_{b2}} = 1.1$, respectively.

Under our model of association γ_{A_b} and γ_{a_b} are independent with prior distribution $\beta(p_1, p_2)$ and the posterior probability under this model is given by

$$\begin{aligned}
\Pr(\mathbf{\Phi}|M_1, x, T_x, \bar{\Lambda}_x, b) &= \int \Pr(\mathbf{\Phi}|T_x, \bar{\Lambda}_x, b, \gamma)\Pr(\gamma)\mathrm{d}\gamma \\
&= \prod_{i\in\{a_b, A_b\}} \int_0^1 \gamma_i^{n_i}(1-\gamma_i)^{m_i}\Pr(\gamma_i)\mathrm{d}\gamma_i \\
&= (\beta(p_1, p_2))^{-2} \prod_{i\in\{a_b, A_b\}} \beta(p_1 + n_i, p_2 + m_i), \quad (5.20)
\end{aligned}$$

where $n_{a_b}, n_{A_b}, m_{a_b}$ and m_{A_b} are defined in Table 5.1.

Under the null model we assume that the allele at d has no impact on phenotype, and there is a single penetrance parameter for all haplotypes with the same prior

distribution $\beta(p_1, p_2)$:

$$\begin{aligned} \Pr(\mathbf{\Phi}|M_0, x, T_x, \bar{\Lambda}_x, b) &= \int \Pr(\mathbf{\Phi}|\gamma_{M_0})\Pr(\gamma_{M_0})\mathrm{d}\gamma_{M_0} \\ &= (\beta(p_1, p_2))^{-1}\beta(p_1 + n, p_2 + m). \end{aligned} \tag{5.21}$$

A discussion on how we should choose p_1 and p_2, and the interpretation of each choice on the relative risk of A_b and a_b is provided in Section 4.5.5. We will be assessing the sensitivity of our method to the penetrance prior in our simulation study later in this chapter.

Combining Equations (5.17), (5.20) and (5.21) we get our Bayes Factor

$$BF(x) = \sum_{b \in \mathbf{B}} BF(x, b)\Pr(b|M_1, x, T_x, \bar{\Lambda}_x), \tag{5.22}$$

where $BF(x, b) = \frac{\Pr(\mathbf{\Phi}|M_1, x, T_x, \bar{\Lambda}_x, b)}{\Pr(\mathbf{\Phi}|M_0, x, T_x, \bar{\Lambda}_x)}$ is the Bayes Factor assessing our model of association against the null model conditional on the disease mutation occurring on branch b:

$$BF(x, b) = \frac{\prod_{i \in \{a_b, A_b\}} \beta(p_1 + n_i, p_2 + m_i)}{(\beta(p_1, p_2))\beta(p_1 + n, p_2 + m)}. \tag{5.23}$$

Let us now consider the prior on which branch, b, carries the putative disease mutation, $\Pr(b|M_1, x, T_x, \bar{\Lambda}_x)$. We will first introduce some more notations; let

- $E = \{e_1, \ldots, e_P\}$ be the set of epochs in T_x between each successive coalescence event, where e_i is the epoch between the $(i-1)$th and ith coalescence events,
- $E_b \subseteq E$ be the set of epochs spanned by branch b,

- τ_{e_i} be the time (scaled by $4N_e$ generations) spanned by e_i, so τ_{e_i} is independently and exponentially distributed with mean $\binom{P-i+1}{2}^{-1}$ under the coalescent model (Kingman, 1982),

- τ_{E_b} be the time spanned by branch b in T_x, i.e. $\tau_{E_b} = \sum_{e_i \in E_b} \tau_{e_i}$, which has mean $\sum_{e_i \in E_b} \mathbb{E}\tau_{e_i}$.

If we assume a constant mutation rate throughout T_x, then the probability, under the coalescent model, that a single mutation lies on a branch, b, conditional on a total of 1 mutation in the tree is proportional to τ_{E_b}, i.e.

$$\Pr(b|M_1, x, T_x, \bar{\Lambda}_x) = \int \Pr(b|\tau_{E_b}, M_1, x, T_x, \bar{\Lambda}_x)\Pr(\tau_{E_b}|M_1, x, T_x, \bar{\Lambda}_x)\mathrm{d}\tau_{E_b} \quad (5.24)$$

To simplify our calculations we make the assumption that

$$\Pr(\tau_{E_b} = \mathbb{E}(\tau_{E_b})|M_1, x, T_x, \bar{\Lambda}_x) = 1. \quad (5.25)$$

This assumption ignores the variance of the branch lengths in T_x and our hope is that the strength of the data, expressed in $BF(x, b)$, is strong enough to overcome this approximation. Thus, to summarise,

$$\Pr(b|M_1, x, T_x, \bar{\Lambda}_x) \propto \sum_{e_i \in E_b} \mathbb{E}\tau_{e_i}, \quad (5.26)$$

where

$$\begin{aligned} \mathbb{E}\tau_{e_i} &= \frac{1.0}{\binom{P-i+1}{2}} \\ &= \frac{2.0}{(P-i+1)(P-i)}, \end{aligned} \quad (5.27)$$

Now, we have just described how to perform inference under a 1-mutation model, which involves a single putative disease mutation. However, our framework can be extended more generally to a k-mutation model. By allowing for multiple mutations we hope that we retain power when there is allelic heterogeneity at x.

To summarise the k-mutation model we need to introduce the following notations; let

- $\mathscr{B} \subset B^{(k)}$ be the set of all sets of k distinct branches in T_x;
- $\mathbf{b} \in \mathscr{B}$ be a set of k distinct branches with a mutation;
- $\mathbf{d} = (d_1, \ldots, d_k)$ be the k hypothetical disease SNPs defined by $\mathbf{b}$;
- $\mathscr{A}^{\mathbf{b}}$ be the set of $k+1$ possible haplotypes, comprising of the alleles at sites $\mathbf{d}$ that are created by the mutations on branches $\mathbf{b}$;
- $\mathbf{H}^{x,\mathbf{b}}$ be the set of panel haplotypes augmented with the putative disease SNPs, $\mathbf{d}$, so that $H^{x,\mathbf{b}}_{i,\mathbf{d}} \in \mathscr{A}^{\mathbf{b}}$ is the disease haplotype carried by H_i at sites $\mathbf{d}$; and
- $m_{\mathbf{a_b}}$, $n_{\mathbf{a_b}}$ be the expected number of control and case haplotypes of $\mathbf{G}$ carrying the disease haplotype $\mathbf{a_b} \in \mathscr{A}^{\mathbf{b}}$.

In the k-mutation model, k mutations are placed on distinct branches of T_x, which define k putative disease SNPs. The data can be summarised a 2 by $(k+1)$ table similar to Table 5.1, which contains the elements $m_{\mathbf{a_b}}$ and $n_{\mathbf{a_b}}$ for each of the $k+1$

disease haplotypes defined by $\mathbf{b}$ and are defined by

$$m_{\mathbf{a_b}} = \sum_{i=1}^{P} m_{i,x} \mathrm{I}(H_{i,\mathbf{d}}^{x,\mathbf{b}} = \mathbf{a_b}), \tag{5.28}$$

$$n_{\mathbf{a_b}} = \sum_{i=1}^{P} n_{i,x} \mathrm{I}(H_{i,\mathbf{d}}^{x,\mathbf{b}} = \mathbf{a_b}), \tag{5.29}$$

where $m_{i,x}$ and $n_{i,x}$ are defined in Equations (5.13) and (5.14). For the example described earlier, illustrated in Figure 5.2, there are two mutations on branches $\mathbf{b} = (b1, b2)$, which create the alleles A_{b1} and A_{b2} at the putative disease SNPs $\mathbf{d} = (d_1, d_2)$. There are 3 possible haplotypes comprised of the alleles at $\mathbf{d}$, $\mathscr{A}^{\mathbf{b}} = \{A_{b1}A_{b2}, A_{b1}a_{b_2}, a_{b1}a_{b_2}\}$. The expected number of cases carrying these haplotypes are $n_{A_{b1}A_{b2}} = 0.4$, $n_{A_{b1}a_{b2}} = 0.5 + 1.1 = 1.6$ and $n_{a_{b1}a_{b2}} = 2.0$, respectively. The expected number of controls carrying the haplotypes are $m_{A_{b1}A_{b2}} = 1.1$, $m_{A_{b1}a_{b2}} = 0.8 + 0.2 = 1.0$ and $m_{a_{b1}a_{b2}} = 1.9$, respectively.

The k-mutation model of association involves $k+1$ independent penetrance parameters, one for each disease haplotype in $\mathscr{A}^{\mathbf{b}}$, but the null model of no association remains unchanged. The Bayes Factor for the k-mutation model is therefore

$$BF(x) = \sum_{\mathbf{b} \in \mathscr{B}} BF(x, \mathbf{b}) \Pr(\mathbf{b} | M_1, x, T_x, \bar{\Lambda}_x), \tag{5.30}$$

where

$$BF(x, \mathbf{b}) = \frac{\prod_{i \in \mathscr{A}^{\mathbf{b}}} \beta(p_1 + n_i, p_2 + m_i)}{(\beta(p_1, p_2))^k \beta(p_1 + n, p_2 + m)} \tag{5.31}$$

For the prior on $\mathbf{b}$ we assume that the mutations on each branch $b \in \mathbf{b}$ occur

independently, so

$$\Pr(\mathbf{b}|M_1, x, T_x, \bar{\Lambda}_x) \propto \prod_{b \in \mathbf{b}} \sum_{e_i \in E_b} \mathbb{E}\tau_{e_i}. \tag{5.32}$$

In the simulation study described in Section 5.2 we will examine the power of using the 1-mutation model and the 2-mutation model. However, in Chapter 6, we have also implemented the 3-mutation model.

Model B

In Section 4.5 we described how, given a haplotype tree, a set of haplotypes can be clustered to create basis clusters and from them groupings of haplotypes that can be tested for association. We can apply the same approach here. The main difference is that the tree at each level defines a partition of the expected haplotype counts, mapped to the leaves of the tree, rather than discrete numbers of haplotypes.

Let

- k be the number of basis clusters, which implies the level of the tree that we cut to construct a set of basis clusters; for details of how the basis clusters are constructed for a given tree, T_x, and k please refer to Section 4.5.2;
- $\mathbf{B}_{x,k} = \{B_1, \ldots, B_k\}$ be the set of k basis clusters constructed at SNP x, and $\mathbf{b} = \{b_1, \ldots, b_P\}$ be the membership vector for haplotypes in $\mathbf{H}$, so that $b_i = j$ if $H_i \in B_j$ for some $H_i \in \mathbf{H}$;
- $C_k = \{c_{k,1}, \ldots, c_{k,s}\}$ be a set of sets that partitions the set $\{1, \ldots, k\}$ into s non-empty subsets, with the properties $c_{k,i} \subset \{1, \ldots, k\}$, $c_{k,i} \neq \emptyset$ for all

$i = 1, \ldots, s$, and $\bigcup_i c_{k,i} = \{1, \ldots, k\}$, $\bigcap_i c_{k,i} = \emptyset$;

- D_k be the set of all possible distinct sets of C_k.

Given a set of basis clusters $B_{x,k}$, we use $C_k = \{c_{k,1}, \ldots, c_{k,s}\}$ to define a grouping, of s groups, for the haplotypes of $\mathbf{G}$. The members of the ith group is defined by the set $c_{k,i}$, which contains the cumulative expected number of case and control haplotypes clustered into the set of basis clusters $\{B_j : j \in c_{k,i}\}$. The expected number of control and case haplotypes in the ith group is given by

$$m_{c_{k,i}} = \sum_{l \in c_{k,i}} \sum_{j=1}^{P} m_{j,x} \mathrm{I}(b_j = l), \tag{5.33}$$

$$n_{c_{k,i}} = \sum_{l \in c_{k,i}} \sum_{j=1}^{P} n_{j,x} \mathrm{I}(b_j = l), \tag{5.34}$$

where $m_{j,x}$ and $n_{j,x}$ are defined in Equations (5.13) and (5.14).

Our model of association specifies a value for k and therefore a set of basis clusters $\mathbf{B}_{x,k}$ defined by the tree T_x. We average over all possible groupings that can be constructed from the basis clusters and model the phenotype information of a haplotype as conditionally independent given its group membership in the grouping C_k:

$$\begin{aligned} \Pr(\mathbf{\Phi}|M_1, x, T_x, \bar{\Lambda}_x) &= \Pr(\mathbf{\Phi}|\mathbf{B}_{x,k}, x, \bar{\Lambda}_x) \\ &= \sum_{C_k \in D_k} \Pr(\mathbf{\Phi}|C_k, \mathbf{B}_{x,k}, x, \bar{\Lambda}_x) \Pr(C_k|\mathbf{B}_{x,k}, x, \bar{\Lambda}_x) \\ &= \sum_{C_k \in D_k} \Pr(\mathbf{\Phi}|C_k) \Pr(C_k|\mathbf{B}_{x,k}, x, \bar{\Lambda}_x). \end{aligned} \tag{5.35}$$

Our approach here is to summarise the topological information of the tree below

level k (the level where there are k distinct branches) into k groups, i.e. the basis clusters, which will be good if the disease mutation occurs in the upper levels of the tree. In return for this simplification, we are able to average over all the possible topologies of T_x above level k, which will account for some of the uncertainty in T_x and would be particularly effective if the variance of T_x is large.

The penetrance parameters of each grouping are independently distributed with prior $\beta(p_1, p_2)$, so

$$\begin{aligned} \Pr(\mathbf{\Phi}|C_k) &= \prod_{c \in C_k} \int_0^1 \gamma_i^{n_c} (1-\gamma_i)^{m_c} \Pr(\gamma_i) \mathrm{d}\gamma_i \\ &= \beta(p_1, p_2)^s \prod_{c \in C_k} \beta(n_c + p_1, m_c + p_2), \end{aligned} \tag{5.36}$$

where m_c and n_c for each $c \in C_k$ are given in Equations (5.33) and (5.34)

The null model remains unchanged from Model A, so is the same as in Equation (5.21). Our Bayes Factor therefore takes the form

$$BF(x) = \sum_{C_k \in D_k} BF(x, C_k) \Pr(C_k | \mathbf{B}_{x,k}, x, \bar{\Lambda}_x), \tag{5.37}$$

where $BF(x, C_k)$ is the Bayes Factor that assesses the model of association conditional on grouping C_k against the null model of no association:

$$BF(x, C_k) = \frac{\prod_{c \in C_k} \beta(n_c + p_1, m_c + p_2)}{(\beta(p_1, p_2))^{s-1} \beta(p_1 + n, p_2 + m)}, \tag{5.38}$$

where s is the number of groups in C_k.

We now need to assign a distribution on C_k, $\Pr(C_k|\mathbf{B}_{x,k}, x, \bar{\Lambda}_x)$. This is a difficult question since it is unclear what is a likely grouping and what is not. Earlier

attempts are described in Section 4.5 but our simulations showed that our method (HAPCLUSTER) is robust to its choice.

Here we will specify a different prior and test the sensitivity of our method by simulation in the next section. We condition on the size of a grouping, s, and assign a truncated geometric distribution as its prior. In addition, we assume that all groupings of the same size are equally likely, i.e. $\Pr(C_k|\mathbf{B}_{x,k}, x, \bar{\Lambda}_x, s) \propto 1$. Thus,

$$\begin{aligned}\Pr(C_k|\mathbf{B}_{x,k}, x, \bar{\Lambda}_x) &= \sum_{s'} \Pr(C_k|\mathbf{B}_{x,k}, x, \bar{\Lambda}_x, s')p(s') \\ &= \Pr(C_k|\mathbf{B}_{x,k}, x, \bar{\Lambda}_x, s)p(s) \\ &\propto (1-p)^{s-1}p, \end{aligned} \tag{5.39}$$

for some parameter p and $s = 2, \ldots, k$. As p decreases, more prior mass is placed on groupings of smaller size. Our model of association conditional on grouping C_k of size s is similar to the $(s-1)$-mutation model, with each group carrying a particular set of disease alleles induced by $s-1$ mutations. Specifying more mass on small s is therefore sensible if the mutation rate is low.

A final question is what value of k we should use, i.e. where should we cut the tree? If we cut the tree too high, i.e. k too small, we lose too much information in the lower topology of the tree and will be underpowered if the disease mutations occurred in the lower part of the tree. However, cutting the tree too low, i.e. k too large, will result in an over-complex model with an excessive parameter space. This will result in a loss of power if the data can be modelled reasonably well using a few basis clusters, since the prior is more concentrated on the observed data (Denison and Holmes, 2005). The simulation study from the last chapter involving HAPCLUSTER suggests that 2-3 basis clusters are underpowered but did not find

any obvious loss of power when using up to 10 basis clusters. We will repeat this comparison, with 2-10 basis clusters, for GENECLUSTER.

5.2 Simulation Study

To assess the approaches described in this chapter we performed two simulation studies. Both used data sets of 2000 control and 2000 case individuals (8000 haplotypes in total) generated from HAPGEN. The first simulation study is designed to assess the power to detect a single mutation and the second simulation study is designed to assess power under allelic heterogeneity.

Data sets were generated conditional on the 120 CEU parental HapMap haplotypes in 5 ENCODE regions (ENr123, ENr213, ENr232, ENr321 and ENm013) as reference panels; each data set is approximately 500Kb in length and contains about 1000 SNPs, approximately 100 of which are on the Affymetrix 500K chip. We used the fine-scale recombination rates and effective population size, $N_e = 11418$, estimated from the HapMap project.

As before, we used the pseudo HapMap panels from Marchini et al. (2007) for use in the various multipoint approaches. The ENCODE regions have a higher SNP density in HapMap than the rest of the genome, and this thinning is required to allow extrapolation to the remainder of the genome. We analysed each data set using the following 3 methods.

1 GENECLUSTER - We conditioned on the 120 CEU HapMap parental haplotypes as our reference panel and estimated their genealogy, based on the

data at the pseudo HapMap SNPs, at positions every 5Kb apart and tested for association. In order to assess power under a realistic setting, only case-control genotype data at the SNPs on the Affymetrix 500K chip was presented to GENECLUSTER. We analysed for association using the 1-mutation and 2-mutation models from Model A and 2-10 basis clusters from Model B.

2 Single SNP - The single SNP approach performed a single SNP Bayesian association test (under a haploid likelihood) at each Affymetrix SNP. Details of this test is given in Section 2.1.2.

3 Multimarker Predictor (MMP) - We used the same set of MMPs as Marchini et al. (2007), which we also used for the simulation study in Chapter 4. For details on how these MMPs were found please refer back to Section 4.6.4. The MMP approach performed a single SNP Bayesian association test at all Affymetrix SNPs and predicted SNPs.

We used a common beta penetrance prior for the single SNP test and GENECLUSTER to ensure results are comparable.

5.2.1 Calculating power

In the WTCCC study (The Wellcome Trust Case Control Consortium, 2007) all regions where a test statistic exceeded a genome-wide significance threshold were reported and put forward for further investigation. In this simulation study we attempt to find out, empirically, the power of each method to report a region with a causal variant(s) as significant.

To evaluate each method we take the maximum Bayes Factor of each analysis as a

summary test statistic of each data set. We compare the power of each method to generate a summary test statistic above a Bayes Factor threshold.

The genome-wide Bayes Factor significance threshold is up to the investigator's discretion and the assumptions (s)he makes. Here we propose 10^4 as a sensible threshold. We have

$$\text{Posterior odds} = \text{Prior odds} \times \text{Bayes Factor}.$$

We assume that the prior odds of true association at any specified locus is of the order 10,000:1 against, for example, on the basis of 1,000,000 "independent" regions of the genome and an expectation of 100 detectable genes involved in the condition. Then if the Bayes Factors for a locus is more than 10^4, the posterior odds in favour of the signal being a true association is at least 1:1. In addition we observe that all loci identified by the WTCCC study (The Wellcome Trust Case Control Consortium, 2007) with a strong signal of association ("tier" 1 signals) have a single SNP Bayes Factor exceeding 10^4, with the lowest Bayes Factor being 4.15. Therefore, 10^4 appear to be a reasonable threshold for genome-wide Bayesian association tests.

A genome-wide association study can also identify regions, for further investigation, by selecting the top n regions ranked by signal strength. In this scenario, the appropriate method to assess power is to consider the Receiver Operating Characteristics (ROC) (de Bakker et al., 2005; Marchini et al., 2007), as we did in the previous chapter, which takes into account of the null distribution for the summary test statistics.

We generated a set of "null" data sets to find a empirical null distribution of the summary test statistics. To do this we took each SNP in the ENCODE marker set

and generated a data set with it as the disease SNP but with a relative risk of 1.0; we generated 4380 data sets this way. We repeated the 3 methods on the null data sets to generate a sample of the summary test statistics under the null.

5.2.2 Computation cost

On an Intel Pentium 4 2.8 GHz desktop, it took 565 seconds to calculate the expected number of haplotypes mapped to each leaf of the trees (step 2 in Section 5.1.3) at every 5Kb across the region, and required 170MB of RAM. To complete the analysis of each data set further running time is required to test for association using the various models we introduced (step 3 in Section 5.1.4). The running cost for this procedure is listed in Table 5.2.

Model	Running time (seconds)
2 Basis Clusters	1
10 Basis Clusters	47
1-Mutation	2
2-Mutation	104
All	173

Table 5.2: The running times of GENECLUSTER models on a single Intel Pentium 4 2.8 GHz desktop for the analysis of 4000 individuals genotyped at approximately 100 SNPs. All methods include the 1-mutation model, 2-mutation model and 2-10 basis clusters models.

The computation cost of a n-mutation model is proportional to $\binom{2P-2}{n}$, the number of distinct sets of n mutations in the genealogical tree for a panel of P haplotypes. The additional computation cost to account for 3 or more mutations was too great for it to be included in the simulation study, although we were able to implement the 3-mutation model over small regions when we analysed the WTCCC data in the next Chapter.

5.2.3 Detecting association in a single mutation disease model

In the first of our simulation studies we generated data sets for every minor allele, at every non-monomorphic SNP, in the 5 ENCODE regions as the single disease causing variant. For every variant we generated a data set with heterozygote relative risks 1.0, 1.5, 2.0 and 2.5 under a log additive model for relative risk. This resulted in 4380 data sets for each effect size. Further we categorised each disease SNP as "tagged" if it is tagged by a MMP or an Affymetrix SNP with $r^2 > 0.8$, and "untagged" otherwise, and "common" if its minor allele frequency (MAF) is greater than 5%, and "rare" otherwise. Under this classification, for each effect size, there are 749 data sets with a rare and untagged variant, 426 data sets with a rare and tagged variant, 1007 data sets with a common and untagged variant and 2201 data sets with a common and tagged variant.

Before we compare the power of the 3 methods we implemented, we will take a moment to compare the power of the GENECLUSTER method under different parameters. The aim is to assess the sensitivity of our methods to these parameters.

Sensitivity to the estimated tree

Here we will briefly compare the two approaches to estimating the tree for the panel haplotypes mentioned in Section 5.1.2, LSTREE and TREESIM. Figure 5.3 shows the inverse cumulative density function of GENECLUSTER Bayes Factors using the 2-mutation model and a penetrance prior of $\beta(15.0, 15.0)$, applied to data sets with a relative risk of 2.0. The x-axis in Figure 5.3 is the $\log_{10}$ Bayes Factor and on the y-axis is the inverse cumulative density of the maximum Bayes Factor from the data sets, which we define as the proportion of data sets with a maximum Bayes Factor

exceeding the threshold specified on the x-axis. The inverse cumulative density is therefore the power to detect a disease locus for a given Bayes Factor threshold and our figures provide a natural illustration of how power decreases with increasing Bayes Factor threshold. The figure is split up according to the properties of the disease SNP.

Figure 5.3 shows that the trees produced from both methods output similar Bayes Factors. We also observe similar comparisons using the 1-mutation model and 2-10 basis clusters, on data sets with different relative risks and when we compare their Receiver Operating Characteristics (ROC). These results suggest that the topologies of the two sets of trees are similar for association testing.

From now on, for consistency, we will use TREESIM trees for our GENECLUSTER analyses.

Sensitivity to the number of basis clusters and prior on grouping size

Figure 5.4 shows the inverse cumulative density function of GENECLUSTER Bayes Factors using 2-10 basis clusters. The data sets contain a single causative mutation of relative risk 2.0. We have used a uniform prior on grouping size and a penetrance prior of $\beta(15.0, 15.0)$.

Figure 5.4 and Table 5.3 show that the maximum Bayes Factor increases with the number of basis clusters for less than 5 basis clusters. Cutting the tree at too high a level, i.e. for k too small, therefore results in loss of power as explained in Section 5.1.4. The observation that more than 5 basis clusters does not gain extra power can be due to two reasons. The first reason is that we are losing power because we

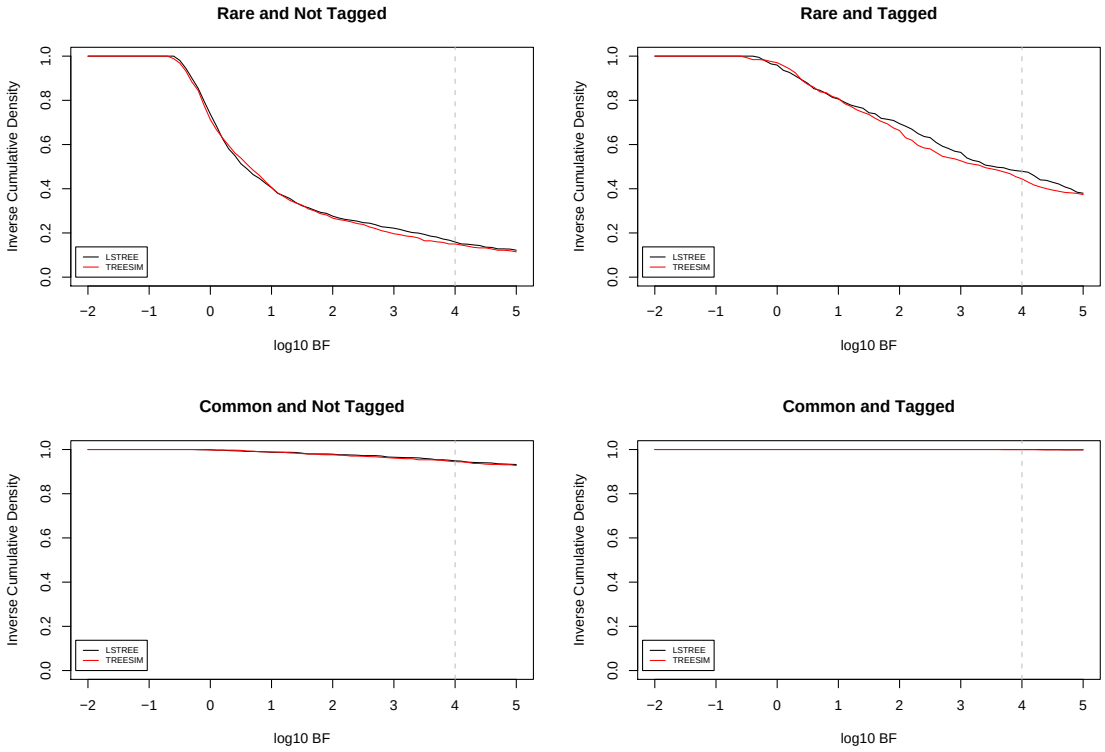


Figure 5.3: The empirical inverse cumulative density function of the maximum Bayes Factor using LSTREE and TREESIM trees. We have used the 2-mutation model from Model A to analyse data sets with a single causal variant of relative risk 2.0. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

mis-specified our prior on grouping size and low probability mass is placed on the more likely grouping(s) in the integration process. The second possible reason is that summarising the tree, T_x, using 5 basis clusters is sufficient to capture most of the signal in our simulations.

Figure 5.5 shows the Bayes Factor distribution using 5 basis clusters and geometric prior distributions with varying parameters on the grouping size. $Geo(1.0)$ is the uniform prior and the smaller the geometric distribution parameter the more prob-

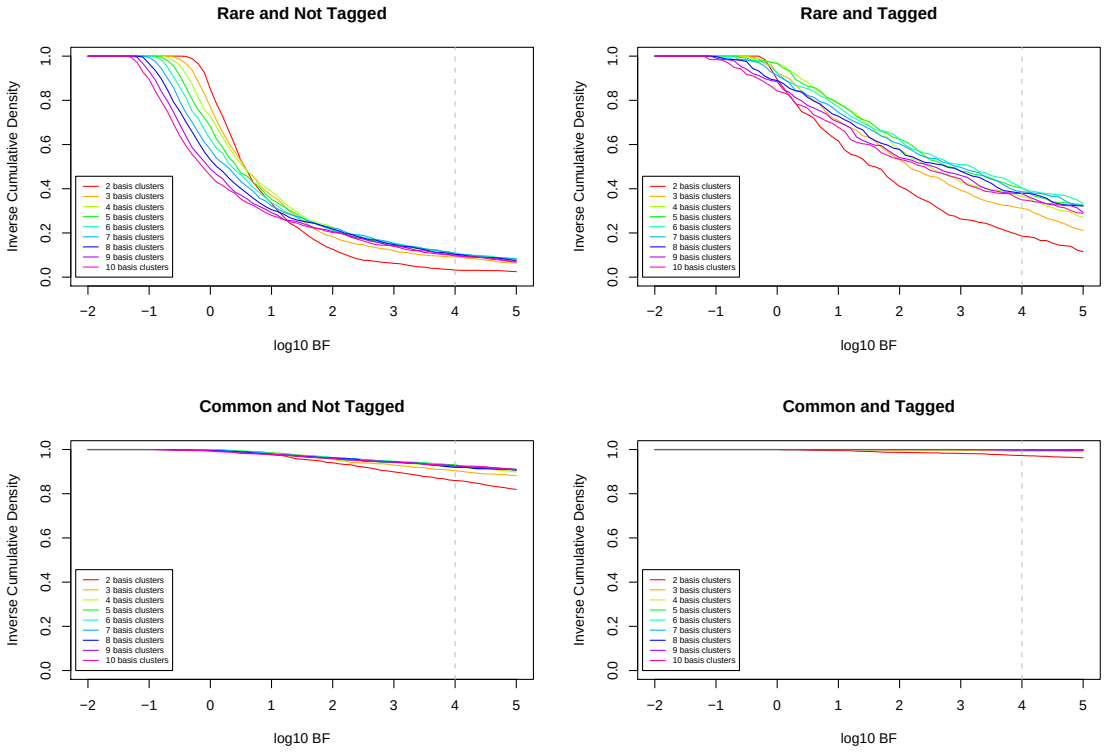


Figure 5.4: The empirical inverse cumulative density function of the maximum Bayes Factor using 2-10 basis clusters from Model B. We have analysed data sets with a single causal variant of relative risk 2.0. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

ability mass is placed on groupings that contain fewer groups. We obtain similar results when we use other number of basis clusters applied to data sets with different effect sizes and when we compared the ROC. These results suggest that the prior on grouping size has little effect on the Bayes Factors and our method is insensitive to its specification.

From now on, for consistency, we will use a prior of $Geo(0.25)$ for grouping size.

Basis clusters	Rare and untagged	Rare and tagged	Common and untagged	Common and tagged
2	0.03	0.19	0.86	0.97
3	0.09	0.31	0.90	0.99
4	0.11	0.37	0.92	1.00
5	0.11	0.40	0.93	1.00
6	0.11	0.40	0.93	1.00
7	0.11	0.39	0.92	1.00
8	0.10	0.38	0.92	1.00
9	0.10	0.38	0.93	1.00
10	0.10	0.35	0.93	1.00

Table 5.3: Power of GENECLUSTER using models from Model B with a Bayes Factor threshold of 10^4. We have analysed data sets containing a single causal variant of relative risk 2.0. The columns refer to the type of disease SNP and power is measured in terms of the frequency of a Bayes Factor exceeding the threshold. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from.

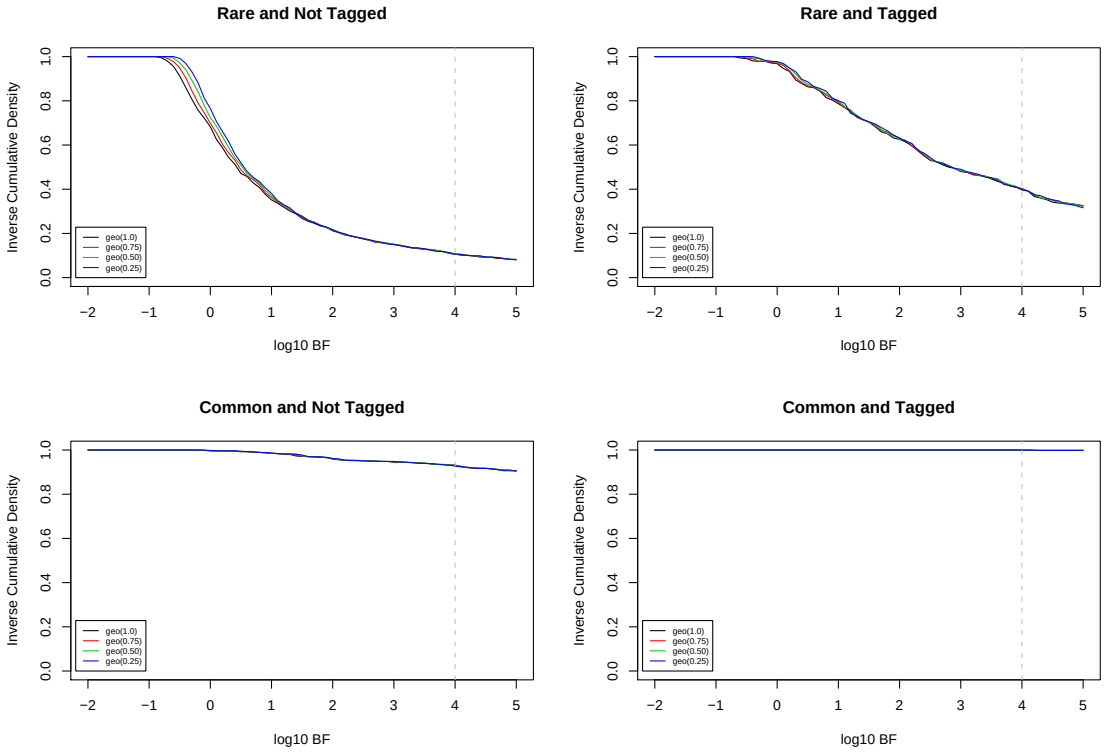


Figure 5.5: The empirical inverse cumulative density function of the maximum Bayes Factor using a set of $Geo(p)$ priors on grouping size. We have used 5 basis clusters from Model B to analyse data sets with a single causal variant of relative risk 2.0. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

Sensitivity to penetrance priors

Under our model of association, we have partitioned the haplotypes of the case-control sample into "penetrance" groups and these groups have independent penetrance parameters with a beta prior distribution. Here we assess the sensitivity of our method to the penetrance prior, by comparing our results with the penetrance prior $\beta(p, p)$ for $p = 1.0, 5.0, 10.0$ and 15.0. Figure 4.3 from the last chapter illustrates the haplotype relative risk between each penetrance group assumed by each prior.

Figure 5.6 compares the distribution of GENECLUSTER Bayes Factor using different penetrance priors on data sets with a causal variant of relative risk 1.5. We have used the 2-mutation model with TREESIM trees to illustrate our result but we see the same trend with the other models of association from models A and B. Results here are consistent with those in Section 4.6.3. The causal variants in our data sets have a relatively small effect, which will be further reduced at the typed SNPs if they are not strongly correlated with the disease SNP. Therefore, priors that place a large mass on large risks are inappropriate as the null model would be favoured over the alternative, because the data would be somewhat consistent with zero effects and completely inconsistent with large effects. Therefore, $\beta(15.0, 15.0)$, which places the most mass on small effects sizes results in the largest Bayes Factors. From now on, for consistency, we will use a penetrance prior of $\beta(15.0, 15.0)$ for all methods.

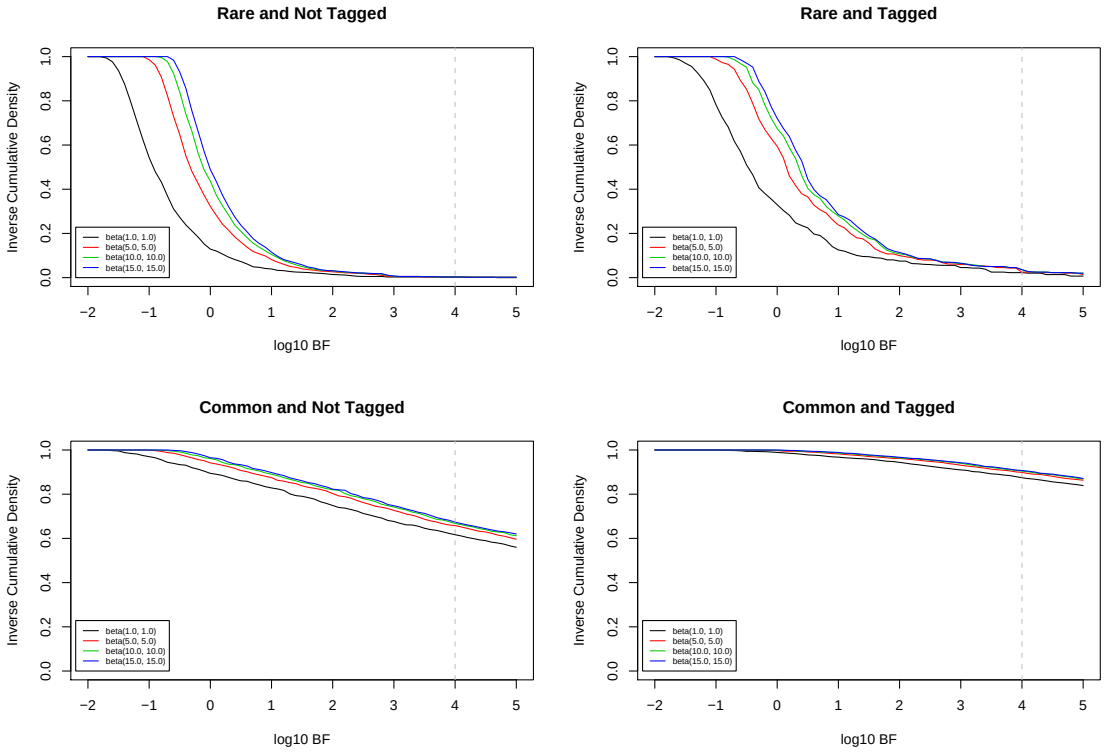


Figure 5.6: The empirical inverse cumulative density function of the maximum Bayes Factor using a set of $\beta(p, p)$ penetrance priors. We have used the 2-mutation model from Model A to analyse data sets with a single causal variant of relative risk 1.5. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

Comparison with SNP based methods

Figure 5.7 shows the distribution of GENECLUSTER Bayes Factors using the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B applied to data sets with a single causal variant of relative risk 2.0. The figure compares the Bayes Factor distributions to the Single SNP and the MMP methods. Figures B.1 - B.2 in Appendix B show the corresponding figures for data sets with a causal variant of relative risk 1.5 and 2.5.

We stratify our results in the same way as in Section 4.6.5 and we observe very similar results. For rare alleles, in particular those that are untagged, we observe very little power when the relative risks are low. This reflects the difficulty in detecting such variants in population data. As expected, it is easier to detect tagged variants and common variants and power increases with relative risk.

The Single SNP and MMP methods perform best when the disease allele is common and tagged and is consistent with Chapman et al. (2003), who reported that single SNP approaches are most powerful for detecting such variants (although those results refer specifically to Frequentist tests). However, for untagged disease alleles, in particular those that are rare, GENECLUSTER provides a large boost in power. The same observations were reported by Marchini et al. (2007) for IMPUTE, which uses the same LD model as GENECLUSTER. Rare and untagged variants are harder to tag using only a few surrogate SNPs but GENECLUSTER (and IMPUTE) uses data at all SNPs, modulated by the local recombination rate, to infer local haplotype structure and is therefore able to detect rare and untagged variants that lie on extended haplotype backgrounds.

The power of each method to detect untagged and rare variants, with a Bayes Factor

threshold of 10^4 is summarised in Table 5.4.

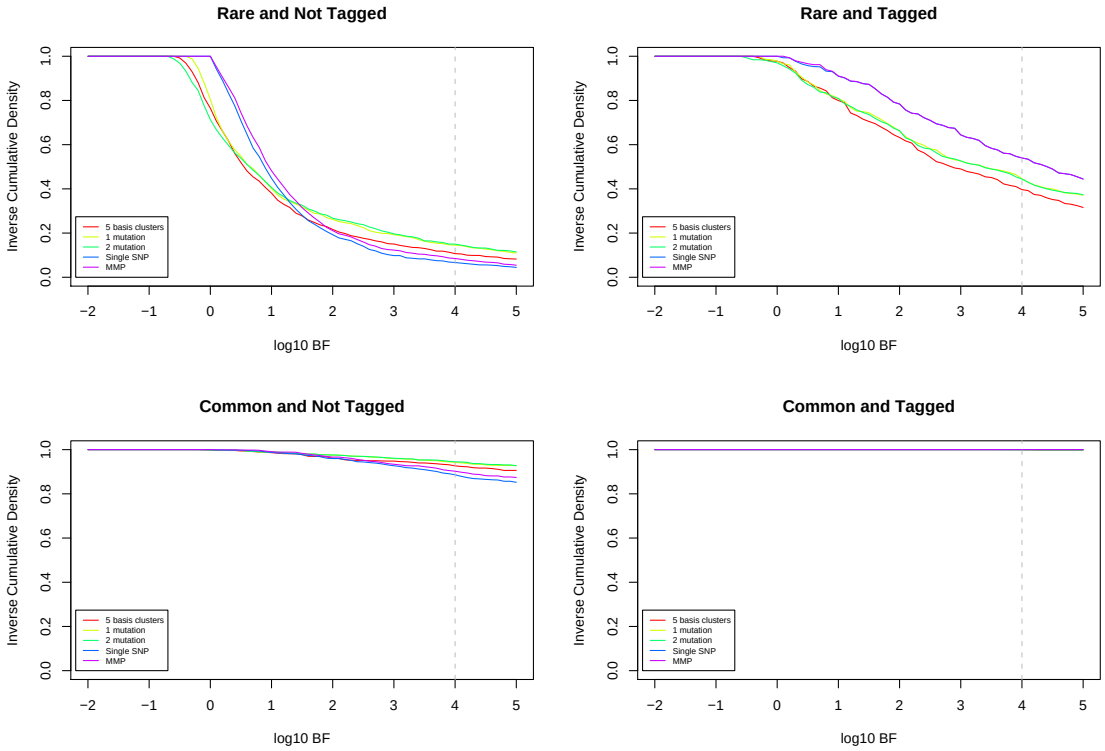


Figure 5.7: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with a single causal variant of relative risk 2.0. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

The models of association used by GENECLUSTER is more complex than those used by SNP based methods. The Bayesian framework contains a natural penalty against over-complex models when the data can be similarly well supported by a simpler model (Denison and Holmes, 2005). For null data, our models of association will incur an increased penalty for complexity because the data can be well-supported

Method	RR = 1.5	RR = 2.0	RR = 2.5
5 Basis Clusters	0.00	0.11	0.25
1-Mutation Model	0.00	0.15	0.34
2-Mutation Model	0.00	0.15	0.35
Singe SNP	0.01	0.07	0.20
MMP	0.01	0.09	0.22

Table 5.4: Power of GENECLUSTER and SNP based approaches to detect a single untagged and rare causal variant with a Bayes Factor threshold of 10^4. The columns refer to the relative risk (RR) of the disease allele and the rows refer to the methods that we used to detect for association. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from.

by the simpler null model. This will result in a more conservative null distribution for GENECLUSTER, which is displayed in Figure 5.8. This boosts the power of GENECLUSTER relative to the Single SNP and MMP methods when we take the null into account and compare the ROC of each method. Figure 5.9 compares the ROC of each method when the causal variant has a relative risk of 2.0. Figures B.3 - B.4 in Appendix B compares the ROC of each method for relative risks of 1.5 and 2.5, respectively. As before, we find that GENECLUSTER provides the greatest boost in power when the disease allele is untagged and rare. At the 5% significance level it is up to 15% more powerful using models from Model A (Table 5.5).

Method	RR = 1.5	RR = 2.0	RR = 2.5
5 Basis Clusters	0.15	0.41	0.58
1-Mutation Model	0.17	0.47	0.64
2-Mutation Model	0.16	0.47	0.64
Single SNP	0.13	0.31	0.49
MMP	0.13	0.32	0.51

Table 5.5: The power of GENECLUSTER and SNP based approaches to detect a single untagged and rare causal variant at the 5% significance level. The columns refer to the relative risk (RR) of the disease allele and the rows refer to the methods that we used to detect for association. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from.

Comparing the 1-mutation and 2-mutation models from Model A with the 5 basis

clusters model from Model B, we find that Model A tend to be more powerful for rare variants. The basis cluster approach cuts the tree at a given level to construct the basis clusters, which means that most of the information on the topology below that level is ignored. Therefore, Model B can be underpowered to detect rare variants, which tend to be caused by mutations at low levels of the tree.

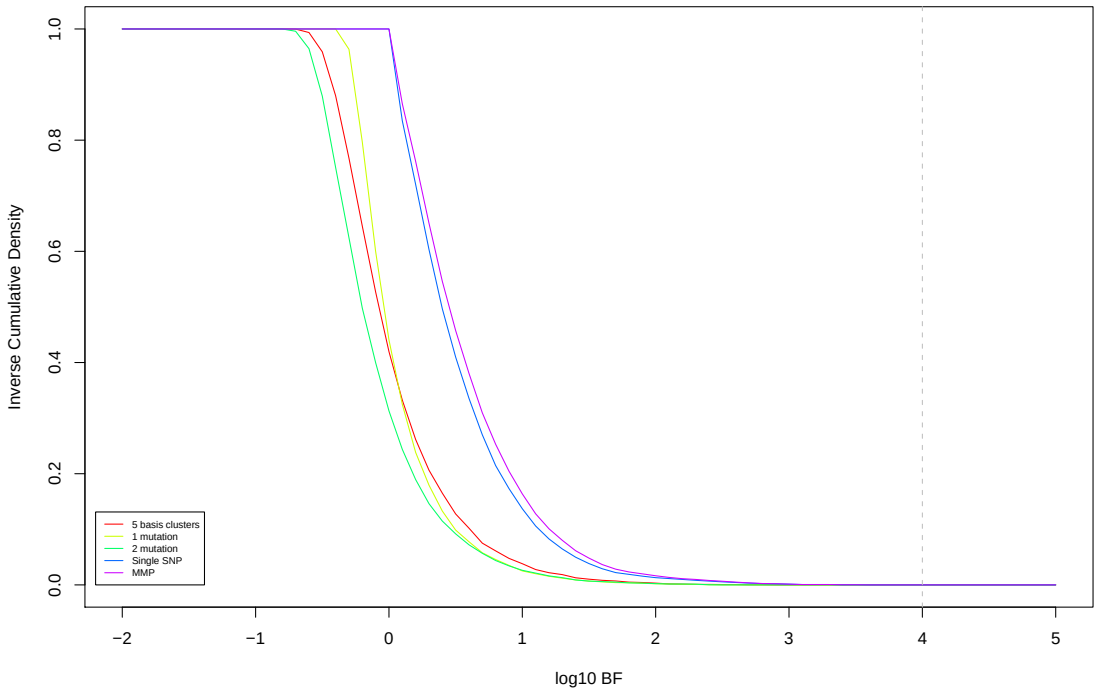


Figure 5.8: The empirical null inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with a single causal variant of relative risk 1.0, which means that the distributions given here are the empirical distributions under the null. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

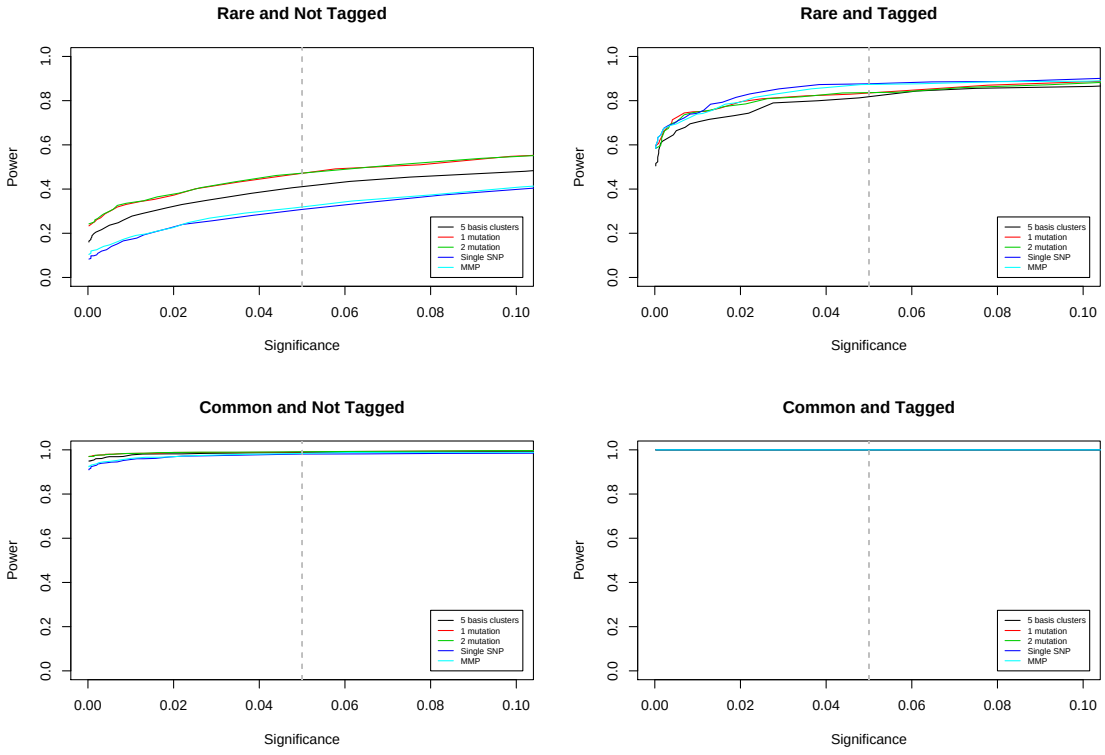


Figure 5.9: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with a single causal variant of relative risk 2.0. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity).The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 5% significance threshold.

5.2.4 Detecting association in a two mutation disease model

In our second simulation study we assess the power of each approach to detect association when there are two causal mutations. We used the version of HAPGEN described in Section 3.3 to create instances of allelic heterogeneity.

We have tried to simulate realistic instances of allelic heterogeneity but this is difficult since not many complex disease loci with allelic heterogeneity are currently known and characterised. There are two exceptions however: the *IL23R* (Duerr et al., 2006) and *NOD2* (Ogura, Y. et al., 2001; Hugot et al., 2001) loci for Crohn's disease, and we have simulated our data based on the observations made there. For more details on these two loci, please skip forward to Section 6.3.3.

As before, we generated data sets conditional on the HapMap haplotypes in the 5 ENCODE regions. Each data set contained two disease SNPs and we simulated under two disease models:

1 two disease alleles: one very rare (allele frequency less than 2%) and one at moderate frequency (allele frequency between 5% and 20%), which are at similar frequencies as two of the causal variants at the *NOD2* locus;

2 two disease alleles: both of moderate frequencies (allele frequencies between 5% and 20%), which are at similar frequencies as two coding variants at the *IL23R* locus, however one of these variants is protective whereas we have simulated both as deleterious.

For GENECLUSTER, we used the 1-mutation and the 2-mutation models from Model A and 5 basis clusters from Model B, with TREESIM trees and a $Geo(0.25)$

prior on grouping size. The penetrance prior for all methods was $\beta(15.0, 15.0)$.

Disease model 1

We took every rare SNP in the ENCODE marker sets and randomly paired it up with a common SNP. We repeated the process twice; once with a tagged common SNP and once with an untagged common SNP. Here, a rare SNP refers to one with a MAF less than 2%, a common SNP refers to one with a MAF between 5% and 20% and tagged means tagged with $r^2 > 0.8$ by a MMP or an Affymetrix SNP in the pseudo HapMap panel. If a rare SNP does not have a suitable common SNP within 15Kb to pair up, then it is discarded. We only generated data sets for each category with regions that have more than 50 suitable SNP pairs.

We stratify our results into 4 categories according to whether each disease SNP in a data set is tagged. For brevity we denote each category as below, and we have also indicated the number of data sets and regions generated for each category.

- **A1** untagged rare SNP and untagged common SNP (429 data sets, 5 regions)
- **B1** untagged rare SNP and tagged common SNP (321 data sets, 4 regions)
- **C1** tagged rare SNP and untagged common SNP (260 data sets, 2 regions)
- **D1** tagged rare SNP and tagged common SNP (176 data sets, 1 region)

We set the common disease allele to have a relative risk of 1.3, varied the relative risk for the rare allele from 1.0 to 2.5, and set the interaction parameter (γ) to 1.0, i.e. no epistasis (see Section 3.3 for an explanation of the risk parameters). Figures 5.10-5.13 compare the inverse cumulative distributions of the Bayes Factors of each

method for the different effect sizes of the rare variant. Figure 5.10 refers to the scenario where there is only one causal variant (since the rare allele has relative risk 1.0) and Figures 5.11-5.13 refer to the scenario where there are two causal variants in the region. It is clear that our results are subject to noise due to the modest number of data sets, however clear trends can be observed. The 2-mutation model of GENECLUSTER explicitly models for allelic heterogeneity and we would expect it to have the largest Bayes Factors in Figures 5.11-5.13. However, from the results it appears that this is not the case, so we will spend a little time explaining this in further detail.

The distribution of Bayes Factors in Figures 5.10 and 5.11 are very similar, which means that for rare variants with a relative risk of 1.5 the Bayes Factors are mainly influenced by the common variant. The Single SNP and MMP methods are powerful for detecting common variants and therefore they have the largest Bayes Factors.

As the relative risk of the rare variant increases, the data is increasingly influenced by two causal variants and therefore the 2-mutation model becomes more powerful relative to the other methods. When the relative risk of the rare variant is 2.5 and using a genome-wide threshold of 10^4, the 2-mutation model can be up to 6% more powerful than the Single SNP and MMP methods (Table 5.6). In addition, we note that the Bayes Factors of the 2-mutation model appear to be significantly larger than the 1-mutation model (compare Figures 5.10 and 5.13), which suggests that the differences in the Bayes Factors under the two models can be an indication of allelic heterogeneity. We will explore the power to detect the presence of allelic heterogeneity in the next section.

The basis clusters model (Model B) allows for multiple risk parameters under its

model of association and we therefore also expect it to be powerful under allelic heterogeneity. However, it appears to be underpowered relative to the 1-mutation and 2-mutation models in Table 5.6. As discussed before, the basis cluster model discards the topology in the lower levels of the tree, so it can be underpowered to detect the rare variant, which is likely to be caused by a mutation in the lower levels of the tree. Using more than 5 basis clusters might improve power, but we did not have time to perform this analysis.

Method	A1	B1	C1	D1
5 Basis Clusters	0.17	0.33	0.67	0.76
1-Mutation Model	0.19	0.33	0.69	0.73
2-Mutation Model	0.22	0.37	0.71	0.81
Single SNP	0.14	0.30	0.74	0.84
MMP	0.16	0.32	0.74	0.84

Table 5.6: Power of GENECLUSTER and SNP based approaches to detect association with two causal variants and a Bayes Factor threshold of 10^4. One disease allele is rare (MAF less than 2%) and the other is common (MAF between 5% and 20%) and they have relative risks of 2.5 and 1.3 respectively. The columns refer to the whether each disease SNP is tagged by an Affymetrix SNP or a MMP and the rows refer to the methods that we used to detect for association. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from.

Since GENECLUSTER is more conservative under the null (Figure 5.8), the comparisons with respect to the ROC show a greater boost in power from GENECLUSTER, as shown in Figure 5.14, which compares the ROC for each method when the rare and common variants have relative risks of 2.5 and 1.3, respectively. Figures B.5-B.6 in Appendix B compares the ROC when the rare variant has a relative risk of 1.5 and 2.0, respectively. When the rare variant has relative risk 2.5, GENECLUSTER using the 2-mutation model is up to 25% more powerful than the Single SNP and MMP methods at the 5% significance level (Table 5.7).

Method	A1	B1	C1	D1
5 Basis Clusters	0.68	0.86	0.97	0.99
1-Mutation Model	0.71	0.88	0.96	0.99
2-Mutation Model	0.73	0.89	0.96	0.99
Single SNP	0.54	0.84	0.97	0.98
MMP	0.58	0.85	0.97	0.98

Table 5.7: Power of GENECLUSTER and SNP based approaches to detect association with two causal variants and a significance threshold of 5%. One disease allele is rare (MAF less than 2%) and the other is common (MAF between 5% and 20%) and they have relative risks of 2.5 and 1.3 respectively. The columns refer to the whether each disease SNP is tagged by an Affymetrix SNP or a MMP and the rows refer to the methods that we used to detect for association. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from.

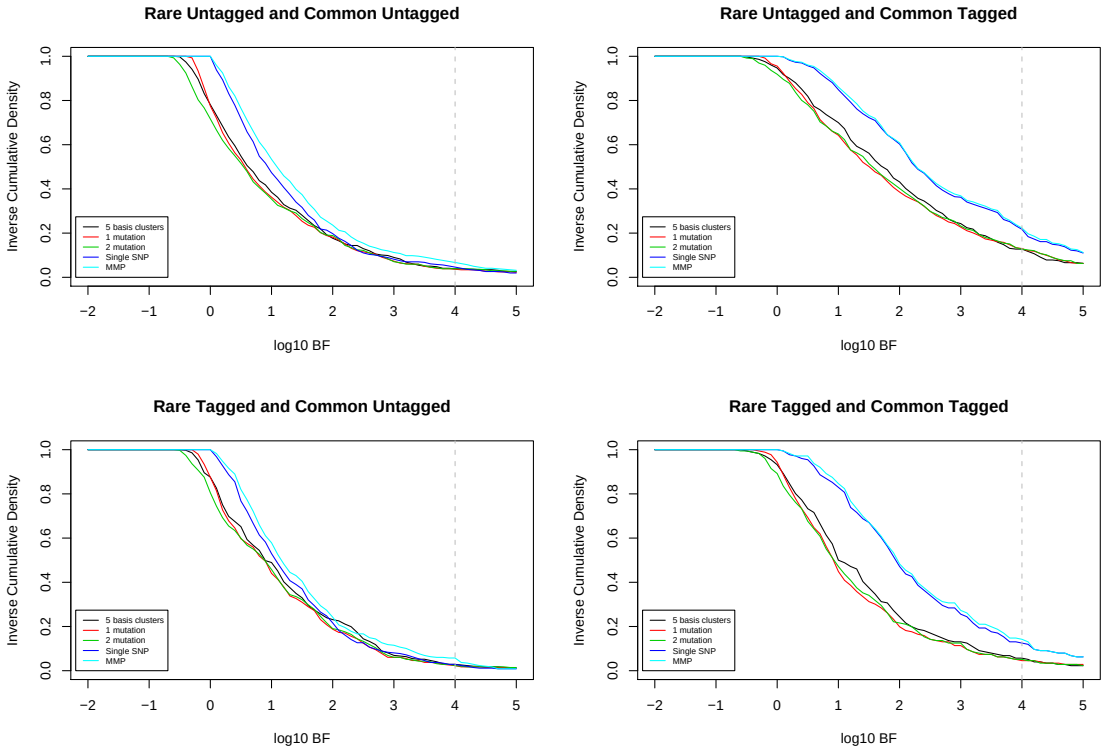


Figure 5.10: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, one rare and one common, of relative risks 1.0 and 1.3, respectively. Since the rare disease allele has a relative risk of 1.0, the data sets effectively have only one causal variant. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

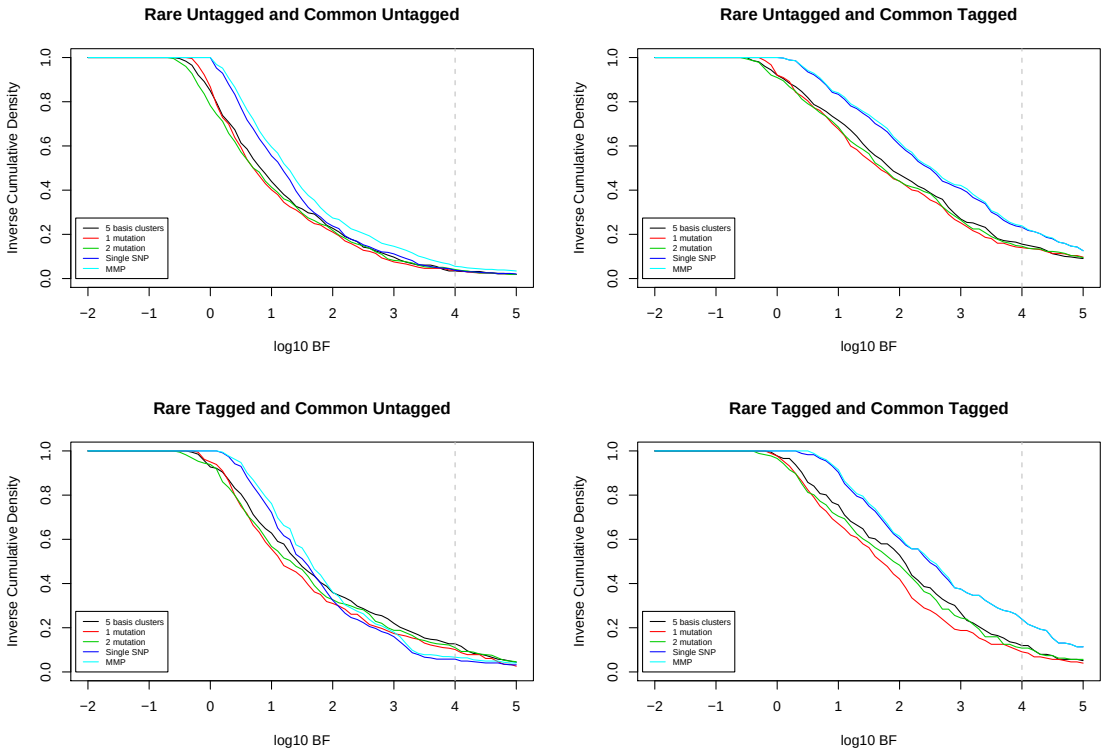


Figure 5.11: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, one rare and one common, of relative risks 1.5 and 1.3, respectively. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

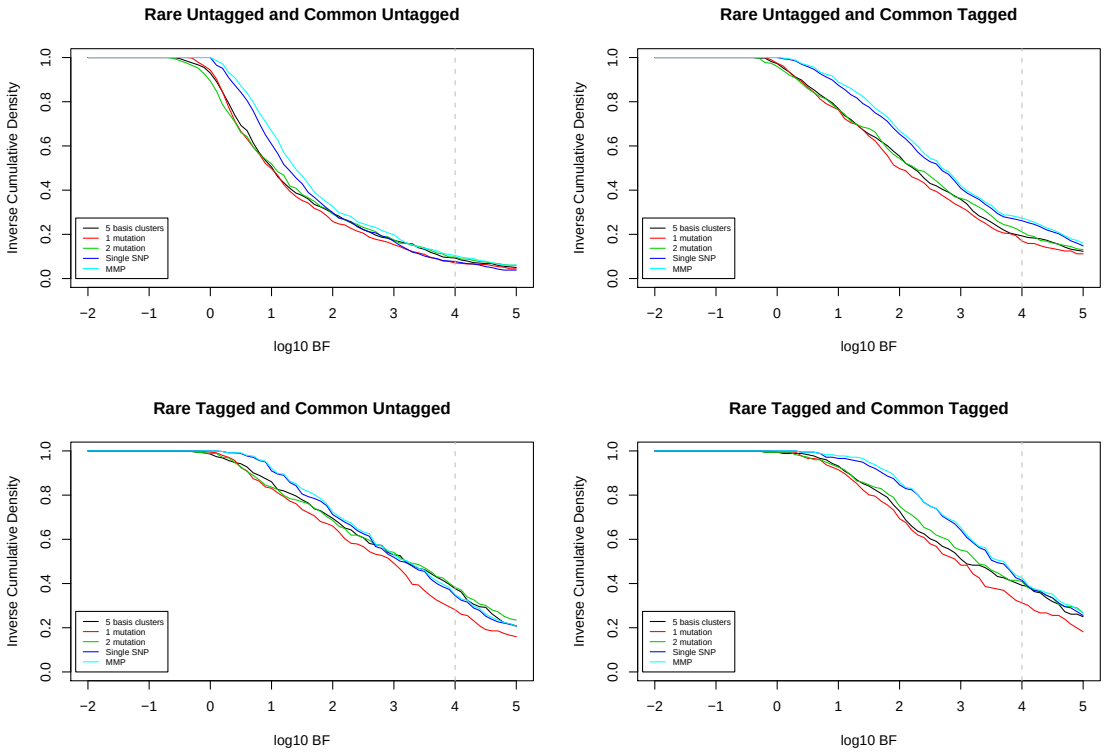


Figure 5.12: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, one rare and one common, of relative risks 2.0 and 1.3, respectively. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

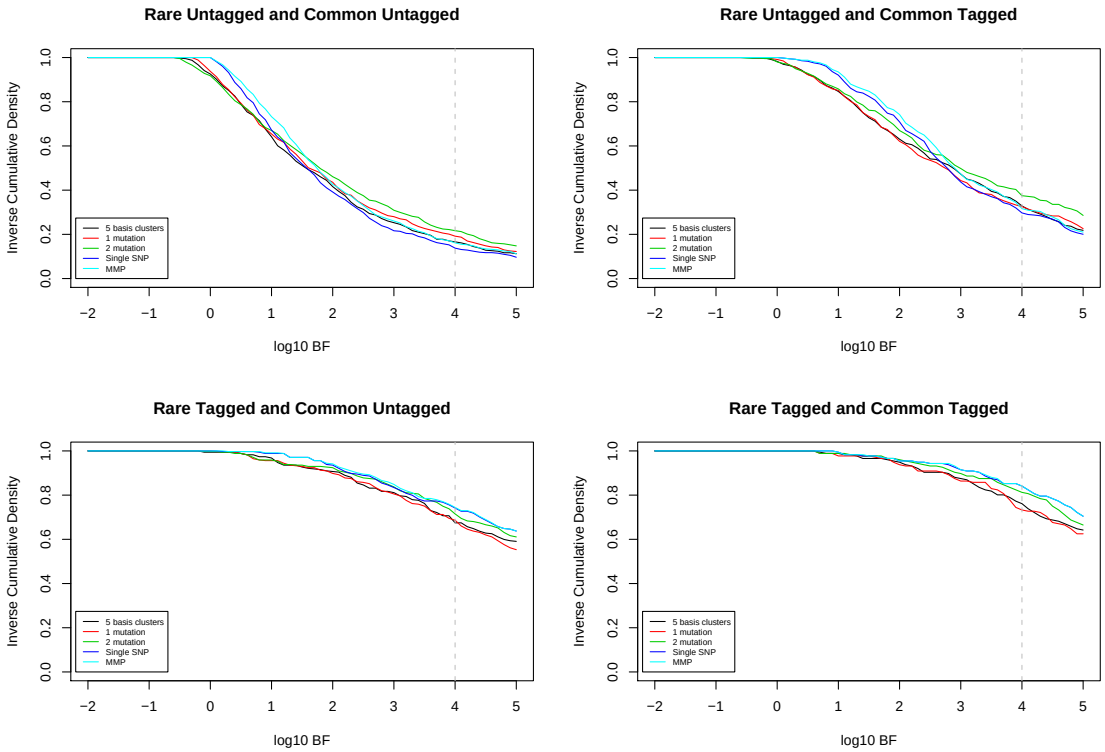


Figure 5.13: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, one rare and one common, of relative risks 2.5 and 1.3, respectively. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

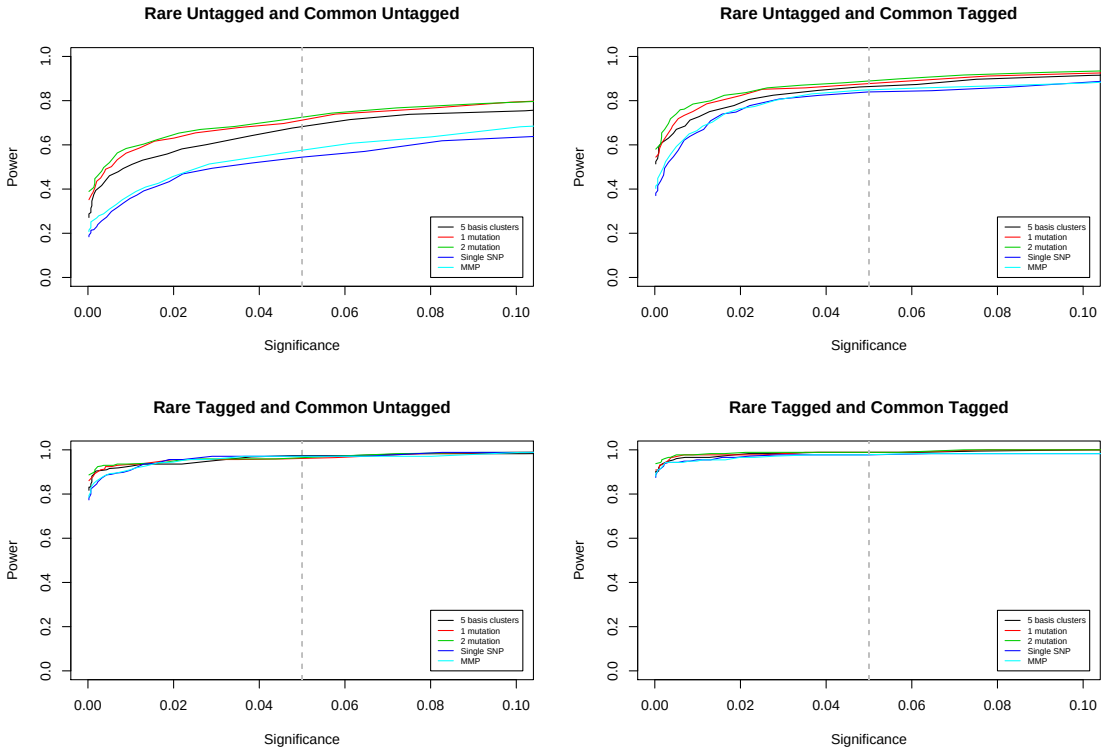


Figure 5.14: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, one rare and one common, of relative risks 2.5 and 1.3, respectively. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity). The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 5% significance threshold.

Disease model 2

We repeated our simulation approach but this time with pairs of disease SNPs that both have a MAF between 5% and 20%. Again we only generated data sets for each category in regions that have more than 50 suitable SNP pairs. The number of data sets and regions generated for each category are give below.

- **A2** untagged common SNP and untagged common SNP (373 data sets, 4 regions)
- **B2** untagged common SNP and tagged common SNP (203 data sets, 2 regions),
- **C2** tagged common SNP and untagged common SNP (618 data sets, 2 regions),
- **D2** tagged common SNP and tagged common SNP (655 data sets, 4 regions).

We set the relative risk of the first allele to be 1.3 and varied the relative risk of the second allele from 1.0 to 1.5; the interaction parameter was set to 1.0, i.e. no epistasis. Our results are illustrated in Figures 5.15-5.18.

As before, the similarity between Figures 5.15 and 5.16 where the second common disease allele has a relative risk of 1.0 and 1.1, implies that power of each analysis is mainly influenced by the disease allele with relative risk 1.3 in those scenarios. The Single SNP and MMP methods are powerful at detecting common variants and have the largest Bayes Factors.

As the relative risk of the second causal variant increases, the signal of association is increasingly influenced by two causal variants and using the 2-mutation model

becomes more powerful relative to the other methods. This is expected since it explicitly models for two mutations. Using a 10^4 Bayes Factor significance threshold, GENECLUSTER is up to 7% more powerful when the second disease allele has a relative risk of 1.5 (Table 5.8). Again, we note that the Bayes Factor for 2-mutation model appears to be significantly larger than the 1-mutation model, which suggests that we have power to detect the presence of allelic heterogeneity.

We note that using 5 basis clusters from Model B is more powerful relative to the other methods under disease model 2 than under disease model 1 (for example compare Figures 5.13 and 5.18). This is consistent with our earlier arguments because now both disease alleles are common and are likely be caused by mutations on higher levels of the genealogical tree (than the mutations responsible for the rare allele simulated under disease model 1), above the level where the basis clusters are defined.

Method	A2	B2	C2	D2
5 Basis Clusters	0.59	0.77	0.61	0.93
1-Mutation Model	0.58	0.72	0.55	0.92
2-Mutation Model	0.60	0.79	0.63	0.93
Single SNP	0.48	0.80	0.49	0.93
MMP	0.53	0.82	0.55	0.94

Table 5.8: Power of GENECLUSTER and SNP based approaches to detect association with two causal variants and a Bayes Factor significance threshold of 10^4. Both disease alleles are common (MAF between 5% and 20%) and they have relative risks of 1.3 and 1.5. The columns refer to the whether each disease SNP is tagged by an Affymetrix SNP or a MMP and the rows refer to the methods that we used to detect for association. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from.

Comparisons of each method based on the ROC are displayed in Figures B.7-B.9 in Appendix B. At the 5% significance level GENECLUSTER is up to 6% more powerful than the Single SNP and MMP methods (Table 5.9).

Method	A2	B2	C2	D2
5 Basis Clusters	0.88	0.99	0.94	0.99
1-Mutation Model	0.89	0.99	0.95	1.00
2-Mutation Model	0.90	0.99	0.95	1.00
Single SNP	0.82	0.99	0.91	1.00
MMP	0.84	1.00	0.93	0.99

Table 5.9: Power of GENECLUSTER and SNP based approaches to detect association with two causal variants and a significance threshold of 5%. Both disease alleles are common (MAF between 5% and 20%) and they have relative risks of 1.3 and 1.5. The columns refer to the whether each disease SNP is tagged by an Affymetrix SNP or a MMP and the rows refer to the methods that we used to detect for association. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from.

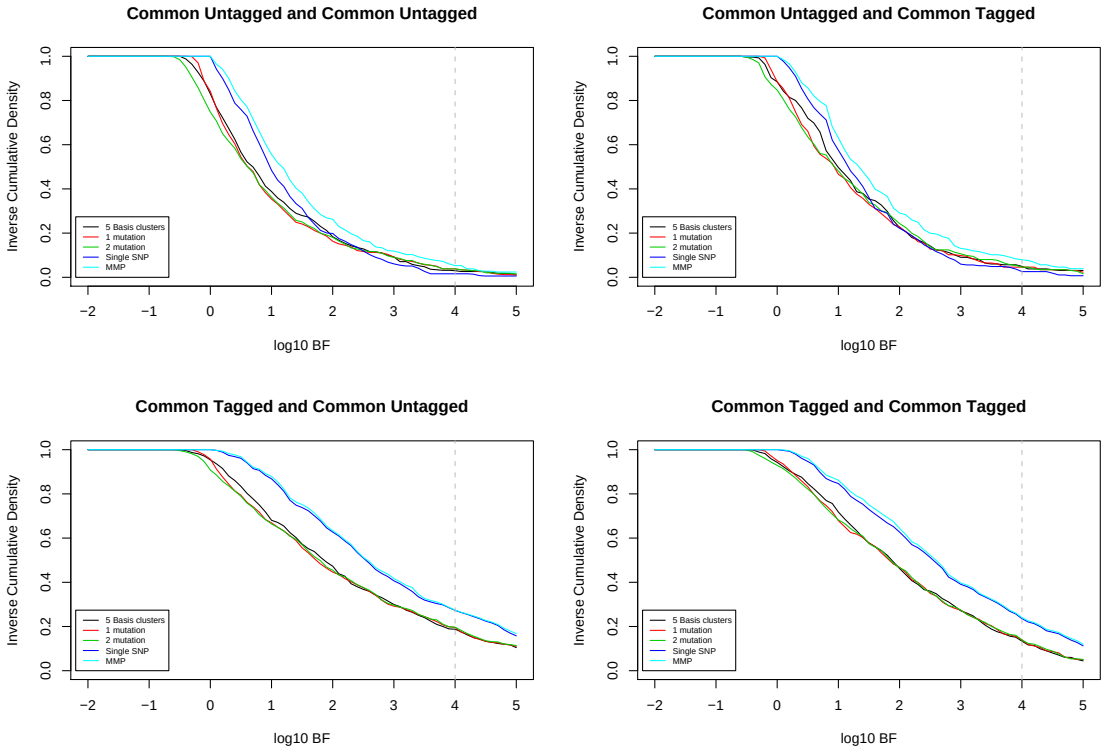


Figure 5.15: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, both common, of relative risks 1.3 and 1.0. Since one of the disease alleles has a relative risk of 1.0, the data sets effectively have only one causal variant. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

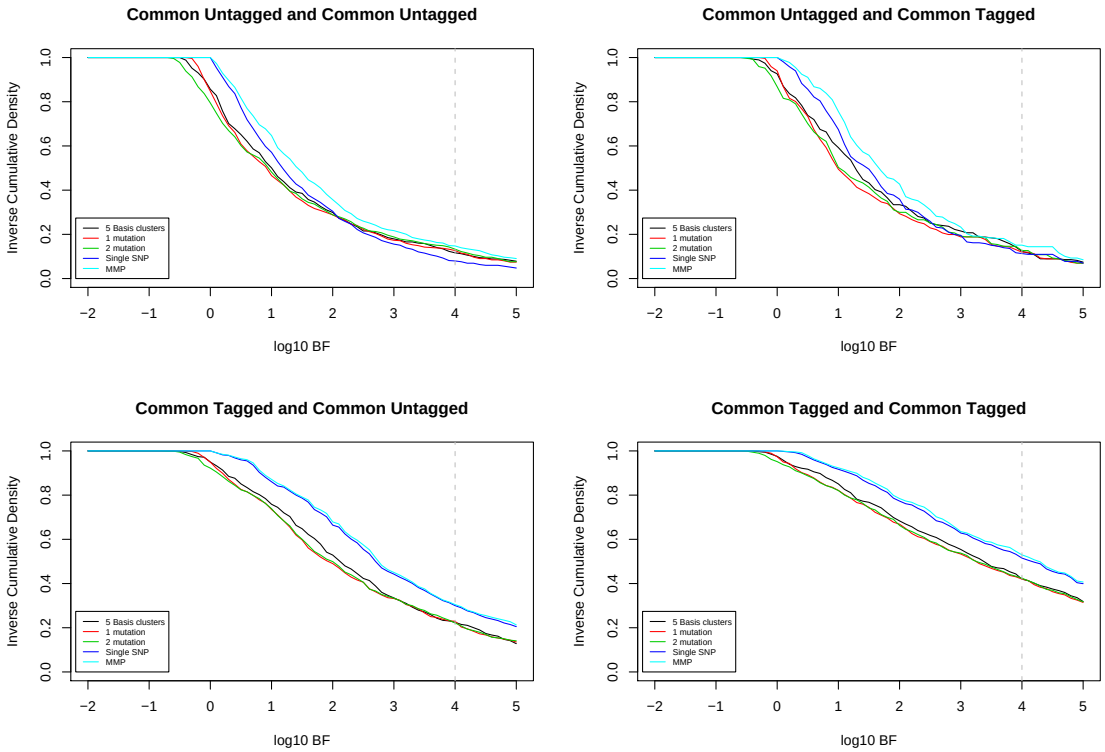


Figure 5.16: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, both common, of relative risks 1.3 and 1.1. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

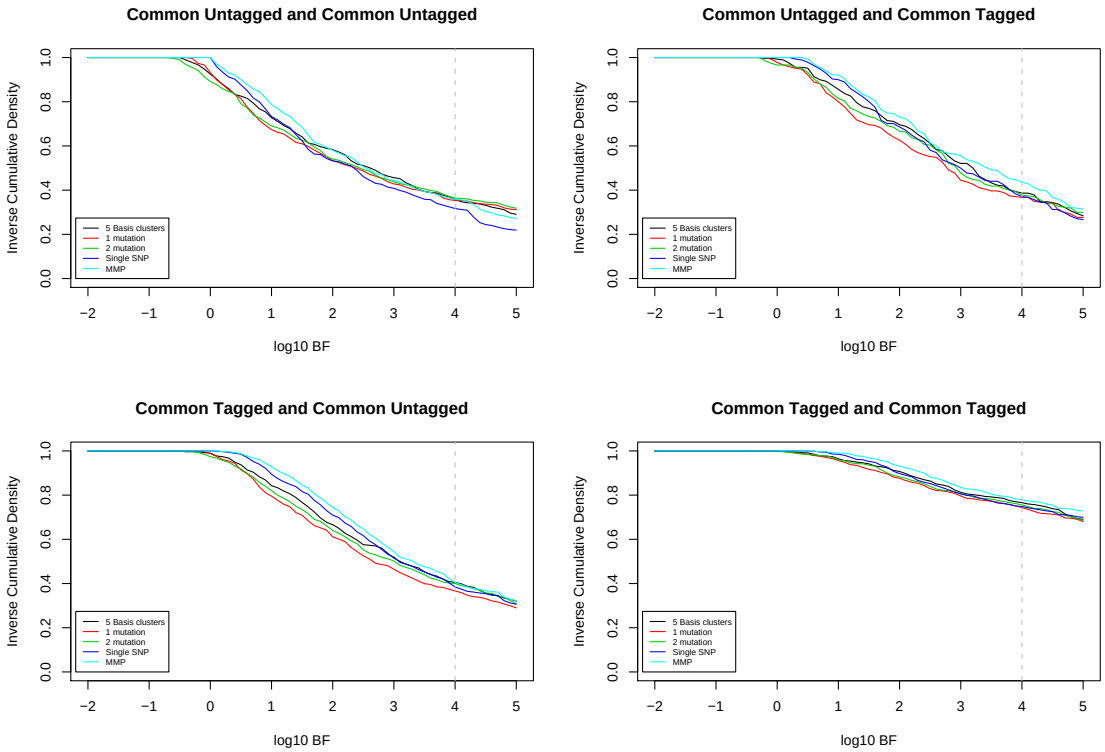


Figure 5.17: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, both common, of relative risks 1.3 and 1.3. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

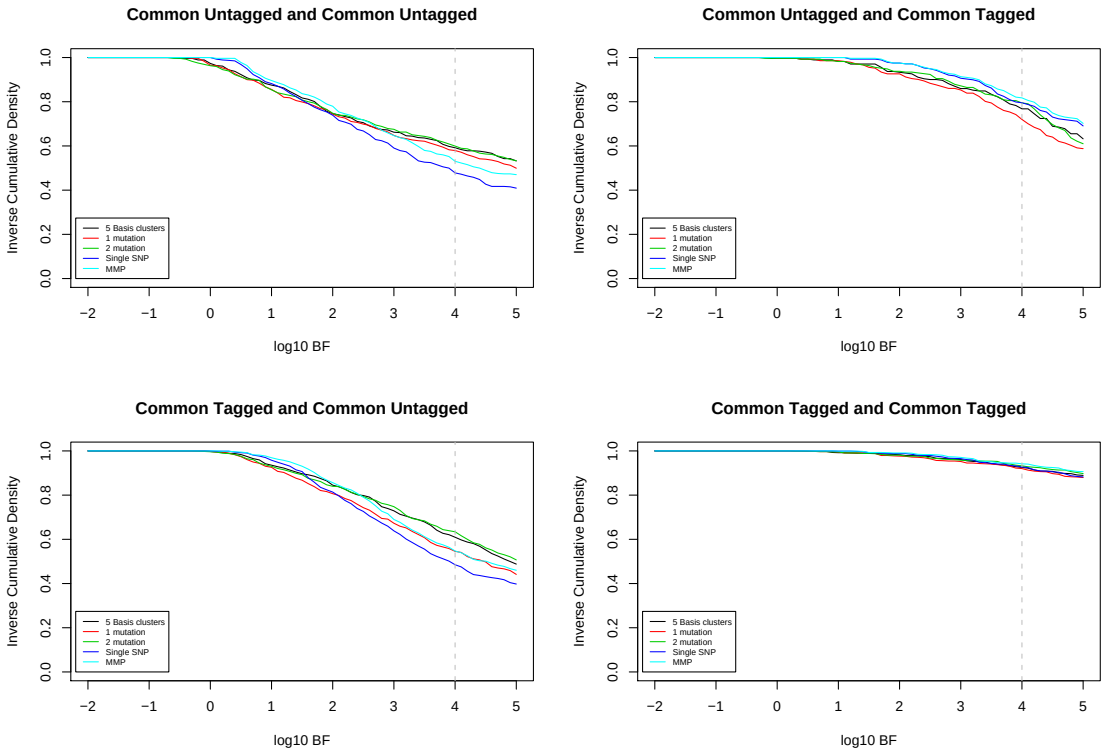


Figure 5.18: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from Model A and 5 basis clusters from Model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with two causal variants, both common, of relative risks 1.3 and 1.5. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

5.2.5 Detecting regions of allelic heterogeneity

By working in a Bayesian framework we can calculate the posterior probability of the 2-mutation model compared to the 1-mutation model based on their Bayes Factors. Let BF_1 and BF_2 be the Bayes Factors under the 1-mutation model and 2-mutation model respectively, then the Bayes Factor comparing the 2-mutation model to the 1-mutation model is given by

$$\begin{aligned} BF_{21} &= \frac{\Pr(D|2 \text{ mutations})}{\Pr(D|1 \text{ mutation})} \\ &= \frac{\Pr(D|2 \text{ mutations})/\Pr(D|0 \text{ mutation})}{\Pr(D|1 \text{ mutations})/\Pr(D|0 \text{ mutation})} \\ &= \frac{BF_2}{BF_1}, \end{aligned} \tag{5.40}$$

where D is the observed data and "0 mutation" is the null model. As an illustration, if we assume a 1:1 prior odds for the 1-mutation model against the 2-mutation model, then $\log_{10} BF_{21} = 0.2$ (1.0) corresponds to a posterior probability for the 2-mutation model of 0.61 (0.91). A better fit with the 2-mutation model should indicate the presence of allelic heterogeneity. In this section we explore the power of using BF_{21} as a way to detect the presence of allelic heterogeneity.

We re-examined the results from the simulation studies in Section 5.2.4 and found the empirical distribution of BF_{21} defined above, where BF_1 and BF_2 are the maximum Bayes Factor in a data set under the 1-mutation model and 2-mutation model, respectively. We have plotted the inverse cumulative distribution for BF_{21} under disease model 1 in Figure 5.19 and disease model 2 in Figure 5.20, which are also summarised in Tables 5.10 and 5.11. In both figures we observe a common trend: as the relative risk of one of the causal alleles increases, we obtain more power.

This is because if the effect of one causal allele is too weak to be detected then the 1-mutation model will provide a better fit. So under disease model 1, a 1-mutation model will be a better model fit to the association with the common allele if the effects of the rare allele can not be detected; similarly in disease model 2, the 1-mutation model can provide a better fit to the association with the first common disease allele if the second disease allele can not be detected.

It is interesting to note that we obtain more power under disease model 1 than disease model 2. A possible reason for this is that the difference in the relative risks between the two causal variants are greater in disease model 1 than in disease model 2. If the difference in relative risks of the two causal variants is large, then we will observe 3 distinct haplotype risk groups in our case-control sample. The 2-mutation model, which allows for 3 independent haplotype penetrance parameters will provide a better model fit to this data compared to the 1-mutation model, which will only allow two independent haplotype penetrance parameters. Conversely, if the difference in the relative risks between the two causal variants is small, then we might obtain a better model fit with the 1-mutation model, which places all haplotypes carrying either causal variant into the same penetrance group. We can see a possible example of this in Figure 6.15 in the next chapter where we applied our methods to the *NOD2* locus for Crohn's disease in the WTCCC data (The Wellcome Trust Case Control Consortium, 2007).

5.3 Discussion

In this chapter we have described a new method for analysing association studies, which is applicable to unphased and missing data. A novelty of our approach is the

Relative Risks	Disease SNP type			
	A1	B1	C1	D1
$(1.0, 1.0)$	0 (0)	0.01 (0)	0.01 (0)	0.01 (0)
$(1.0, 1.3)$	0.04 (0)	0.07 (0)	0.05 (0)	0.07 (0)
$(1.5, 1.3)$	0.07 (0.01)	0.14 (0.01)	0.17 (0.01)	0.30 (0.03)
$(2.0, 1.3)$	0.19 (0.03)	0.24 (0.07)	0.41 (0.17)	0.52 (0.16)
$(2.5, 1.3)$	0.26 (0.06)	0.43 (0.13)	0.51 (0.25)	0.65 (0.34)

Table 5.10: Power to detect allelic heterogeneity when there are two causal variants, one rare (MAF less than 2%) and one common (MAF between 5% and 20%). Each row refers to relative risks of the two disease alleles, for example (2.0, 1.3) refers to the scenario where the rare and common variants have relative risks of 2.0 and 1.3, respectively. The columns refer to the whether each disease SNP is tagged by an Affymetrix SNP or a MMP (we have used the abbreviation defined in Section 5.2.4). The numbers in the table refer to the frequency of data sets with the maximum $\log_{10}$ Bayes Factor of at least 0.2 greater (1.0 in brackets) under the 2-mutation model than the 1-mutation model. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from.

construction of a genealogy for the case-control sample conditional on a reference panel. This construction relies on combining a genealogical tree for the panel with a HMM, which averages over the phase uncertainty to map the haplotypes of each case-control individual to a leaf in the tree. This process is computationally efficient and provides us a way to perform inference. We proposed two models of association based on the constructed genealogy.

One of the models, Model B, is adopted from the previous chapter, where we summarise the information in the lower levels of the tree using a set of basis clusters and integrate over the set of haplotype groupings that can be constructed from merging these clusters. Each grouping is given its own penetrance parameter, so this method can remain powerful when there are multiple disease mutations. Our approximation using the basis clusters allows us to integrate over the topology of the upper parts of the tree and thus accounting for some of the uncertainty in estimating the tree for the panel. One other advantage of Model B, over Model A, is that the k basis

Relative Risks	Disease SNP type			
	A2	B2	C2	D2
(1.0, 1.0)	0 (0)	0.01 (0)	0 (0)	0 (0)
(1.3, 1.0)	0.05 (0)	0.08 (0.01)	0.03 (0)	0.06 (0)
(1.3, 1.1)	0.10 (0.01)	0.10 (0.01)	0.06 (0)	0.05 (0.01)
(1.3, 1.3)	0.16 (0.02)	0.27 (0.03)	0.28 (0.05)	0.16 (0.02)
(1.3, 1.5)	0.28 (0.09)	0.47 (0.08)	0.48 (0.18)	0.21 (0.06)

Table 5.11: Power to detect allelic heterogeneity when there are two causal variants, both common (MAF between 5% and 20%). Each row refers to relative risks of the two causal variants, for example (1.3, 1.5) refers to the scenario where the causal variants have relative risks of 1.3 and 1.5. The columns refer to the whether each disease SNP is tagged by an Affymetrix SNP or a MMP (we have used the abbreviation defined in Section 5.2.4). The numbers in the table refer to the frequency of data sets with the maximum $\log_{10}$ Bayes Factor of at least 0.2 greater (1.0 in brackets) under the 2-mutation model than the 1-mutation model. The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from.

cluster model allows up to k independent penetrance parameters under its model of association. If there are many causal mutations at a locus, then the k cluster model might be a better alternative to Model A, since the 1-mutation and 2-mutation models do not allow for enough mutations in the tree and the k-mutation model might not be computationally tractable. However, by cutting the tree at a certain level, analysis can be underpowered if the true causal mutation occurred in lower levels of the tree.

Our second model, Model A, is similar to the ones presented by Minichiello and Durbin (2006), where we place mutations on the branches of the constructed genealogical tree to assess how well they explain the phenotypic status observed at the leaves. Modelling association this way naturally relates to how the disease mutation arises and propagates in a population and we can allow for allelic heterogeneity when more than one mutation is modelled. In addition, analysing the likely locations of each mutation in the tree can yield useful information about the underlying disease model, for example the haplotype backgrounds that disease alleles reside on. We

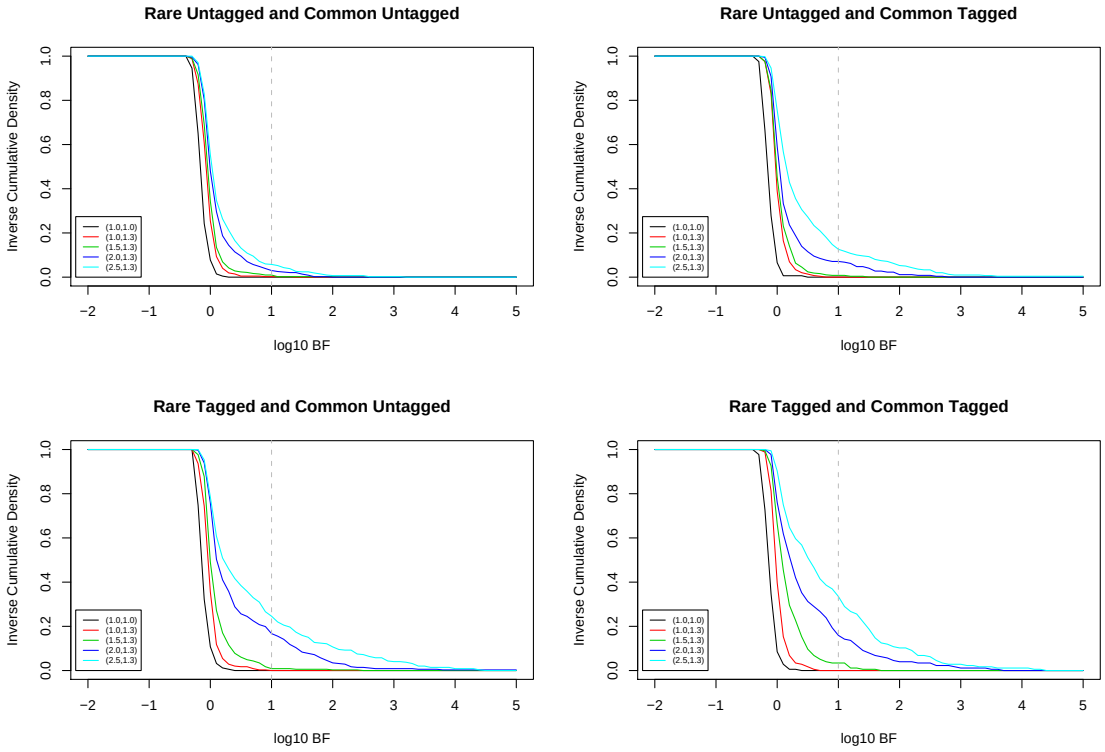


Figure 5.19: The empirical inverse cumulative density function of the difference between the maximum $\log_{10}$ Bayes Factors under the 2-mutation and 1-mutation models per data set. We have analysed data sets with two causal variants, one rare and one common, and the legend indicates the relative risks of each variant. For example, the red line refers to the scenario where the rare and common variants have relative risks of 1.0 and 1.3, respectively. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a difference exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale.

will demonstrate this feature in the next chapter (in Section 6.3.3) when we apply our method to real population data.

To assess the performance of our methods, we conducted two simulation studies. We first simulated data sets under a single mutation model and a range of effect sizes. The marker density, sample size and the way the data sets were analysed are typical of current genome-wide association studies, for example the WTCCC study. Our results show that GENECLUSTER is more powerful than the Single SNP and

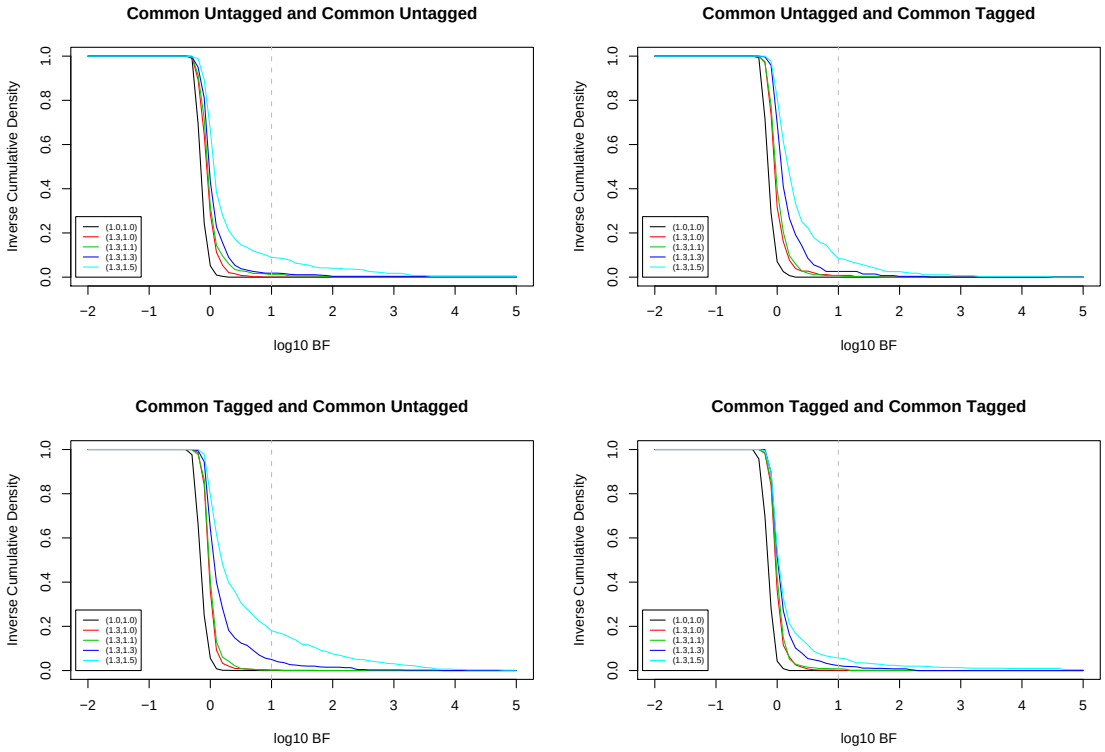


Figure 5.20: The empirical inverse cumulative density function of the difference between the maximum $\log_{10}$ Bayes Factors under the 2-mutation and 1-mutation models per data set. We have analysed data sets with two causal variants, both common, and the legend indicates the relative risks of each variant. For example, the red line refers to the scenario where the variants have relative risks of 1.3 and 1.0. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a difference exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale.

MMP methods for detecting rare and untagged variants. However, to detect common causal variants that are either typed or tagged, the signal of association is better captured by single SNP tests. Therefore, our method represents a complementary rather than replacement approach to scanning the genome for association. Very similar results were reported by Marchini et al. (2007), which compared IMPUTE with the single SNP and MMP methods. Since IMPUTE uses the same underlying model of LD as GENECLUSTER, it is unsurprising, and reassuring, to find that the two sets of studies are consistent. Due to constraints on time and resources, compar-

isons between GENECLUSTER and IMPUTE were not possible under simulation. However, we do expect differences between the two methods. Analyses using IMPUTE rely on SNP based approaches to test for association at the imputed SNP and will suffer some of the same drawbacks as the Single SNP and MMP methods that we have evaluated. Consider a situation where the causal variant is untyped and not well tagged by any SNP in the reference panel, but is well identified by the its local haplotype background. The IMPUTE analysis will be underpowered in this situation but GENECLUSTER, which can test for association at any given position, will detect the high risk haplotypes that carry the disease allele. A second situation where GENECLUSTER can outperform IMPUTE is when there is allelic heterogeneity since it models for it.

We simulated scenarios of allelic heterogeneity involving two disease SNPs under two disease models. The first one involved a rare and one common disease allele, which is loosely based on the disease model observed at the *NOD2* locus for Crohn's disease (Hugot et al., 2001; Ogura, Y. et al., 2001); the second involved two common disease alleles, which is loosely based on the disease model at the *IL23R* locus for Crohn's disease (Duerr et al., 2006). We found that for both disease models, when both causal variants confer a strong risk for disease, the 2-mutation model can provide a significant boost in power over the SNP based approaches. In addition, our simulations show that using Model B can detect allelic heterogeneity: if the Bayes Factor is significantly larger for the 2-mutation model than the 1-mutation model, then it is an indication for the presence of allelic heterogeneity.

There are, however, weaknesses in our simulation design. We selected the disease SNPs in each data set based purely on MAF and their physical distance apart, without consideration for the local LD structure. The relative risks that we have

assigned to each disease allele were also arbitrary. Both of these problems stem from the fact that we have very little idea about the scenarios of allelic heterogeneity that we are likely to encounter in population data. Further, the number of data sets we simulated was limited, which will have added noise to our results. Therefore, a larger and more realistic simulation study is required.

More generally, we have used HAPGEN for simulation, which uses the same LS model for LD as GENECLUSTER. Although we have shown in Chapter 3 that our simulated data sets appear realistic, we can not be sure that there are no biases in our simulations, which might favour GENECLUSTER.

One final point on simulation design is that we have based all our simulations on the log additive disease model in terms of genotype relative risk. We have done this because the majority of causal variants found in the population, by The Wellcome Trust Case Control Consortium (2007) for example, conform to this model. We have established that GENECLUSTER performs well under this model but we have assumed a haploid disease model, which is log additive in genotype relative risk. Other disease models are known to exist, so we need to know how robust GENECLUSTER is to the underlying disease model. Without a simulation study, it is difficult to know exactly how the current implementation of GENECLUSTER will perform against a dominant disease model. We would expect to retain some power since there will be a higher frequency of disease haplotypes being carried by case individuals (although this increase in frequency will be less compared to an additive disease model because of the dominance effect). Those disease haplotypes will be mapped to a small subset of leaves in the estimated tree for the panel, resulting in an excess of cases there. Against a recessive disease model, however, one can imagine that we would achieve very low power if the disease haplotype is common

in the population. In this scenario, a large number of disease haplotypes are carried by heterozygote controls and will be mapped to the same leaves as the haplotypes carried by the homozygotes cases, which will reduce our signal greatly. However, if the disease allele is rare in the population, the oversampling of case individuals will result in a large proportion of disease haplotypes being carried by homozygote cases, which will be mapped to the same leaves. Therefore, when this is the case, we will observe an excess of case haplotypes mapped to a subset of leaves and retain some power.

Another potential weakness of our approach is that we have performed inference based only on a single tree at each position, which ignores the uncertainty in our estimated trees. It is difficult to know how much uncertainty there is in our trees and the answer to this question is largely dependent on the accuracy of the methods we have used to construct them. We are, however, encouraged by the fact that GENECLUSTER appears to produce substantial power despite this approximation, which suggests that the strength in our data may be strong enough to overcome any mis-specification in the trees. We hope that this will be the case for population data as well.

GENECLUSTER requires approximately 10 minutes to examine a 500Kb data set, comprising of 2000 case and 2000 control individuals typed at Affymetrix 500K SNPs, at SNPs every 5Kb apart. The use of the reference panel means that the number of states in our HMM is fixed to P^2, where P is the size of the panel. Computation cost therefore scales linearly with the number of genotyped SNPs and the number of individuals in a study, which means that GENECLUSTER is applicable to large scale studies as we will demonstrate in the next chapter.

Data from large-scale genome-wide association studies are becoming available, which allows us to assess our method on population data. In the next chapter we will apply GENECLUSTER to real genome-wide case-control data from the WTCCC study and compare our results with the single SNP test and IMPUTE.

A C++ implementation of GENECLUSTER is currently available from the Marchini group software website at `http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html`.

Chapter 6

Analysis of the WTCCC Data

6.1 Introduction

The main WTCCC study (The Wellcome Trust Case Control Consortium, 2007) was the genome-wide association study of approximately 2,000 cases and 3,000 shared controls for 7 complex human diseases of major public health importance – bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). Individuals included in the study were living in Great Britain and the vast majority had self-identified themselves as white Europeans. The control individuals came from two sources: 1,500 individuals from the 1958 British Birth Cohort and 1,500 individuals selected from blood donors recruited as part of the project.

The 17,000 samples were genotyped with the GeneChip 500K Mapping Array Set (Affymetrix chip), which comprise of 500,568 SNPs. 809 samples were excluded after checks for contamination, false identity, non-Caucasian ancestry and relatedness.

Further tests were performed to check for ascertainment bias in the control samples and population structure in the case and control samples, and results showed that neither confounding effects are strong enough to affect the findings. Of the remaining 16,179 individuals, data for 469,557 SNPs, which passed quality control filters on missing data rates and departures from Hardy-Weinberg Equilibrium, were put forward for analysis.

The WTCCC assessed evidence of association in several ways, drawing on both classical and Bayesian statistical approaches. For polymorphic SNPs on the Affymetrix chip, they performed trend tests (1 degree of freedom) and general genotype tests (2 degrees of freedom) between each case collection and the pooled controls, and calculated analogous Bayes Factors. The study reported 21 ("tier 1") signals, from the trend and genotypic tests for Affymetrix SNPs, across the 7 diseases that exceeded a P value threshold of 5×10^{-7}. All of these signals were subjected to visual inspection of cluster plots, which means that genotyping artifacts are unlikely to be responsible for these associations. 12 of the 21 signals represent replications of previously reported findings. A further 58 ("tier 2") signals were also reported with P values above 1×10^{-5}.

In this chapter we will describe the results of using GENECLUSTER to analyse the WTCCC data. Our purpose is twofold:

1 to provide an assessment of GENECLUSTER applied to real population data;

2 to uncover novel disease loci.

With respect to these goals we report that GENECLUSTER can accurately identify regions of allelic heterogeneity and the haplotype backgrounds that contain the dis-

ease alleles; and we find regions of association that previously have not been reported using SNP based methods. The ability to identify the risk haplotype backgrounds at a disease locus has potential applications in fine-mapping and we will explore this further in the next chapter.

6.2 Method

We used GENECLUSTER conditioned on the 120 CEU HapMap Phase 2 haplotypes (Frazer et al., 2007), and built trees using TREESIM and LSTREE at positions 5Kb apart across the genome. We assume that the CEU HapMap sample match the ancestry and reflect the haplotypic variation of the Caucasian European case-control samples. We used the corresponding fine-scale recombination rates and effective population size, $N_e = 11418$, estimated in the HapMap study.

We tested for association at all SNPs where a tree was constructed using the 1-mutation and 2-mutation models of Model A, and the 5 basis clusters model of Model B with a $Geo(0.25)$ prior on grouping size. We chose to use 5 basis clusters because it appeared to perform well in our simulation studies of the last chapter. Model B appeared to be robust against the prior for grouping size, so the prior $Geo(0.25)$ should not have much of an effect on our analyses.

We used a $\beta(20.0, 30.0)$ penetrance prior for both the null and the alternative models. We have chosen this prior because it has a mean of $\frac{2}{5}$, which is approximately the proportion of cases in our case-control sample for each disease. Our prior assumes a mean haplotype (and heterozygote) relative risk of 1.03 with variance 0.07, according to a simulation of 1,000,000 samples (see Figure 6.1). A glance down the tier 1 and

tier 2 table suggests that the heterozygote relative risk for the disease loci range from 0.5 and 2, with most between 0.67 and 1.5. Figure 6.1 therefore suggests that our penetrance prior matches well to the effect sizes that we expect to find in the data.

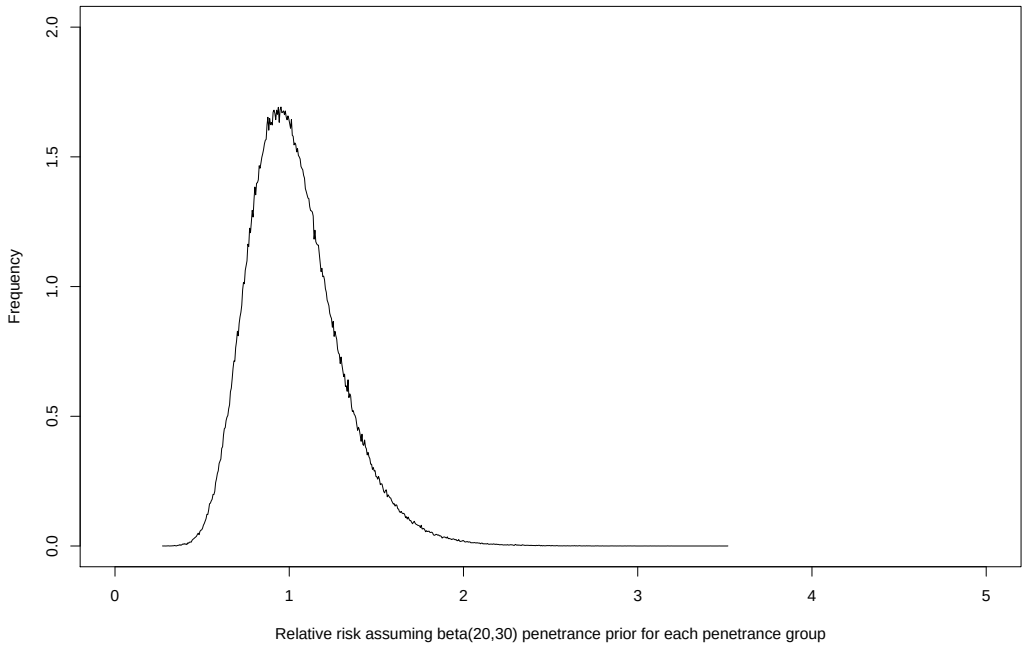


Figure 6.1: The density of the haplotype relative risk of assumed by our prior risk of $\beta(20.0, 30.0)$. The distribution was obtained through simulation of 10^6 samples of $\frac{a}{b}$, where a and b are sampled from the $\beta(20.0, 30.0)$ distribution.

On a multinode computer, comprising of 74 dual processor Opteron 244 nodes with 2 GB RAM, analyses of all 7 diseases were completed in under 4 days. Less than 3 days were needed to calculate the expected number of case and control haplotypes mapped to each leaf of the estimated trees (step 2 in Section 5.1.3) and less than 1 day was needed to test for association under the various models we implemented (step 3 in Section 5.1.4).

6.3 Results

The genome-wide results from GENECLUSTER using TREESIM trees are displayed in Figures 6.2 to 6.4. We set a Bayes Factor significance threshold of 10^4, which we explained in Section 5.2. We find 26 signals (coloured green) that exceed our Bayes Factor threshold (for at least one model) and Table 6.1 shows the location and maximum Bayes Factor of each signal. The table also includes the Bayes Factors at the Affymetrix SNPs and imputed SNPs (using IMPUTE) reported by the WTCCC using a single SNP Bayesian association test under the log odds additive model. A region of association is defined as a collection of SNPs, within 500Kb of each other, where we observe a Bayes Factor greater than the threshold; this definition should separate our signals into independent regions, except perhaps in the HLA region of chromosome 6 where there is extensive LD (de Bakker et al., 2006). Of the 26 signals, 18 are tier 1 signals in the WTCCC paper (coloured green in the table), 4 are tier 2 signals in the WTCCC paper (coloured blue in the table) and 4 were not identified by the WTCCC study.

In Table 6.2 we have summarised the signals of association detected using GENECLUSTER with LSTREE trees. We find the same 24 regions of association that is found with TREESIM trees. This suggests that the topologies of the two sets of trees, which GENECLUSTER uses to detect association, are similar.

collection	chr	region (Mb)	1-mut model	2-mut model	5 basis clusters	Affy SNPs	imputed SNPs
BD	13	79.09	3.53	3.50	4.09	1.10	1.26
CAD	9	21.98-22.11	11.04	11.01	10.57	11.66	11.58
CD	1	67.25-67.47	12.96	17.99	18.43	10.07	15.82
CD	2	233.93-233.99	10.38	10.29	9.89	11.11	11.55
CD	5	40.33-40.65	10.45	14.68	14.79	10.41	10.93
CD	5	131.65-131.83	5.82	6.13	6.24	4.54	7.18
CD	5	150.16-150.3	4.94	4.89	4.85	5.43	5.51
CD	6	31.38	4.61	4.69	5.04	1.96	6.52
CD	6	32.16-32.52	4.00	4.75	6.05	1.40	3.33
CD	10	101.27-101.28	5.32	5.40	5.08	5.91	6.05
CD	16	49.16-49.44	11.44	13.33	12.09	12.00	11.42
CD	18	12.77-12.87	5.56	5.52	5.09	5.42	5.53
RA	1	113.59-114.26	20.73	20.81	14.91	22.36	11.87
RA	6	29.66-33.77	102.92	124.67	95.15	74.84	91.19
T1D	1	113.59-114.23	20.76	20.70	16.43	23.07	13.21
T1D	4	123.59-123.86	4.59	4.58	5.07	4.42	5.63
T1D	6	25.98-33.93	290.18	>300	> 300	306.95	202.71
T1D	10	6.13-6.15	4.35	4.85	3.11	3.31	4.58
T1D	12	54.66-54.78	7.65	7.72	7.84	8.89	8.02
T1D	12	109.83-111.48	10.98	10.98	10.95	12.53	12.74
T1D	15	58.57-58.58	3.20	4.06	4.10	1.08	1.98
T1D	16	10.97-11.12	5.20	5.24	5.50	5.76	6.27
T2D	6	20.79-20.81	4.04	4.15	3.65	4.15	4.35
T2D	9	22.12	5.61	5.71	5.92	1.53	2.90
T2D	10	114.72-114.81	9.73	9.89	9.96	10.14	11.09
T2D	16	52.36-52.38	5.01	5.11	4.97	5.89	5.74

Table 6.1: Regions of association found by GENECLUSTER with TREESIM trees using a Bayes Factor threshold of 10^4. All Bayes Factors are given on the $\log_{10}$ scale. The disease, chromosome and physical region are coloured according to whether the region was reported in the WTCCC paper. Green indicates a reported tier 1 signal, blue a reported tier 2 signal and red an unreported signal. We also include the Bayes Factors from the single SNP Bayesian association test at Affymetrix SNPs and SNPs imputed by IMPUTE. Bayes Factors exceeding a threshold of 10^4 are in black, and in grey otherwise.

collection	chr	region (Mb)	1-mut model	2-mut model	5 basis clusters	Affy SNPs	imputed SNPs
CAD	9	21.99-22.11	11.22	11.17	10.91	11.66	11.58
CD	1	67.25-67.47	12.40	17.95	11.74	10.07	15.82
CD	2	233.93-234	10.47	10.41	10.13	11.11	11.55
CD	5	40.33-40.65	10.50	15.09	14.54	10.41	10.93
CD	5	131.71-131.84	5.53	5.65	6.36	4.54	7.18
CD	5	150.16-150.29	4.57	4.51	4.30	5.43	5.51
CD	6	31.34-31.36	2.87	4.38	3.13	1.96	6.52
CD	6	32.11-32.52	3.62	4.58	4.61	1.40	3.33
CD	10	101.27-101.28	5.35	5.31	5.03	5.91	6.05
CD	16	49.12-49.44	10.55	14.17	11.53	12.00	11.42
CD	18	12.77-12.87	4.56	4.67	4.71	5.42	5.53
RA	1	113.58-114.26	20.19	20.33	8.51	22.36	11.87
RA	6	29.62-33.92	119.42	129.59	128.11	74.84	91.19
T1D	1	113.56-114.25	20.26	20.22	12.49	23.07	13.21
T1D	4	123.61-123.87	3.67	3.93	4.88	4.42	5.63
T1D	6	25.55-33.95	> 300	> 300	> 300	306.95	202.71
T1D	10	6.13-6.16	4.89	5.94	5.85	3.31	4.58
T1D	12	54.66-54.77	7.78	7.71	7.74	8.89	8.02
T1D	12	109.83-111.48	11.69	11.68	11.61	12.53	12.74
T1D	16	10.97-11.12	5.08	5.09	5.37	5.76	6.27
T2D	6	20.79-20.8	3.97	4.17	3.62	4.15	4.35
T2D	9	22.12	5.69	5.88	5.84	1.53	2.90
T2D	10	114.72-114.81	9.19	9.30	9.53	10.14	11.09
T2D	16	52.36-52.38	4.70	4.88	4.97	5.89	5.74

Table 6.2: Regions of association found by GENECLUSTER with LSTREE trees using a Bayes Factor threshold of 10^4. All Bayes Factors are given on the $\log_{10}$ scale. We also include the Bayes Factors from the single SNP Bayesian association test at Affymetrix SNPs and SNPs imputed by IMPUTE. Bayes Factors exceeding a threshold of 10^4 are in black, and in grey otherwise.

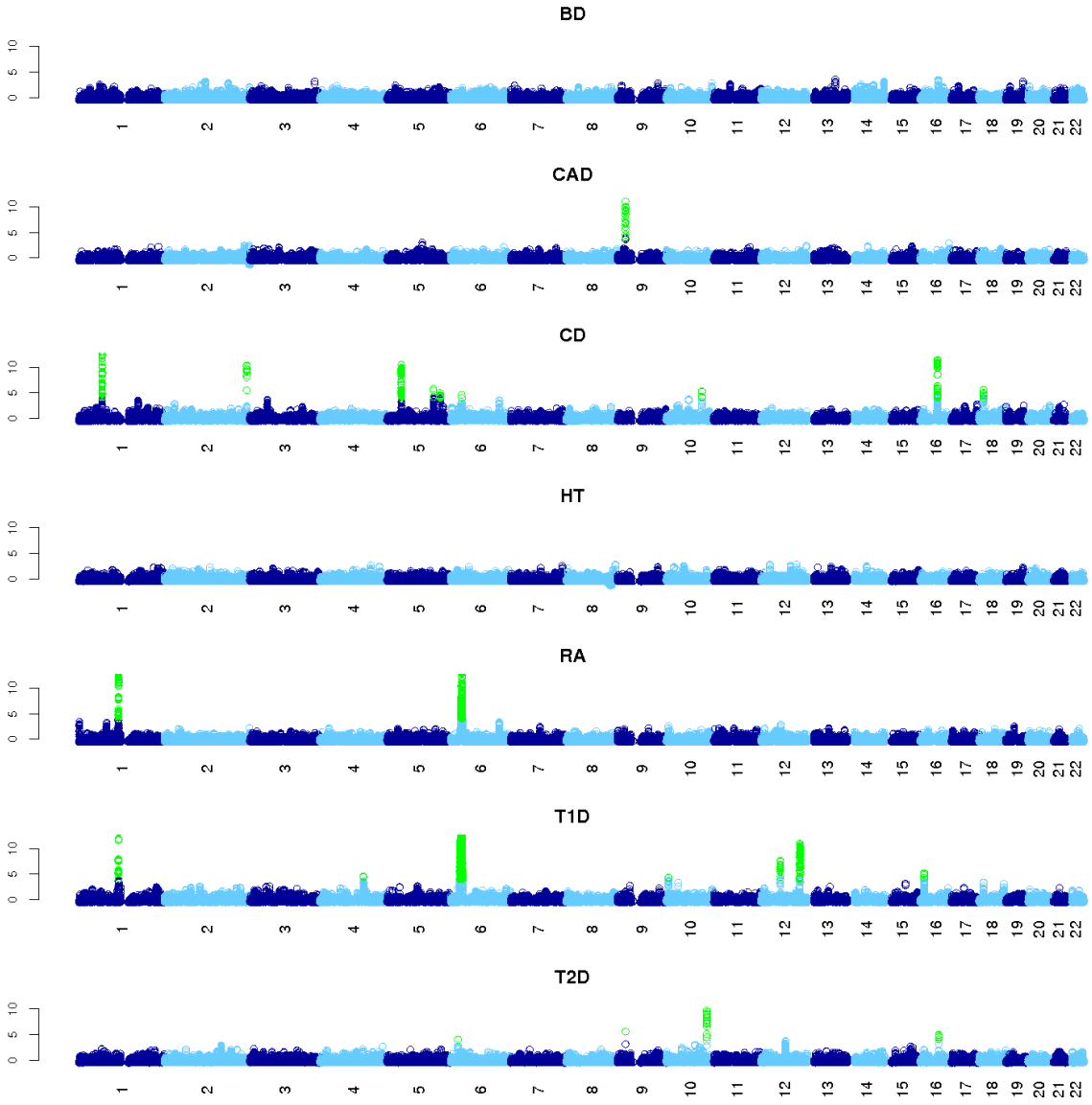


Figure 6.2: Results of genome-wide analysis using the 1-mutation model for all seven diseases. On the y-axis is the $\log_{10}$ BF and the x-axis is the chromosomal location. Signals exceeding a Bayes Factor threshold of 10^4 are coloured green.

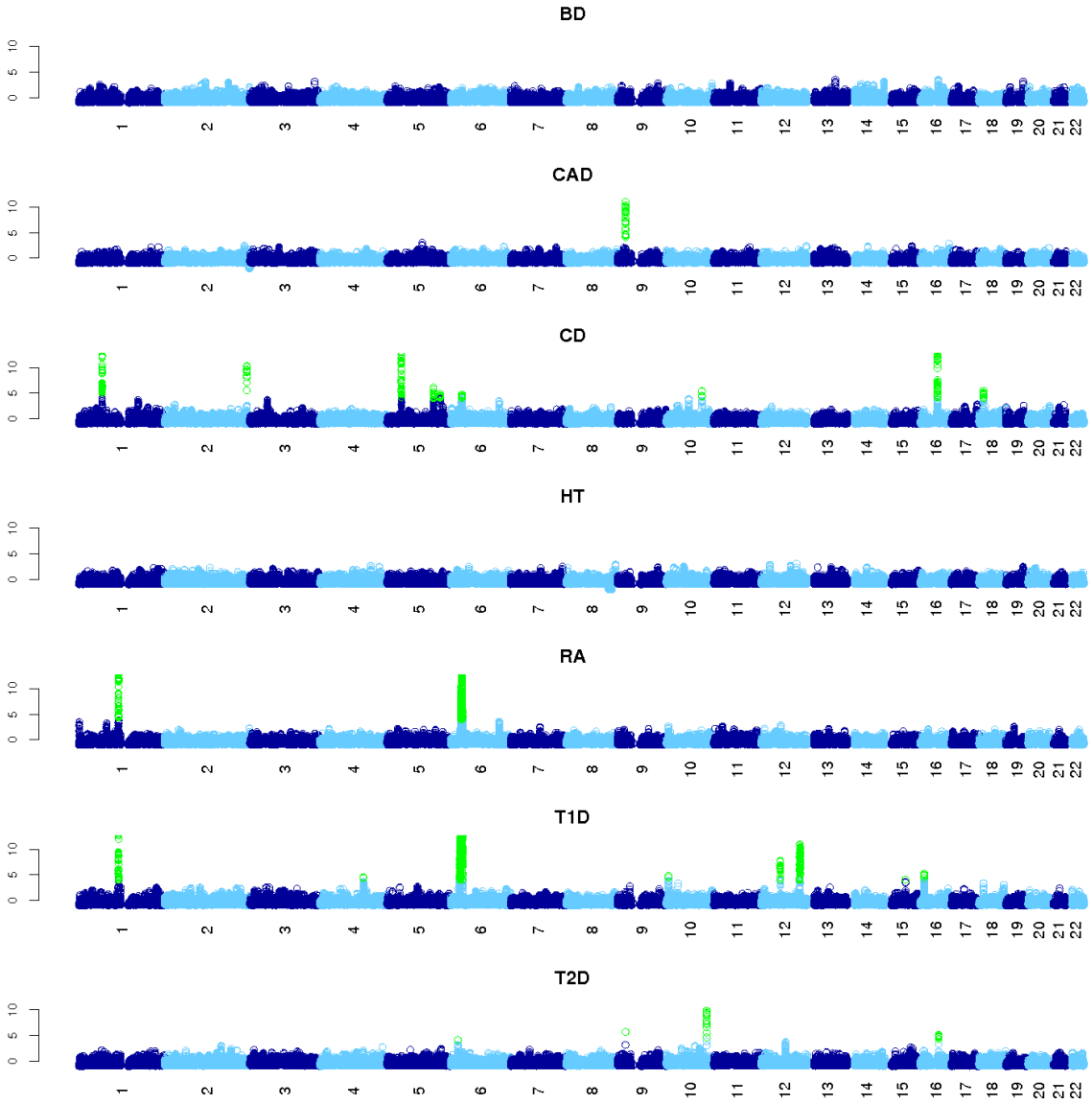


Figure 6.3: Results of genome-wide analysis using the 2-mutation model for all seven diseases. On the y-axis is the $\log_{10}$ BF and the x-axis is the chromosomal location. Signals exceeding a Bayes Factor threshold of 10^4 are coloured green.

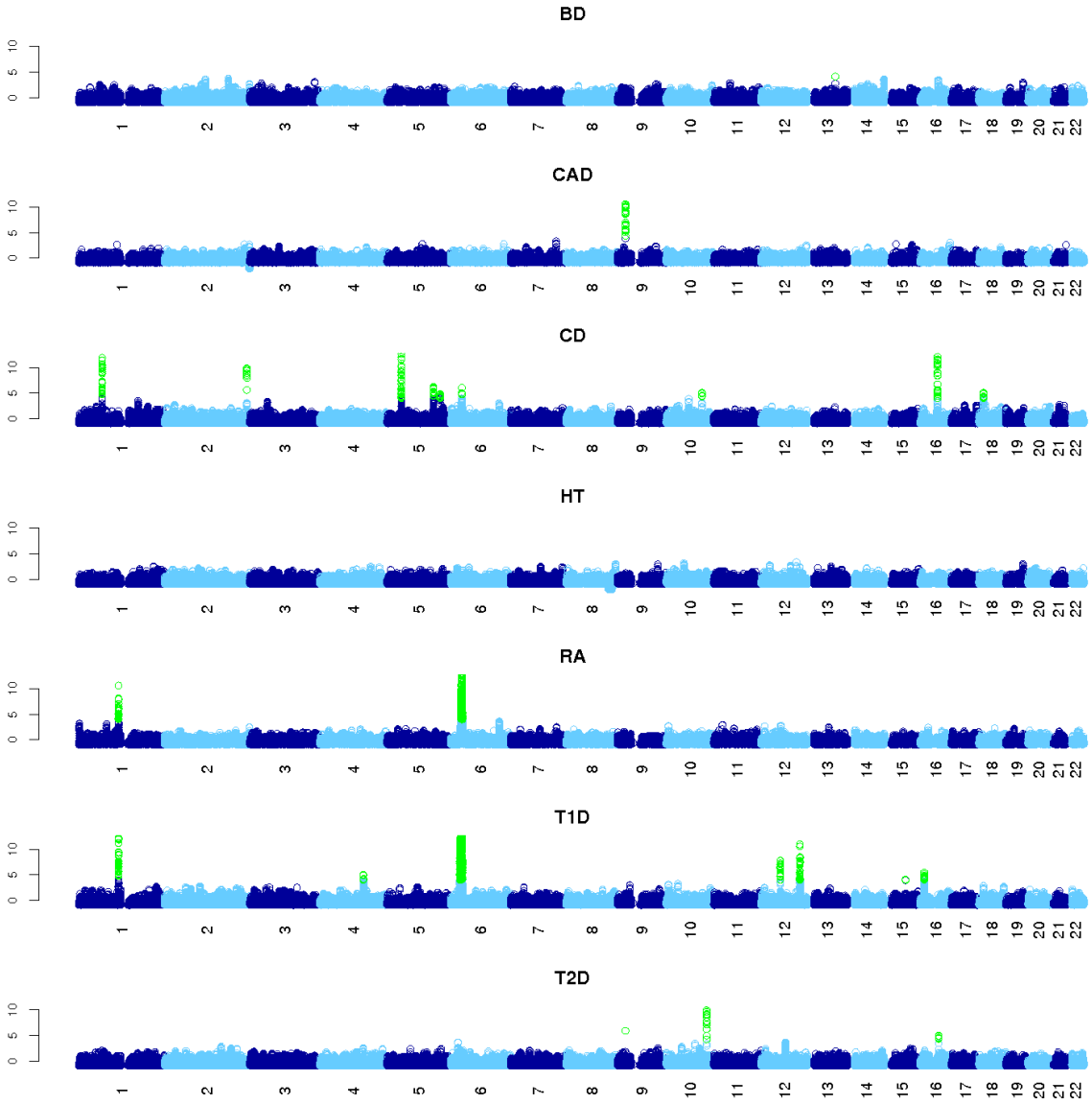


Figure 6.4: Results of genome-wide analysis using 5 basis clusters for all seven diseases. On the y-axis is the $\log_{10}$ BF and the x-axis is the chromosomal location. Signals exceeding a Bayes Factor threshold of 10^4 are coloured green.

6.3.1 Comparing GENECLUSTER with SNP based approaches

The Bayes Factor threshold, above which a signal is declared significant, is dependent on the investigator's choice and the context of the study. It is therefore important to know the performance of our method at all disease loci not just those that exceed our thresholds. We therefore examined the 3 tier 1 signals that are not in Table 6.1 and they are summarised in Table 6.3.

The WTCCC reported the BD locus to have a very weak signal under the additive model but is significant under a genotype model, which allows for dominance and recessive effects. GENECLUSTER uses a haploid disease model, which assumes a log additive disease model for genotype relative risk. The absence of a signal here suggests that we may be underpowered when the underlying disease model deviates away from the log additive disease model for genotype relative risk.

We do observe a signal with GENECLUSTER for CD on chromosomes 3 and 10 but not large enough to exceed our significance threshold. Interestingly, using 2 basis clusters actually increases the Bayes Factors in both regions ($\log_{10} BF = 4.31$ for chromosome 3 and $\log_{10} BF = 4.34$ for chromosome 10) to above the threshold. The 2 basis cluster model and the single SNP test share a similar model of association, where case and control individuals are bipartitioned into 2 penetrance groups. This model is considerably simpler than the 1-mutation, 2-mutation and 5 basis cluster models. The Bayesian model contains a natural penalty against over-complex models, so if association between observed data can be supported by the simpler model then it will have a higher posterior probability, because its priors are more concentrated around the the observed data (Denison and Holmes, 2005), which might explain why our methods are underpowered here.

collection	chr	region (Mb)	1-mut model	2-mut model	5 basis clusters	Affy SNPs	imputed SNPs
BD	16	23.3-23.62	0.64	0.64	1.02	1.96	1.43
CD	3	49.3-49.87	3.66	3.7	3.31	4.24	4.22
CD	10	64.06-64.31	3.73	3.87	3.86	4.69	3.96

Table 6.3: Tier 1 signals that GENECLUSTER did not detect with a Bayes Factor threshold of 10^4. The Bayes Factors are obtained using TREESIM trees and are given on the $\log_{10}$ scale.

Meta-analyses of genome-wide association data, including the WTCCC data, have facilitated the discovery and replication of a large number of disease loci for CD and T2D. At the time of writing, there are 30 CD (Barrett et al., 2008) and 18 T2D loci (Zeggini et al., 2008) that have been replicated. We can therefore extend our comparisons to these CD and T2D loci, which are summarised in Figures 6.5 and 6.6 respectively. We have plotted the maximum $\log_{10}$ Bayes Factor observed at each replicated disease locus at an Affymetrix SNP on the x-axis, and on the y-axis we have plotted the maximum $\log_{10}$ Bayes Factor observed at an imputed SNP and using our methods at the same locus. We observe that at most loci the single SNP tests (at Affymetrix and imputed SNPs) outperforms our method. The disease models at these loci are presumably similar to those described above (chromosomes 3 and 10 for CD), where association can be captured by the simpler models used by SNP based approaches. There are 4 notable exceptions, however, on chromosomes 1 (the *IL23R* locus), 5 and 16 (*NOD2* locus) for CD, and chromosome 9 for T2D. They represent scenarios where our method can provide a boost in power and we will describe each of these loci in more detail in the subsequent sections of this chapter.

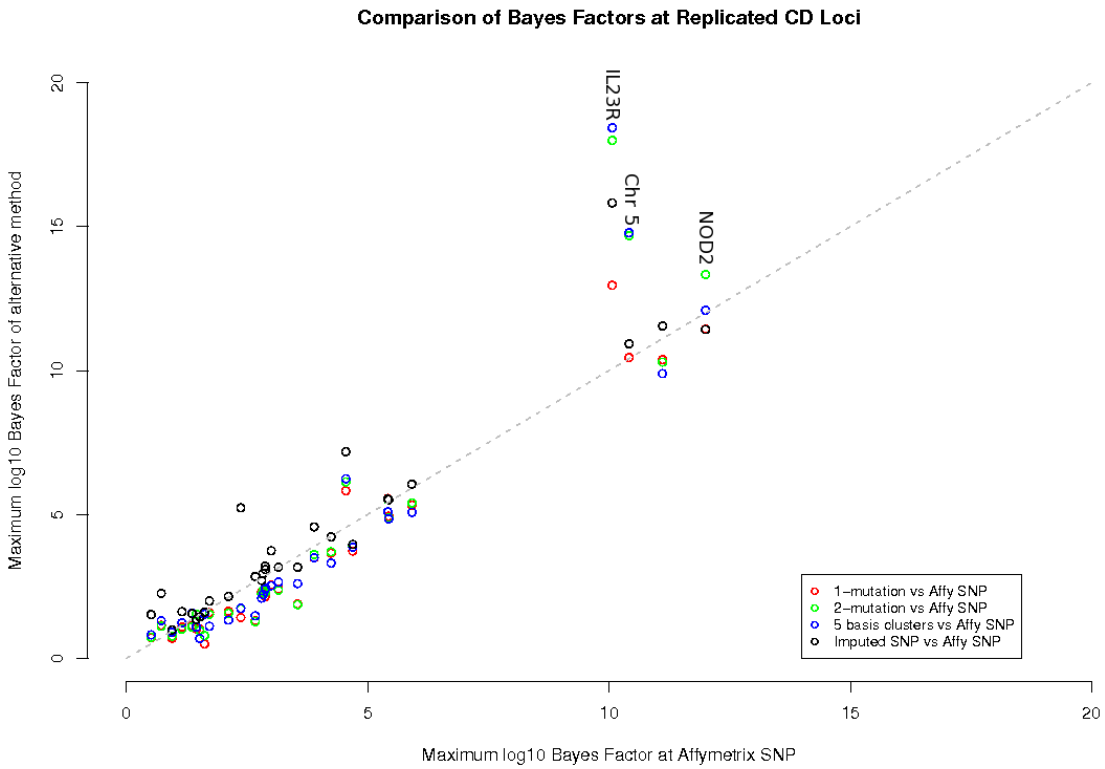


Figure 6.5: A comparison of $\log_{10}$ Bayes Factors at replicated CD loci. The x-axis is the maximum single SNP $\log_{10}$ Bayes Factor at each replicated CD locus and on the y-axis is the corresponding $\log_{10}$ Bayes Factor using imputation (black circles), 1-mutation model (red circles), 2-mutation model (green circles) and 5 basis clusters model (blue circles). We observe a boost in signal using the 2-mutation and 5 basis clusters models over the SNP based methods, at loci on chromosome 1 (*IL23R*), chromosome 5 and chromosome 16 (*NOD2*).

6.3.2 Detecting new signals of association

In this section we will take a closer examination of the 4 significant signals (coloured red in Table 6.1) that we have found, which were not reported by the WTCCC study. Cluster plots of the genotype calls at SNPs located in each locus were subjected to visual inspection. The results presented in this section are based on the analyses of the data after SNPs with bad cluster plots were removed.

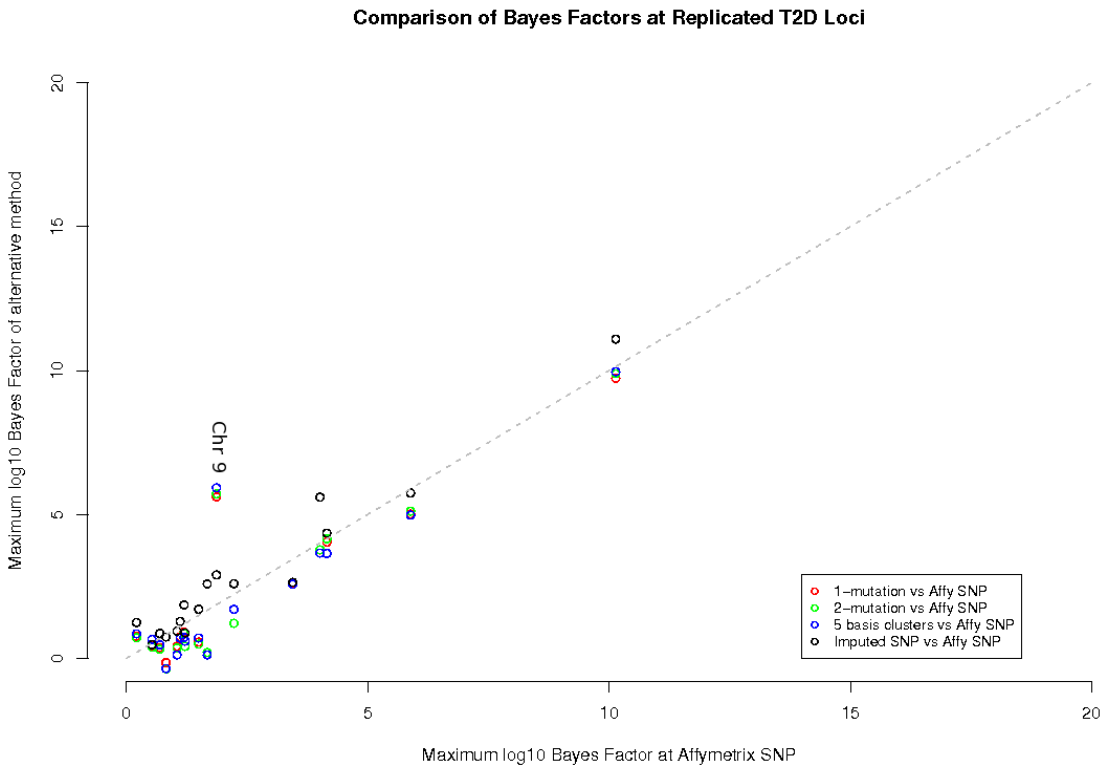


Figure 6.6: A comparison of $\log_{10}$ Bayes Factors at replicated CD loci. The x-axis is the maximum single SNP $\log_{10}$ Bayes Factor at each replicated T2D locus and on the y-axis is the corresponding $\log_{10}$ Bayes Factor using imputation (black circles), 1-mutation model (red circles), 2-mutation model (green circles) and 5 basis clusters model (blue circles). We observe a boost in signal using the 1-mutation, 2-mutation and 5 basis clusters models over the SNP based methods, on chromosome 9.

T2D: chromosome 9

The T2D signal on chromosome 9 resides within a 9Kb region flanked by recombination hot-spots. This locus was recently identified and confirmed by 3 independent T2D genome-wide association studies, which reported rs10811661 with the strongest signal of association (Zeggini et al., 2007; The Diabetes Genetics Initiative, 2007; Scott, L. J. et al., 2007). Each study followed a two-stage design, so that in the first stage signals are identified in a genome-wide scan and taken forward to stage 2 for replication using additional case and control subjects. Two studies (Zeggini et

al., 2007; The Diabetes Genetics Initiative, 2007) observed significant P values at rs10811611 in both stages (7.6×10^{-4} and 3.6×10^{-5} in stage 1, and 1.7×10^{-4} and 2.2×10^{-5} in stage 2) and meta-analysis of the pooled samples from all 3 studies, which comprised of 14,586 cases and 17,968 controls, yielded a P value of 7.8×10^{-15}. A second independent and weaker signal was detected at rs564398 100Kb away from rs10811661 (Zeggini et al., 2007).

Single SNP analyses of the WTCCC data revealed a moderate signal at rs10811611 of $\log_{10}$ Bayes Factor 1.53, which is the strongest within the 9Kb region flanking the recombination hotspots (stronger signals are located approximately 100Kb away but are likely to be related to the signal associated with rs564398). However, haplotype-based analysis using GENEBPM (Morris, 2006) provided Bayes Factors more than 4 orders of magnitude greater than the equivalent single SNP analyses and identified haplotypes sharing alleles T and T at rs10811661 and rs10757283 as high risk (Zeggini et al., 2007). The second stage replication study confirmed the association with that haplotype and combining the replication sample with the WTCCC sample provided very strong evidence of association (P value 2.9×10^{-11}). Analysis of CEU HapMap data has failed to identify any single SNP that can account for this haplotypic association, suggesting that the association may be explained by an as yet unidentified variant in the region.

Figure 6.7 summarises our results in the region of rs10811661 using the TREESIM tree estimated at the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model, which we will refer to as the focal position. The top left panel of the figure shows the $\log_{10}$ Bayes Factors across the region. The maximum $\log_{10}$ Bayes Factor peak at 5.61 and 5.71 for the 1-mutation and 2-mutation models and 5.92 for the 5 basis clusters model, which represent significant boosts in signal over the SNP

based approaches used by the WTCCC. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this plot. The bottom left panel shows the 120 CEU HapMap Phase 2 haplotypes, which was used as the reference panel across the region. Each row of this panel is a haplotype and each column is a SNP. The vertical blue dashed line indicates the focal position. The bottom right panel shows the TREESIM tree at the focal position. We refer to the best (fitting) 1 mutation (or single mutation) as the mutation on the branch b, which makes the largest contribution to the 1-mutation Bayes Factor, i.e. b with the largest $BF(x, b)\Pr(b|M_1, x, T_x, \bar{\Lambda}_x)$ as defined by Equation (5.22). Similarly, the best (fitting) 2 mutations are on branches $\mathbf{b}$ with the largest $BF(x, \mathbf{b})\Pr(\mathbf{b}|M_1, x, T_x, \bar{\Lambda}_x)$ in Equation (5.30). The best 1 and 2 mutations are indicated by the blue and red/green dots on the tree, respectively. The top right panel shows the tables of expected allele counts for the best fitting mutations, together with a summary of the Bayes Factors (averaged over all possible mutations) that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel. The bottom column of each table indicates an estimate, based on the expected allele counts, of the haplotype risk conferred by each best fitting mutation relative to the haplotype without a mutation.

In the bottom left plot of Figure 6.7 we have coloured the panel haplotypes to indicate the 3 haplotypes induced by the SNPs rs10811661 and rs10757283. The best fitting single mutation (blue dot) and one of the best fitting 2 mutations (red dot) exactly identifies all but one of the high risk TT haplotypes (coloured purple). The one TT haplotype that was is identified by the red mutation falls under an adjacent branch and is grouped with the CT haplotypes (coloured cyan) identified by the other best fitting 2 mutation (green dot). There is no single branch in

the estimated genealogical tree that can encompass all of the TT haplotypes, so with respect to identifying the high risk haplotypes GENECLUSTER has chosen the more appropriate mutations. The relative risk estimate indicates that the TT haplotype confers high risk (the estimated relative risk for haplotypes carrying the blue mutation on the tree, relative to haplotypes not carrying that mutation, is 1.33) and is driving the GENECLUSTER signal in at this locus.

Figure 6.8 summarises a similar result using the estimated LSTREE tree at the same focal SNP. This tree has a branch that encompasses all of the TT haplotypes, which GENECLUSTER has identified as the best fitting single mutation and one of the best fitting 2 mutations.

The disease model underlying this T2D locus is the realisation of a scenario where our methods provide a boost in power. Given that the signals from SNP based approaches are weak, it is likely that the causal variant is weakly correlated with nearby Affymetrix or imputed (HapMap) SNPs. However, we are able to detect a signal from the structure of the local haplotypes, which could have acted as a surrogate for the causal variant.

T1D: chromosome 15

Figure 6.9 summarises our results for T1D at 15q22.2. This is a highly conserved region, in which no signal has previously been found for T1D. Our signal is located in the *RORA* gene, which encodes ROR, an evolutionarily related transcription factor and belongs to the steroid hormone receptor superfamily. There is evidence that *RORA* is involved in cellular stress response and human tumorigenesis (Zhu et al., 2006), and also immunomodulatory activities (Missbach et al., 1996), which would

make *RORA* a candidate gene for autoimmune diseases such as T1D.

The signals from the single SNP tests are weak in this region (peaks of 1.08 and 1.98 at Affymetrix and imputed SNPs respectively). The signals from the 2-mutation model (4.06) and the 5 basis clusters model (4.10) are significantly stronger than the 1-mutation model (3.20), which suggest that multiple causal variants are involved (the posterior probability for the 2-mutation model, compared against the 1-mutation model, is 0.88). The best fitting 2 mutations (represented by the green and red dots on the tree at the bottom right) delineates the panel haplotypes (bottom left) into 3 groups of distinct haplotype structure around the focal position, which we have highlighted in the bottom left plot with the colours cyan, yellow and pink.

The signals using LSTREE trees are a little weaker with the 1-mutation, 2-mutation models peaking at 2.98 and 3.90, respectively. On inspection of the signal plot in Figure 6.10 we find that the haplotype groups delineated by best fitting 2 mutations closely correspond to the 3 groups identified with the TREESIM tree. This suggests that the same haplotype groups are driving the signal in the TREESIM and LSTREE analyses, which comprises of a deleterious group (cyan) and a protective group (pink and some yellow).

In addition to the T2D locus on chromosome 9 described earlier, this T1D locus, if confirmed, is another representation of a scenario (possibly involving allelic heterogeneity) where our methods outperform the SNP based approaches. We are in the process of carrying out a replication study to confirm this locus.

BD: chromosome 13

Figure 6.11 summaries the signal found for BD at 13q31.1 using TREESIM trees. The peak signal is close to the *NDFIP2* gene, which is involved in endocytosis (Shearwin-Whyatt et al., 2004; Konstas et al., 2002) but is not linked to BD as far as we are aware.

The signal from the single SNP tests are very weak in this region (peaks of 1.10 and 1.26 at Affymetrix and imputed SNPs, respectively). There is moderate signal from the 1-mutation model (3.53) and 2-mutation model (3.50) but only the 5 basis clusters model (4.09) produces a signal greater than a threshold of 4.0. Results of using 2-10 basis clusters in the region, which are shown in Table 6.4, reveal more moderate signals when 6-10 basis clusters are used. We observe similar signals using LSTREE trees but they are weaker than those obtained using TREESIM trees.

Number of Basis Clusters	2	3	4	5	6	7	8	9	10
Max $\log_{10}$ Bayes Factor	1.54	1.15	2.87	4.09	3.79	3.56	3.46	3.21	3.09

Table 6.4: The $\log_{10}$ Bayes Factors under Model B at a region on chromosome 13 of the BD analysis.

Comparison between the mutation models (Model A) and basis clusters models (Model B) have been discussed in the previous chapter, which lead to two possible explanations for the stronger signal with Model B. The first is that using 5 (or more) basis clusters model allows up to 5 (or more) risk parameters, so if there are more than two causal variants at this locus then it can be a more powerful approach. However, analysing the focal SNP using a 3-mutation model at the focal SNP produces a $\log_{10}$ Bayes Factor of 3.20, which is lower than the 1-mutation and 2-mutation models and therefore suggests against this explanation. The second explanation is that Model B integrates over the different topologies in the upper

levels of the estimated tree and therefore accounts for its uncertainty. One way to test this is to repeat our analysis but using a sample of estimated trees to see if we observe an increase in signal with the 1-mutation and 2-mutation models. Unfortunately, due to time constraints this was not possible before the submission of this thesis.

CD: chromosome 6

Figure 6.12 summarises our results for CD in the HLA region of chromosome 6. There are numerous genes in this region related to immune function, which are candidates genes for autoimmune diseases such as CD. However, it is unclear if this is a genuine new association since the WTCCC reported a signal only 1.5Mb away and this is a region that contains extensive LD.

6.3.3 Detecting regions of allelic heterogeneity

There are 4 signals in Table 6.3 where the 2-mutation model provides a much better model fit than the 1-mutation model (a difference of at least 1.0 in $\log_{10}$ Bayes Factors between the 2-mutation and 1-mutation models). As we demonstrated by simulation in Section 5.2.5, this is an indication of allelic heterogeneity. We have summarised the 4 signals in Table 6.5.

Each locus in Table 6.5 is an established disease locus with evidence of allelic heterogeneity (Undliem et al., 2001; Weyand et al., 1995; Duerr et al., 2006; Libioulle et al., 2007; Hugot et al., 2001; Ogura, Y. et al., 2001). The CD loci on chromosome 1 (*NOD2*) and chromosome 16 (*IL23R*), and chromosome 5 to some ex-

collection	chr	region (Mb)	BF_{12}	Post. prob. 2 mutations
CD	1	67.25-67.47	5.03	1.00
CD	5	40.33-40.65	4.23	1.00
CD	16	49.16-49.44	1.89	0.987
RA	6	29.66-33.77	21.75	1.00
T1D	6	25.98-33.93	> 9	1.00

Table 6.5: Regions where GENECLUSTER results indicate the presence of more than 1 mutation. The fourth column gives the posterior probability of the 2-mutation model versus the 1-mutation model assuming prior odds of 1:1.

tent, are well-characterised, which allow us to further evaluate the performance of GENECLUSTER.

IL23R

The *IL23R* locus on chromosome 1 is an established disease locus for Crohn's disease with extensive allelic heterogeneity (Duerr et al., 2006). A plot showing the results of our method using TREESIM trees in this region is given in Figure 6.13. The maximum $\log_{10}$ Bayes Factors are 12.96, 17.99 and 18.43 for the 1-mutation, 2-mutation and 5 basis clusters models respectively, which compare favourably with the best Affymetrix SNP (10.07) and the best imputed SNP (15.82). The difference between the 2-mutation and 1-mutation Bayes Factors implies a posterior probability of 1.00 for the 2-mutations model, indicating evidence of allelic heterogeneity as described in Section 5.2.5.

The original paper identified two SNPs in functional regions of the *IL23R* gene. The first SNP (rs11209026) is the non-synonymous SNP (c.1142G>A, p.Arg381Gln) identified as the strongest signal in the original study. The second SNP (rs10889677) is in the 3' UTR of the *IL23R* gene and the only other associated non-intronic SNP

found in the original study. When we look at these two SNPs in the CEU HapMap panel we identify 3 distinct haplotypes coloured green, purple and cyan in Figure 6.14. These haplotypes are almost precisely those that are delineated by the best fitting 2 mutations (green and red dots). One of the mutations on the tree (red dot) identifies all the CEU HapMap haplotypes that carry the A allele at rs11209026 (haplotypes coloured cyan) and the second mutation (green dot) identifies all but one of the haplotypes that contain the A allele at rs10889677 (haplotypes coloured green). Relative risk estimates for haplotypes carrying the red and green mutations on the tree, relative to haplotypes carrying neither mutations, are 0.39 and 1.29 respectively. These risk estimates are consistent with the A allele at rs11209026 being protective as reported by Duerr et al. (2006)

Our analysis using LSTREE trees (Figure 6.14) is very similar. One of the best 2 mutations (green dot) precisely identifies all panel haplotypes carrying the deleterious A allele at rs10889677 (haplotypes coloured green). The second mutation (red dot) identifies all haplotypes carrying the A allele at rs11209026 (haplotypes coloured cyan) but the tree does not contain a branch that allows their precise identification and a haplotype (coloured purple) carrying the GC alleles at rs11209026 and rs10889677 is included in the same group.

NOD2

We have detected strong signals of association in the *NOD2* locus for CD, which is an established CD locus with extensive allelic heterogeneity (Hugot et al., 2001; Ogura, Y. et al., 2001).

Our results are illustrated in Figures 6.15 (using TREESIM trees) and 6.16 (using

LSTREE trees). All the methods show a substantial signal at the locus but the signal for our methods are higher and broader. The $\log_{10}$ Bayes Factors allowing for 1-mutation, 2-mutation and 5 basis cluster models peak at 11.44, 13.33 and 12.09 respectively. These compare favourably with the $\log_{10}$ Bayes Factors at the best Affymetrix SNP (12.00) and the best imputed SNP (11.42). The $\log_{10}$ Bayes Factor for a 2-mutation model compared with a 1-mutation model is 1.89, which implies a posterior probability of 0.987 for the 2-mutation model, indicating allelic heterogeneity.

There are 3 known coding SNPs in this region (rs2066844, rs2066845 and rs2066847) and two of them (rs2066844 and rs2066845) are in the HapMap Phase 2 marker set. We have coloured the panel haplotypes carrying the CC and TG alleles at the two SNPs yellow and blue, respectively. The haplotypes carrying the CG alleles are coloured cyan and pink. As before, our signal plots indicate the best fitting mutation (blue dot) and 2 mutations (red and green dots) that makes the greatest contribution to the 1-mutation and 2-mutation model Bayes Factors. We have also implemented the 3-mutation model at the focal SNP and have indicated the best 3 mutations (red, green and blue triangles) that make the largest contribution to the 3-mutation model Bayes Factor.

Focusing on the LSTREE trees signal plot in Figure 6.16, we find that one of the best fitting 2 mutations (green dot) precisely identifies the TG haplotypes (coloured blue). Interestingly, the other best fitting mutation (red dot) lies on a branch that identifies the group of pink haplotypes carrying the CG alleles as high risk, and is not on the branch that identifies the yellow CC haplotypes, which we know are high risk haplotypes. For brevity, let us denote the branch identified by the red dot as b_1, the branch encompassing the CC haplotypes as b_2 and the group of (pink) haplotypes

identified by the mutation on b_1 as $\mathbf{H_1}$. It appears that although b_1 is more than 6 times shorter than b_2, the excess of case haplotypes mapped to $\mathbf{H_1}$ means that there is more evidence for selecting b_1 as one of the best fitting 2 mutations. The 3-mutation model yields a $\log_{10}$ Bayes Factor of 16.28, which supports the 3-mutation model (posterior probability of 1) over either the 1-mutation or 2-mutation models. The best fitting 3 mutations identify perfectly the 3 groups of haplotypes defined by the two coding SNPs in addition to $\mathbf{H_1}$. We have exhaustively searched, without success, for a SNP that delineates $\mathbf{H_1}$ from the other CG (cyan) haplotypes. These results suggest that $\mathbf{H_1}$ are high risk haplotypes and are possibly associated with another causal variant at this locus and not in the HapMap marker set. It would be very interesting to find out, should the resources become available, the association between $\mathbf{H_1}$ and rs2066847 (the causal variant identified for this locus that is not in the HapMap marker set).

The best fitting 2 mutations on the TREESIM tree is displayed in Figure 6.15. One mutation precisely identifies the yellow CC haplotypes in the panel. However, the second mutation, nor any of the best 3 mutations, are able to delineate the TG haplotypes from the group of haplotypes, which we denote $\mathbf{H_2}$, comprising of the TG haplotypes (haplotypes coloured blue), haplotypes in $\mathbf{H_1}$ (haplotypes coloured pink) and other CG haplotypes (haplotypes coloured cyan). The Bayesian model naturally penalises against over-complex models, and in this case it appears that the simpler model of fitting a single mutation to the excess of cases mapped to $\mathbf{H_2}$ is more favourable than the more complex model of fitting 2 mutations to delineate the TG and $\mathbf{H_1}$ haplotypes from each other in $\mathbf{H_2}$.

The TREESIM analysis here illustrates a potential weakness of the mutation model: if two groups of panel haplotypes, each carrying independent causal variants but

exhibiting similar risks, are placed close together under the estimated tree, then GENECLUSTER may be underpowered to identify both groups as independently associated with disease. In this instance, GENECLUSTER will be underpowered to detect allelic heterogeneity. For example, under the 3-mutation model the $\log_{10}$ Bayes Factor, using TREESIM, is 13.42 and therefore it is only marginally more favourable than the 2-mutation model.

CD: Chromosome 5

The third Crohn's disease signal is located within an approximately 250Kb region on chromosome 5, flanked by recombination hotspots. Numerous SNPs within this region have been identified and replicated (P values up to 10^{-12}) (Libioulle et al., 2007). The LD structure separates this region into 5 LD blocks and the strongest associations (single SNP and haplotype) are found in a central 122Kb block. However, multivariate haplotype analysis conditional on the effect of the central block showed that the two flanking LD blocks remain significantly associated, which suggests that multiple variants in the region may account for the observed effects on CD.

Single SNP analysis in the WTCCC data reveals strong associations at both Affymetrix and imputed SNPs (maximum $\log_{10}$ Bayes Factors are 10.41 and 10.92, respectively). Figure 6.17 illustrates the results of our analysis. The 2-mutation model provides a large boost in signal (14.68) and compared to the 1-mutation model (10.45) indicate allelic heterogeneity at this locus (posterior probability for the 2-mutation model is 1.0). Further, the best 2 mutations appear to delineate the panel haplotypes into 3 groups (coloured cyan, yellow and purple in the signal plot) with distinct LD pattern approximately 100Kb either side of the focal position at 40430000.

Analysis with LSTREE trees, in Figure 6.18, yields a similar result. The best fitting 2 mutations identifies the same 3 groups identified in the TREESIM analysis and therefore suggests that they are driving the signal at this locus. However, the Bayes Factor using LSTREE (15.09 for the 2-mutation model) is greater than using TREESIM, partly because the best fitting mutations lie on longer branches.

The 3-mutation model, which yields a $\log_{10}$ Bayes Factor of 14.43 and 15.18 with TREESIM and LSTREE trees respectively, do not provide strong indications for more than two causal variants

6.4 Discussion

As stated in the beginning of this chapter we had two goals:

1 to provide an assessment of our approach applied to real population data;

2 to uncover novel disease loci.

We will now discuss in turn the evidence we have gathered with respect to both of them.

With respect to the first goal, we have demonstrated that we are able to apply GENECLUSTER to genome-wide association data using a multinode computer. We have found signals of association that correspond well to the results of the WTCCC study, which provide validation for the assumptions that we made in Section 5.1 to devise our method.

We have provided a comparison of our method with the SNP based approaches that were used in the WTCCC study, which is reflective of approaches currently used in most association studies. It appears that for most disease loci, SNP based approaches in general produce a larger Bayes Factor. The successes of SNP based approaches may be the result of using a simpler model of association. However we also found loci where our method outperforms the SNP based approaches. The disease model at these loci can be largely classified into two categories:

- the causal SNP is not in the HapMap marker set, so it can not be imputed, and is not in strong LD with any of the Affymetrix or imputed (HapMap) SNPs, but it is correlated with local haplotypes (T2D locus on chromosome 9);
- the presence of multiple causal variants at the same locus resulting in allelic heterogeneity (CD loci at *IL23R* on chromosome 1, *NOD2* on chromosome 16 and a gene desert in chromosome 5).

The relative strengths of the mutation model, Model A, to the basis cluster model, Model B, have been discussed in the last chapter and in this chapter we were able to assess the practicality of applying them to real population data. There appears to be a clear correlation between the signals reported by both approaches in Tables 6.1 and 6.2, and Figures 6.5 and 6.6, and there does not appear to be a most powerful approach.

A problem with the current implementation of Model B is that it is difficult to know how many basis clusters to use. The BD locus on chromosome 13 is a good example where signal is observed using a number of basis clusters but the signal only exceeds our threshold when using 5 basis clusters. It is not clear how this should be

interpreted and this is something we will discuss in the final chapter about how our methods can be improved.

In contrast, Model A provides a natural interpretation of the underlying disease model at a disease locus. The difference in the 1-mutation model and 2-mutation model Bayes Factors is a natural way to summarise the evidence for allelic heterogeneity and we have correctly predicted regions with allelic heterogeneity this way. In addition, without any prior knowledge of the disease model, the best fitting mutations have correctly identified the branches that induce the distinct haplotype backgrounds in the panel haplotypes, which harbour the causal alleles. This is a further validation of our method and suggests that the haplotype trees, which we use to perform inference, retain important features of the true genealogies of the case-control sample for the purposes of detecting association. This added functionality of GENECLUSTER may prove to be useful for fine-mapping studies, where the aim is to identify the actual causal variant in a region after a signal of association has been detected. We will explore this further in the next chapter. For the above reasons, we would recommend using Model A in the current implementation of GENECLUSTER for genome-wide association scans.

There is a close correspondence between our analysis using the two sets of trees estimated by TREESIM and LSTREE. Our signal plots show that the best fitting mutations identify similar groups of haplotypes in the panel, which are responsible for driving the signal at a disease locus. Variation in Bayes Factors between the two sets of analyses are caused by differences in the tree topology, mostly in the lower levels, and in branch lengths. It is unclear which set of trees is the best to use, and at this stage we suggest that both are used in tandem since combining the results from both sets of analyses can uncover further information, as demonstrated by our

analysis of the *NOD2* locus for CD.

Our second goal of finding possible novel associations is a challenging one. Since the WTCCC study, a number of large meta-analyses on T2D and CD, using sample sizes up to ten thousand individuals, have been conducted and have found numerous new disease loci. It is therefore unclear how many new signals can still be extracted from the WTCCC data. As an example, without the meta-analysis on T2D, we would have been able to report the T2D locus on chromosome 9 as novel. Nevertheless, we have identified a T1D locus on chromosome 15 that is potentially novel. We are further encouraged by the fact that it lies within a candidate gene. The discoveries of these novel signals in the WTCCC data lead us to believe that there is great value in performing genome-wide scans for association using our method.

For optimal results GENECLUSTER requires a haplotype reference panel of sufficient size to fully represent the local genetic variation of the case-control populations. The marker resolution should be sufficient to allow the accurate estimations of the panel's genealogy across the genome. At the moment the most suitable panel, for genome-wide scans, is the HapMap haplotypes. Our results here demonstrate that the HapMap Phase 2 CEU haplotypes can be successfully used for genome-wide association studies involving subjects of Caucasian European descent. Chinese+Japanese and Yoruban panels are also available, which should be applicable to studies involving subjects closely related to one of those populations. Panels of both greater size (such as HapMap Phase 3) and finer marker resolution (such as the 1000 Genome Project), are becoming available, which can only make GENECLUSTER more powerful in the future.

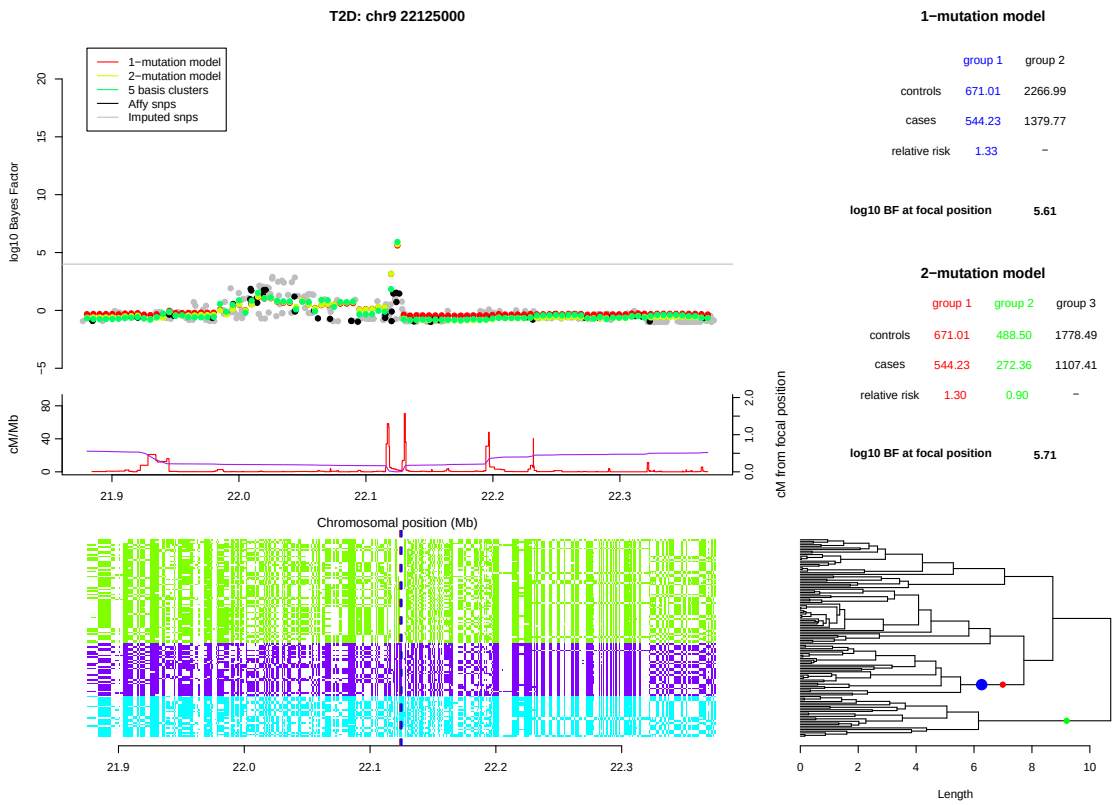


Figure 6.7: GENECLUSTER analysis of T2D on chromosome 9 using TREESIM trees. The top left panel of the plot shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within a chromosome 9 region of the T2D analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 haplotypes that occur at the SNPs rs10811661 and rs10757283 (cyan=CT, purple=TT, green = TC). The dashed vertical blue and brown lines indicate the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position) and those two SNPs, respectively, although on this scale the 3 SNPs are too close together to distinguish. The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

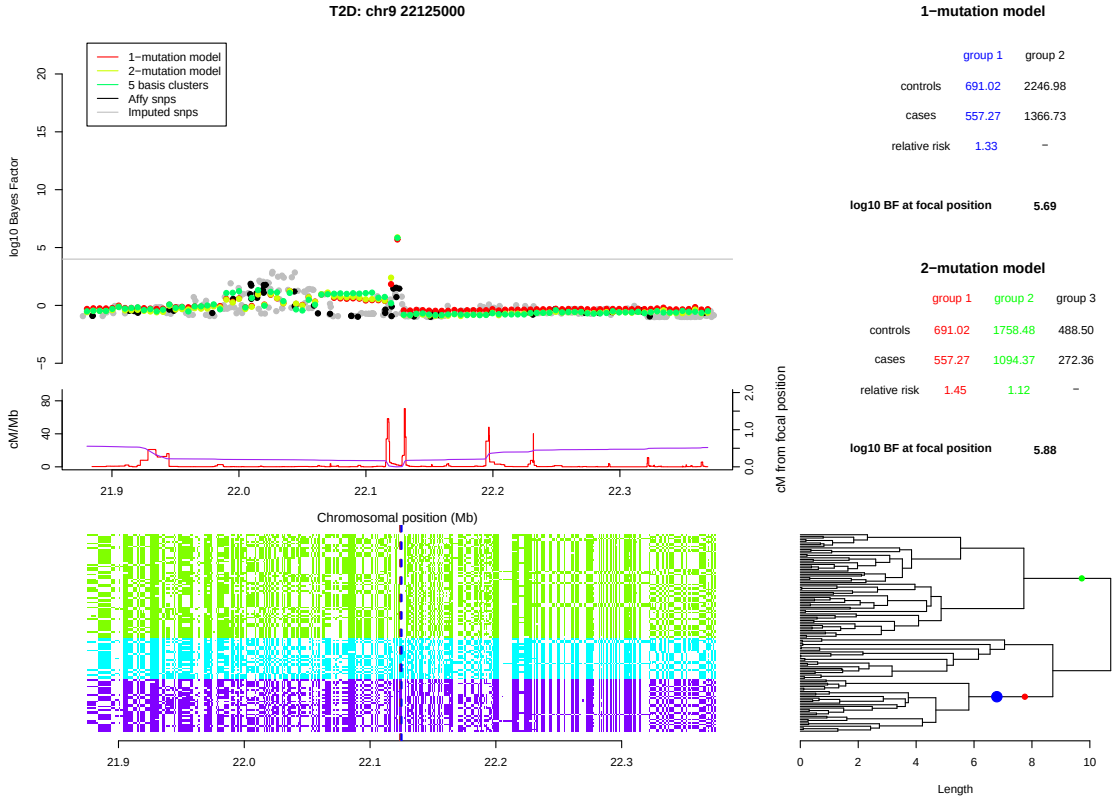


Figure 6.8: GENECLUSTER analysis of T2D on chromosome 9 using LSTREE trees. The top left panel of the plot shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within a chromosome 9 region of the T2D analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 haplotypes that occur at the SNPs rs10811661 and rs10757283 (cyan=CT, purple=TT, green = TC). The dashed vertical blue and brown lines indicate the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position) and those two SNPs, respectively, although on this scale the 3 SNPs are too close together to distinguish. The bottom right panel shows the estimated LSTREE tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

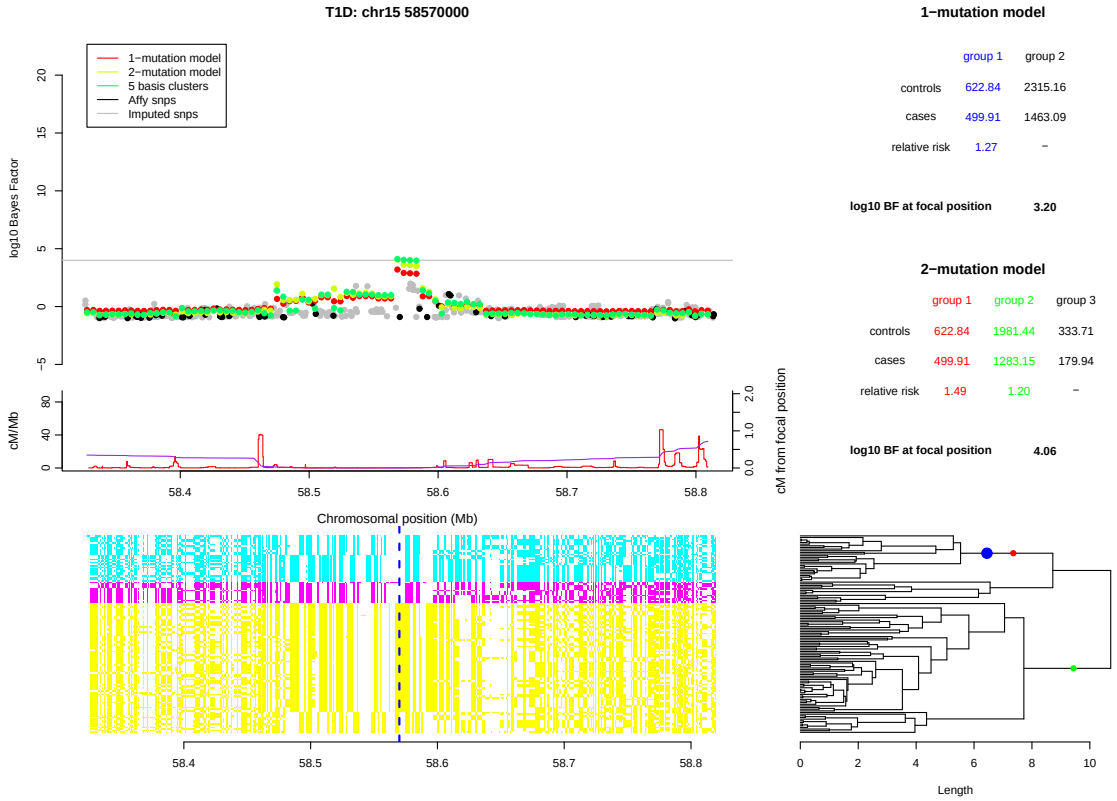


Figure 6.9: GENECLUSTER analysis of T1D on chromosome 15 using TREESIM trees. The top left panel of the plot shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within a chromosome 15 region of the T1D analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom right panel shows the estimated TREESIM tree at the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position). The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 groups defined by the best fitting 2 mutations (green and red dots). The dashed vertical blue line indicate the focal position.

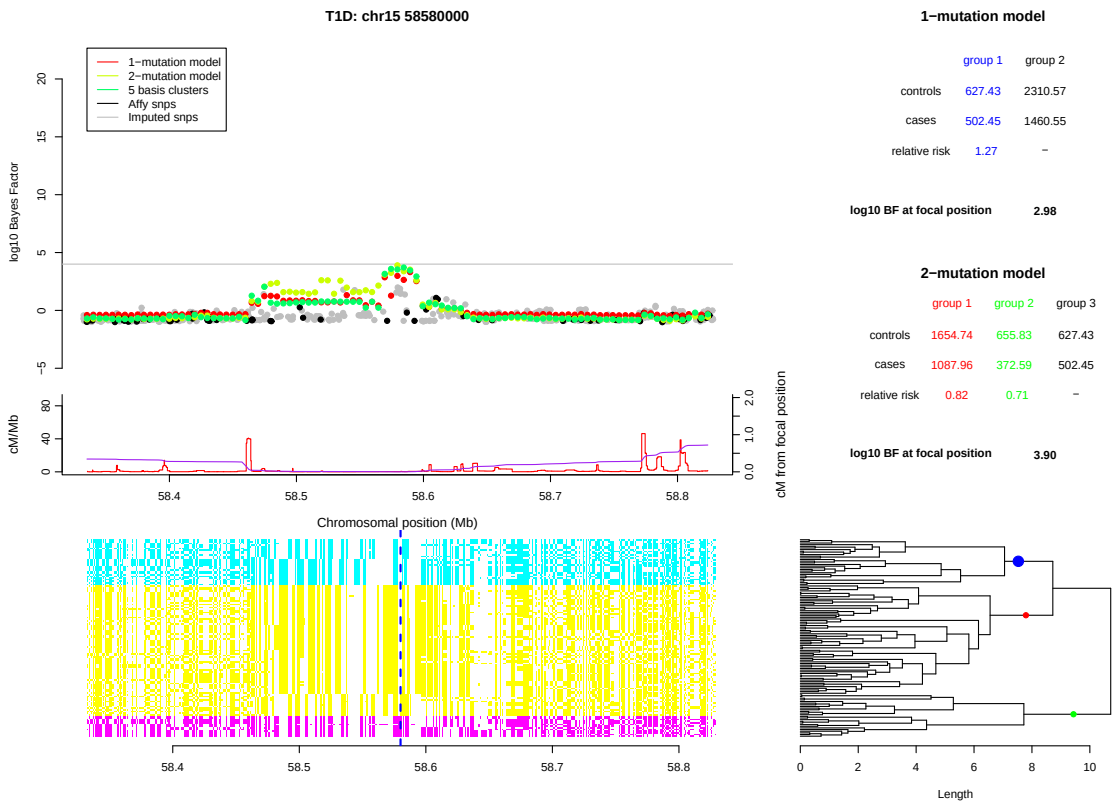


Figure 6.10: GENECLUSTER analysis of T1D on chromosome 15 using LSTREE trees. The top left panel of the plot shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within a chromosome 15 region of the T1D analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom right panel shows the estimated LSTREE tree at the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position). The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 groups identified by the best 2 mutations on the TREESIM tree at the same location (green and red dots in Figure 6.9), which corresponds closely the 3 groups delineated by the best 2 mutations on the LSTREE tree shown on the bottom right plot (green and red dots). The dashed vertical blue line indicate the focal position.

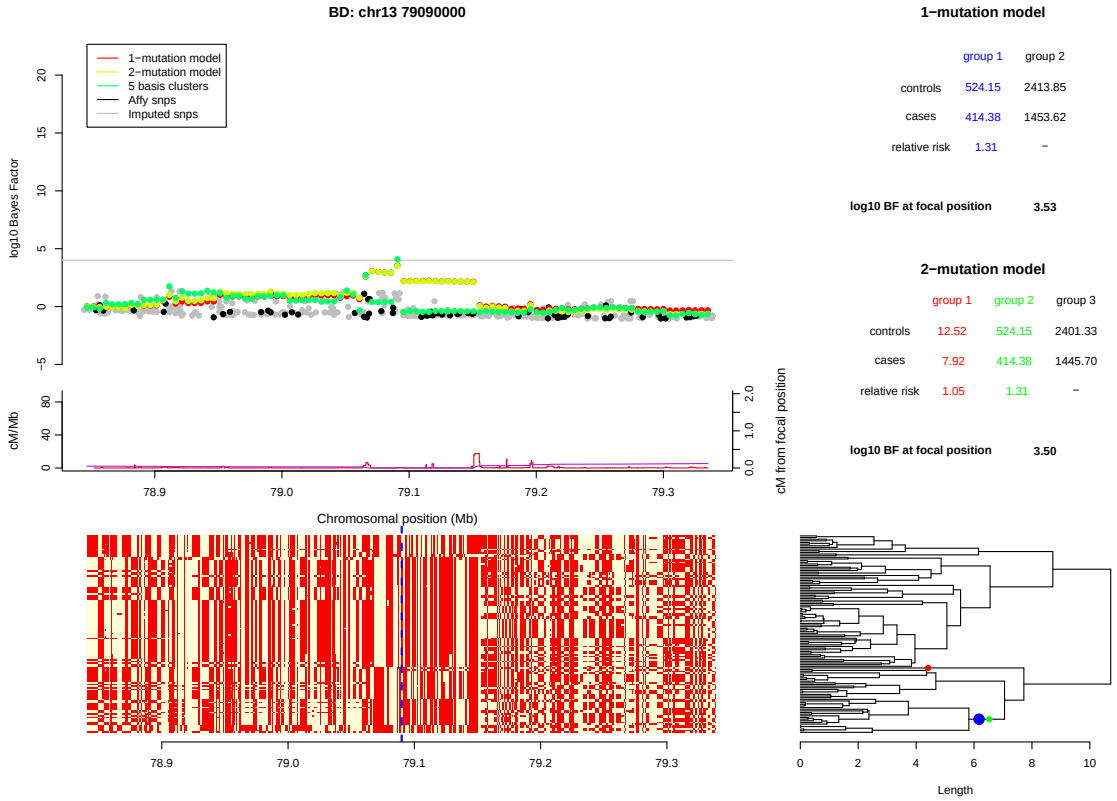


Figure 6.11: GENECLUSTER analysis of BD on chromosome 13 using TREESIM trees. The top left panel of the plot shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within a chromosome 13 region of the BD analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured red and beige to represent the two allele types at each SNP. The dashed vertical blue line indicate the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position). The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

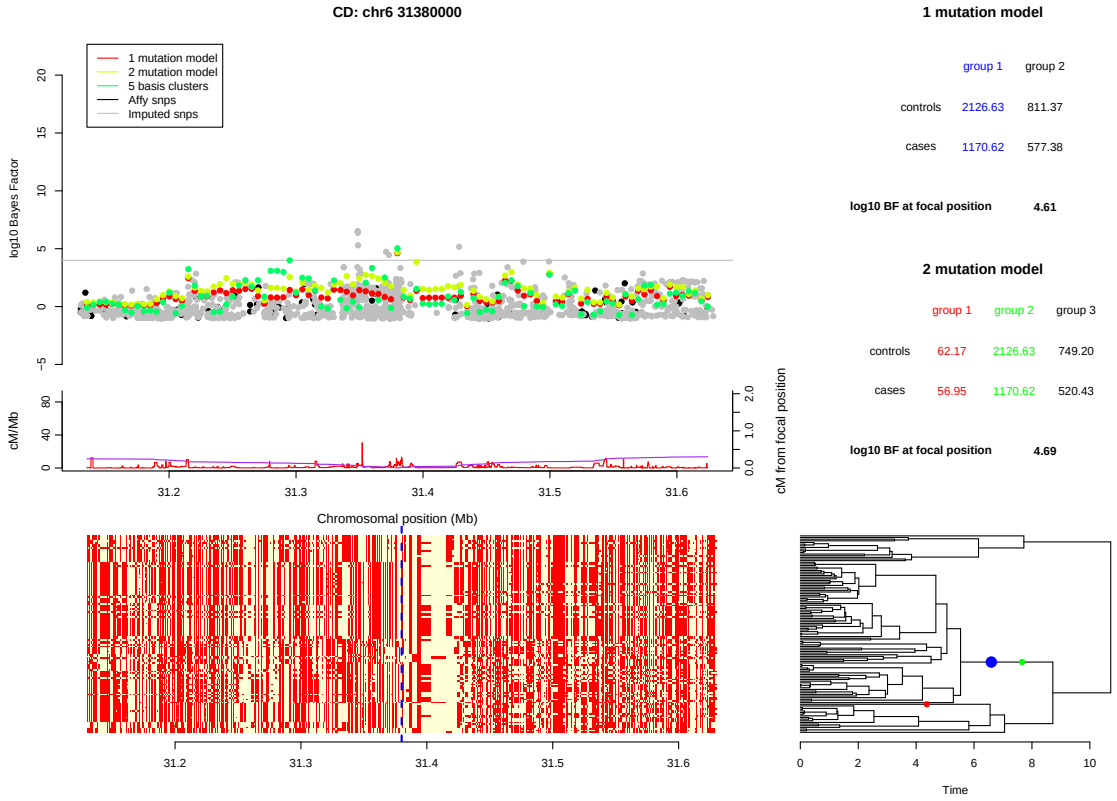


Figure 6.12: GENECLUSTER analysis of CD on chromosome 6 using TREESIM trees. The top left panel of the figure shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within a chromosome 6 region of the CD analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured red and beige to represent the two allele types at each SNP. The dashed vertical blue line indicate the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position). The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

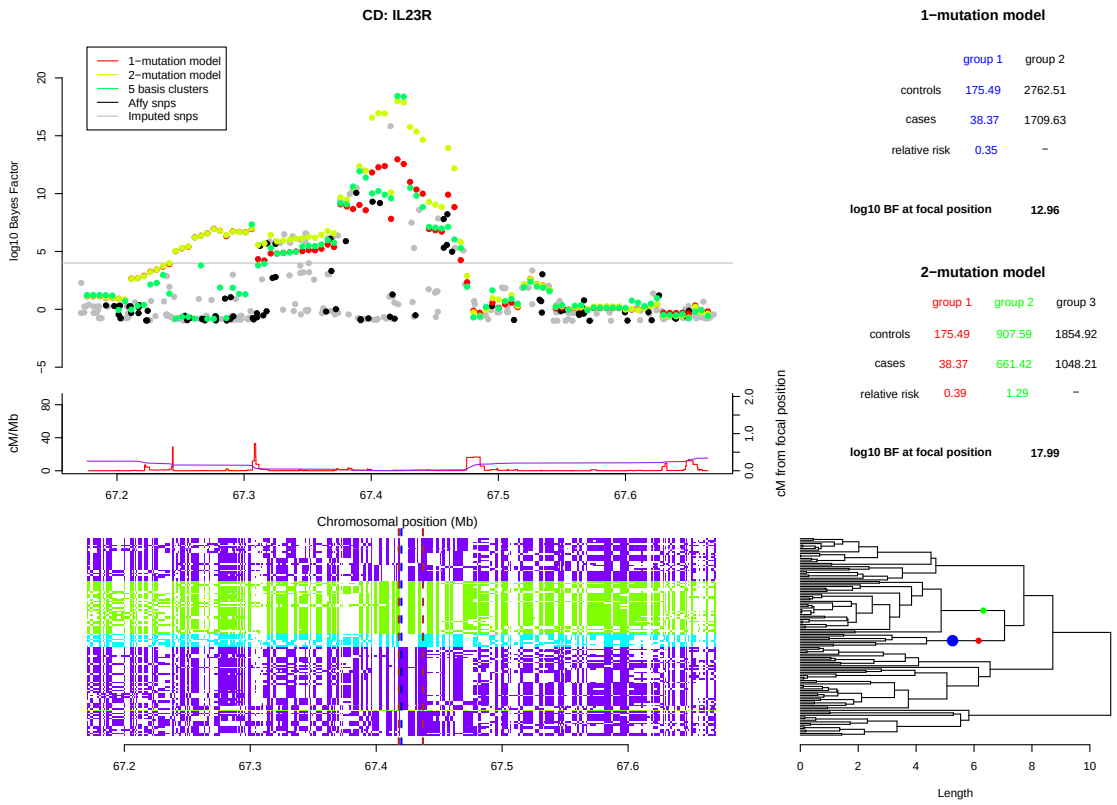


Figure 6.13: GENECLUSTER analysis of CD at the *IL23R* locus using TREESIM trees. The top left panel of the figure shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within the *IL23R* locus of the CD analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 haplotypes that occur at the two coding SNPs rs11209026 and rs10889677 (cyan=AC, purple=GC, green = GA). The dashed vertical blue and brown lines indicate the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position) and the two coding SNPs respectively. The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

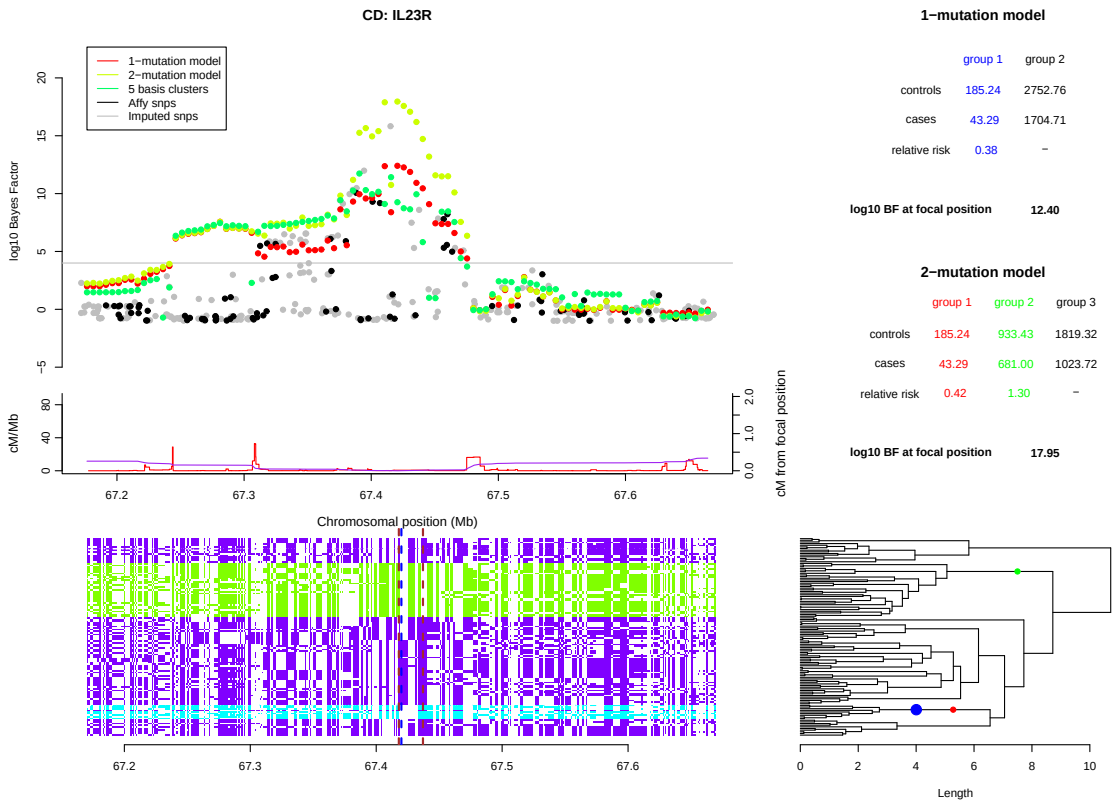


Figure 6.14: GENECLUSTER analysis of CD at the *IL23R* locus using LSTREE trees. The top left panel of the figure shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within the *IL23R* locus of the CD analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 haplotypes that occur at the two coding SNPs rs11209026 and rs10889677 (cyan=AC, purple=GC, green = GA). The dashed vertical blue and brown lines indicate the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position) and the two coding SNPs respectively. The bottom right panel shows the estimated LSTREE tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

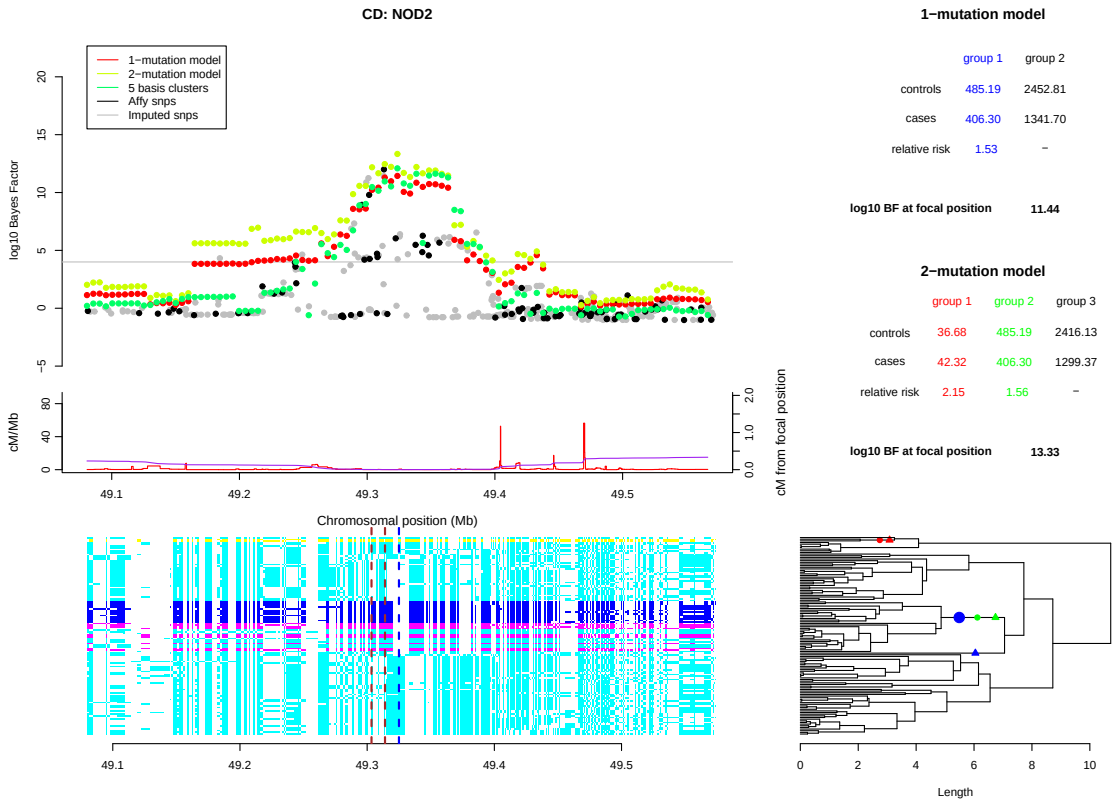


Figure 6.15: GENECLUSTER analysis of CD at the *NOD2* locus using TREESIM trees. The top left panel of the figure shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within the *NOD2* locus of the CD analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The haplotypes are coloured to indicate the 3 haplotypes that occur at the two coding SNPs rs2066844 and rs2066845 (yellow=CC, blue=TG, cyan=CG, pink=CG). Haplotypes coloured pink are identified as high risk in the corresponding analysis with a LSTREE tree. The dashed vertical blue and brown lines indicate the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position) and the two coding SNPs respectively. The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. We have also implemented the 3-mutation model for this region and the best fitting 3 mutations are indicated by the red, blue and green triangles. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

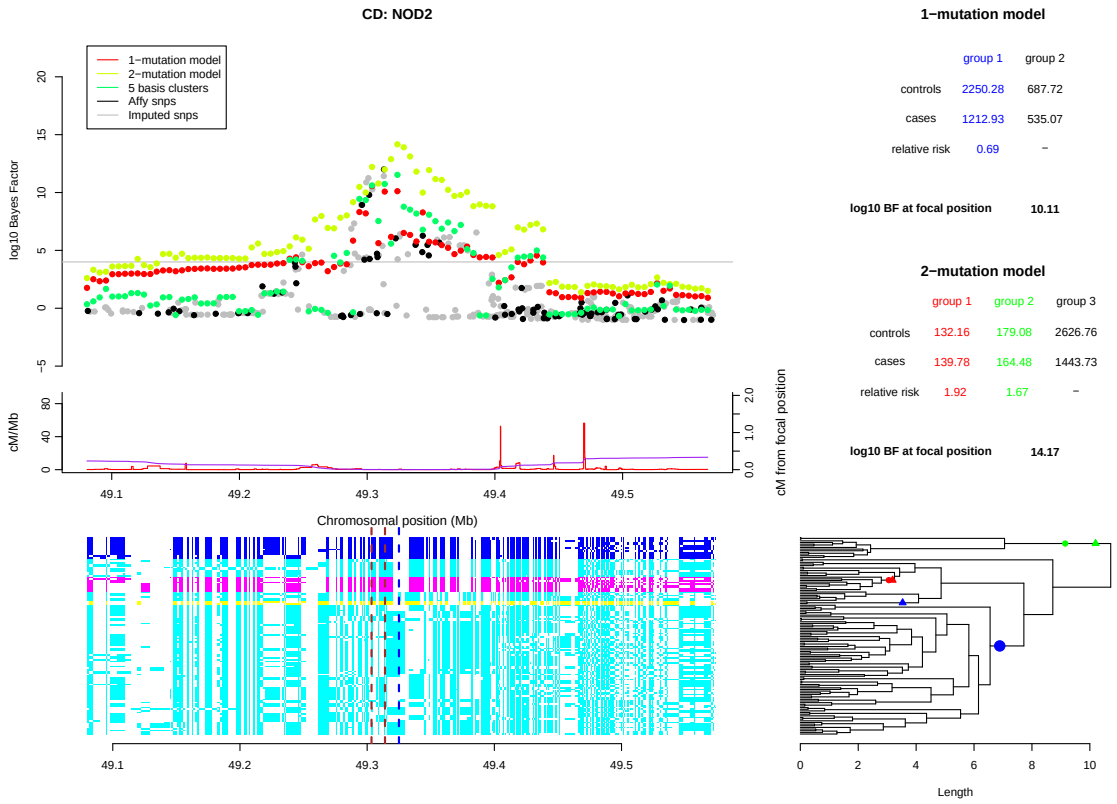


Figure 6.16: GENECLUSTER analysis of CD at the *NOD2* locus using LSTREE trees. The top left panel of the figure shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within the *NOD2* locus of the CD analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The haplotypes are coloured to indicate the 3 haplotypes that occur at the two coding SNPs rs2066844 and rs2066845 (yellow=CC, blue=TG, cyan=CG, pink=CG). Haplotypes coloured pink indicate a group of 9 haplotypes under one of the best fitting 2-mutations (indicated by the red dot). The dashed vertical blue and brown lines indicate the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position) and the two coding SNPs respectively. The bottom right panel shows the estimated LSTREE tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. We have also implemented the 3-mutation model for this region and the best fitting 3 mutations are indicated by the red, blue and green triangles. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

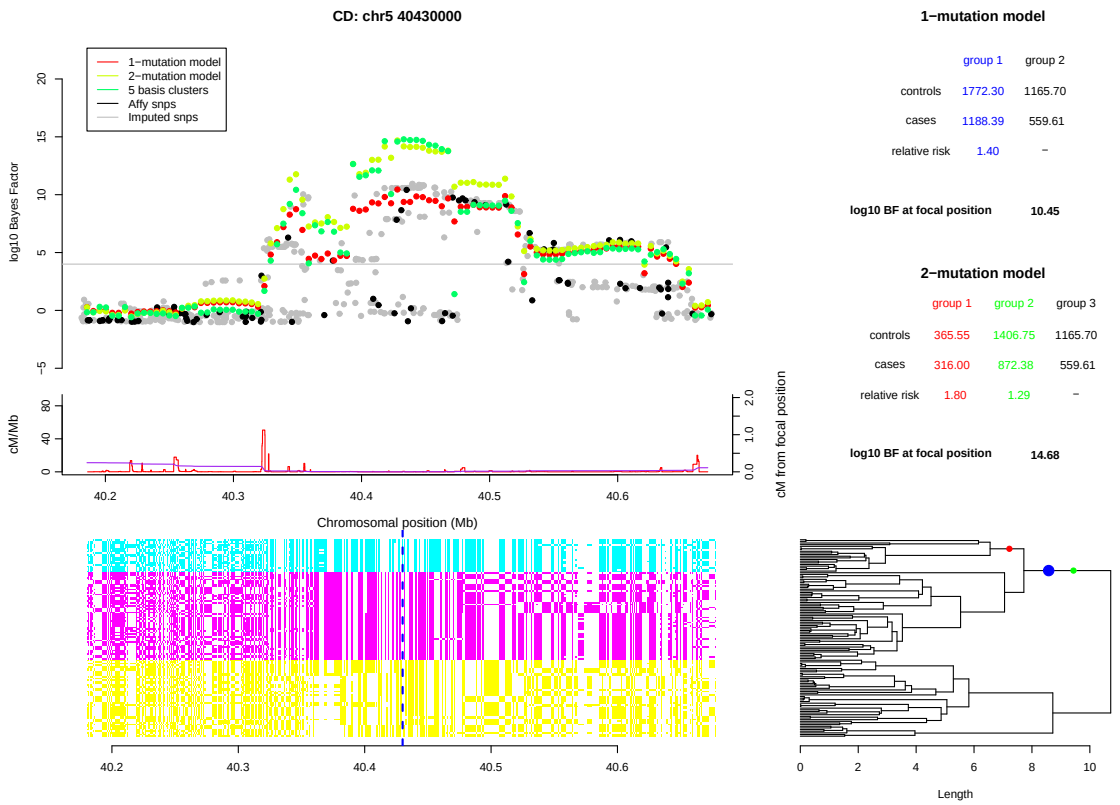


Figure 6.17: GENECLUSTER analysis of CD on chromosome 5 using TREESIM trees. The top left panel of the figure shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within a gene desert on chromosome 5 of the CD analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 groups of haplotypes that are identified by the best fitting 2 mutations, which match exactly to the groups identified in the corresponding analysis with a LSTREE tree in (Figure 6.18). The dashed vertical blue line indicates the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position). The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

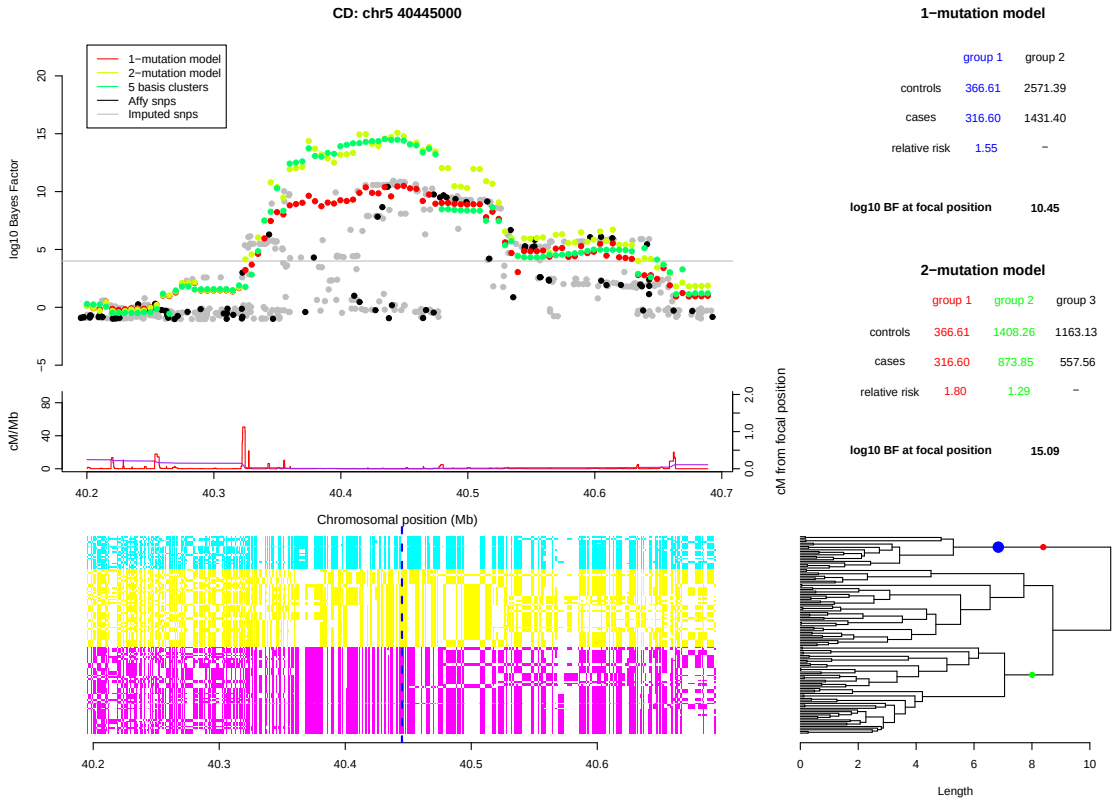


Figure 6.18: GENECLUSTER analysis of CD on chromosome 5 using LSTREE trees. The top left panel of the figure shows the $\log_{10}$ Bayes Factors for the 1-mutation model (red), 2-mutation model (light green), 5 basis clusters model (dark green) within a gene desert on chromosome 5 of the CD analysis. The single SNP Bayes Factors are displayed in black for Affymetrix SNPs and grey for imputed SNPs. The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 groups of haplotypes that are identified by the best fitting 2 mutations, which match exactly to the groups identified in the corresponding analysis with a TREESIM tree in (Figure 6.17). The dashed vertical blue line indicates the position of the largest $\log_{10}$ Bayes Factor for the 2-mutation model (the focal position). The bottom right panel shows the estimated LSTREE tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes Factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes Factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.

Chapter 7

Application to Fine-mapping

7.1 Introduction

Once detected and replicated the next stage in the study of a disease locus is to identify the specific allele(s) that confers an elevated risk of disease. An example of a fine-mapping study is the WTCCC+, which is a follow up from the WTCCC study. The WTCCC+ study focuses on regions of association found in the WTCCC study that have since been replicated. The study is done in 3 stages:

1 select 80 individuals from the case and control samples for resequencing to identify the segregating sites in each region of association;

2 impute or genotype all case and control individuals at segregating sites identified in stage 1;

3 conduct an association scan to identify the causal variant(s) using the data collected in stage 2.

In this chapter we will focus on the challenge of choosing the subsample to resequence in stage 1. The subsample must be chosen so that there is segregation at the disease SNP(s) otherwise there is no chance of identifying the disease allele(s) in the subsequent association study.

In the previous chapter we showed that GENECLUSTER has the ability to identify the haplotype backgrounds that harbour a causal variant. In this chapter we will use this feature to develop a strategy for choosing a subsample of case and control individuals that segregate at the disease SNPs.

We will attempt to find out, through simulation, the power of various selection strategies to identify a disease SNP as segregating when (i) there is a single disease SNP in the region of association, and (ii) when there are two disease SNPs in the region of association.

7.2 Sampling strategies

In this section we will list the possible strategies for selecting a subsample of N (where $N = 80$ for the WTCCC+ study) individuals. We seek the strategy that can maximise the chances of segregation at the disease SNP(s). We propose 3 strategies, labelled A, B and C:

A random selection,

B selection based on the results of an association scan using a single SNP association test, and

C selection based on the best fitting mutations identified by the Model A of

GENECLUSTER.

7.2.1 Strategy A: random sampling

The simplest strategy is to select a subsample of N individuals randomly from the case and control samples.

If the minor allele frequency at the disease SNP is f, then the power of this strategy, p, is given by $p = 1 - (1 - f)^{2N} - f^{2N}$. The power is therefore determined by the probability of sampling the minor disease allele since f^{2N} becomes insignificant for $f < 0.5$ and $N > 10$. Thus, this strategy is underpowered for rare alleles. For example, if $N = 80$ and $f = 0.05$ then $p = 1.0$, but the power drops to $p = 0.80$ when $f = 0.01$.

If there are two disease SNPs at a disease locus, in close proximity so no recombination events have occurred between them in the case-control sample, then we observe 3 disease haplotypes comprised of the two alleles at the disease SNPs. In order to observe segregation at both SNPs, all 3 haplotypes need to be sampled. If the frequency of the haplotypes are $\mathbf{f} = (f_1, f_2, f_3)$, then the power to capture both SNPs as segregating is $p = 1 - (f_1 + f_2)^{2N} - (f_2 + f_3)^{2N} - (f_1 + f_3)^{2N}$. Again the power of this strategy drops if one of the haplotypes is rare. For example, if $N = 80$ and $\mathbf{f} = (0.1, 0.1, 0.8)$ then $p = 1.0$, but when $\mathbf{f} = (0.02, 0.018, 0.8)$ the power falls to $p = 0.96$.

The random strategy represents a non-parametric method since it makes no assumptions about the underlying disease model.

7.2.2 Strategy B: sampling based on single SNP tests

A more intelligent strategy is to make use of the results that identified the region in the genome-wide scan. The power to detect a causal variant at a typed SNP increases with the correlation coefficient between it and the causal SNP (Pritchard, 2001). Thus, SNPs showing the greatest signal of association in a region are likely to be highly correlated with the true causal SNP, so sampling equal numbers of the two alleles at an associated SNP can improve the chances of capturing the minor allele at the disease SNP.

Suppose that at the SNP with greatest signal of association (the focal SNP) the alleles are a and A and that case individuals preferentially carry A. We investigate the following selection strategy to select a subsample of N individuals, defined by 3 steps:

1 As far as possible, balance the number of a and A alleles in the subsample to maximise the chances of segregation at the disease SNP. So, if there are N or more of both a and A alleles, then we can skip to step 2; however if, for example, there are less than N a alleles then we select all individuals carrying at least one a allele and make up the rest of the subsample using AA homozygotes according to step 2.

2 Conditional on the selections made in step 1, resolve selection choices by selecting as many homozygotes as possible because they are easier to sequence. For example, if after applying condition 1 we have a set of suitable individuals carrying the A allele then sample as many AA homozygotes as possible and make up the rest with Aa heterozygotes. Proceed to step 3.

3 Conditional on the selections made in 1 and 2, resolve any selection choices by preferentially selecting case AA homozygotes, which are assumed to be more likely to carry the true risk allele, control aa homozygotes, which are assumed to be more likely to carry the protective allele, and balance the number of case and control Aa heterozygotes. For example, if after steps 1 and 2, we have a set of AA homozygotes suitable for selection, sample as many AA case individuals as possible.

If after applying all 3 steps selection is still not resolved, then this strategy will select randomly from the set of suitable individuals. In the subsequent simulation studies we will use the log additive Trend Test to implement this strategy.

To illustrate the above strategy, consider the example in Table 7.1, which summarises the genotypes counts in a hypothetical case-control sample. For $N = 80$, no selection is resolved in step 1 since there are more than 80 A and a alleles in the case-control sample. Step 2 implies that all 34 AA homozygotes should be selected, which means 12 Aa heterozygotes and 34 aa homozygotes must be selected to balance the number of A and a alleles (as specified by step 1). Finally, step 3 implies that all 34 aa homozygotes should be randomly selected from the control sample and 6 Aa heterozygotes should be randomly selected from each of the case and control samples.

Phenotype	AA	Aa	aa
Control	4	30	100
Case	30	40	160

Table 7.1: An example contingency table for genotype data at a SNP.

The strategy described here makes the assumption that the focal SNP is correlated with the disease SNP and oversamples two subsets of individuals, $S_1 = \{AA$ cases$\}$

and $S_2 = \{aa \text{ controls}\}$. If the assumption is true, then the frequency of the deleterious allele is elevated in one of those sets, S_1 say, and oversampling S_1 leads to a greater probability of capturing that allele. Similarly, oversampling S_2 will increase the probability of sampling the protective allele. For example, if the focal SNP is perfectly correlated with the disease SNP then the frequency of the deleterious allele in S_1 is 1, and is 0 in the other, so sampling just one individual in each of S_1 and S_2 will capture both alleles at the disease SNP. Moreover, in a single SNP disease model this strategy can be no worse than the random strategy, since if the focal SNP is in linkage equilibrium with the disease SNP, then the frequencies of the protective and deleterious alleles in S_1 and S_2 are the same as their frequencies in whole case-control sample.

However, when there are two disease SNPs, which define 3 haplotypes, then all 3 haplotypes need to be sampled to observe segregation at both disease SNPs, which can cause this strategy to be underpowered. This is best illustrated by example: suppose that

- d_1, d_2 are the two disease SNPs, with alleles $\{A_{d_1}, a_{d_1}\}$ and $\{A_{d_2}, a_{d_2}\}$, respectively, and
- the haplotypes comprising of the alleles at the focal SNP and the two disease SNPs are $\{h_1 = (A, A_{d_1}, A_{d_2}), h_2 = (A, a_{d_1}, A_{d_2}), h_3 = (a, A_{d_1}, a_{d_2})\}$, with frequencies $f_1 = 0.88, f_2 = 0.02$ and $f_3 = 0.1$.

This represents a situation where the focal SNP is perfectly correlated with d_2 but weakly correlated with the rare SNP, d_1. Sampling 40 AA homozygotes and 40 aa homozygotes has power $p = 1 - (\frac{0.88}{0.88+0.02})^{80} - (\frac{0.02}{0.88+0.02})^{80} = 0.83$ to capture all 3 haplotype backgrounds and hence identify both d_1 and d_2 as segregating, but

random sampling has power $p = 0.96$ as shown earlier. In effect, over-sampling S_1 and S_2 to capture both alleles at d_2 resulted in less power to capture the alleles at d_1.

7.2.3 Strategy C: sampling based on GENECLUSTER

Consider the situation where we know the true genealogical tree for our case-control sample, T_x^G, at position x and the branch, b_G, where the causal mutation occurred. Given T_x^G and b_G we would know exactly the allele type of every individual at the disease locus, d. If we wanted to sample individuals segregating at the disease SNP, we would just sample a subset of the individuals that are under branch b_G and a subset of the individuals that are not.

In a realistic setting we will not know T_x^G and b_G. However, GENECLUSTER approximates T_x^G, for the purpose of association testing, with the estimated genealogical tree for a reference panel, T_x, and a function, $\mathbf{z}_x$, that maps the case and control haplotypes to its leaves. In Section 6.3.3 we analysed the locations of the set of best fitting mutations in T_x, which has the maximum posterior probability out of all other sets of mutations under our model of association. We found that the best fitting mutations appear to accurately predict the panel haplotypes that carry the causal mutation(s). Based on this observation, we propose a novel strategy, which assumes the best fitting single mutation is the real disease mutation. The sampling strategy conditional on T_x and the best fitting mutation on branch, b, is to sample equal numbers of individuals carrying both alleles at the disease SNP defined by b.

To describe our strategy more precisely, we need to refer back to the approach we proposed in Section 5.1 and below are the notations that we will use. Let

- x be the position where we performed inference using GENECLUSTER;
- T_x be the estimated genealogical tree at x for the panel haplotypes $\mathbf{H} = \{H_1, \ldots, H_P\}$;
- $\mathbf{G} = \{G_1, \ldots, G_M\}$ be the case-control sample that we subsample from;
- b be the branch of the best fitting single mutation, which we will assume is the disease causing mutation and defines the disease SNP d';
- $\mathbf{H}^{x,b}$ be the set of panel haplotypes augmented with the SNP d' so that $H^{x,b}_{i,d'} \in \{A_b, a_b\}$ is the allele carried by the haplotype H_i at d';
- $Z_{i,x} = (z^{(1)}_{i,x}, z^{(2)}_{i,x})$ be the copying states of the pair of haplotypes of individual i at position x;
- $\{A, a\}$ be the alleles at the *true* disease SNP, d;
- N be the size of our subsample that we can resequence, i.e. $N = 80$ in the case of WTCCC+ study.

Under the 1-mutation model stated in Section 5.1.4, the expected number of A_b and a_b alleles carried by individual i is defined by

$$n_{i,A_b} = \sum_{k_1=1}^{P}\sum_{k_2=1}^{P}(\Pr(Z_{i,x} = (k_1, k_2))(\mathrm{I}(H^{x,b}_{k_1,d'} = A_b) + \mathrm{I}(H^{x,b}_{k_2,d'} = A_b)) \quad (7.1)$$

$$n_{i,a_b} = \sum_{k_1=1}^{P}\sum_{k_2=1}^{P}(\Pr(Z_{i,x} = (k_1, k_2))(\mathrm{I}(H^{x,b}_{k_1,d'} = a_b) + \mathrm{I}(H^{x,b}_{k_2,d'} = a_b)) \quad (7.2)$$

Our strategy to capture the alleles at a single disease SNP is to select a sample of $N/2$ individuals, S_1, that carry the greatest expected number of A_b alleles (i.e. select individuals i with the largest n_{i,A_b}) and similarly select a set, S_2, of $N/2$ individuals

that carry the greatest expected number of a_b alleles. If b approximates the branch of the true disease mutation well, then individuals in S_1 will correspond to the set of individuals carrying one of the disease alleles, A say, at d. Similarly, individuals in S_2 will correspond to the set of individuals carrying the other disease allele, a.

Our strategy to capture the alleles at k SNPs is to use the k-mutation model. We assume the disease SNPs are in close proximity so that no recombination events have occurred between any of them in our case-control sample. This means that the disease SNPs define $k+1$ disease haplotypes and our strategy is to try to select $k+1$ subsamples of individuals, with each subsample comprising of the individuals carrying the greatest expect number of one particular disease haplotype. Let

- $\mathbf{b}$ be the branches of the best fitting k mutations;
- $\mathbf{d} = (d_1, \ldots, d_k)$ be the k disease SNPs defined by mutations on the branches $\mathbf{b}$;
- $\mathscr{A}^{\mathbf{b}}$ be the set of $k+1$ possible haplotypes comprising of the alleles at the disease SNPs, $\mathbf{d}$;
- $\mathbf{H}^{x,\mathbf{b}}$ be the set of panel haplotypes augmented with the disease SNPs, $\mathbf{d}$, so that $H^{x,\mathbf{b}}_{i,\mathbf{d}} \in \mathscr{A}^{\mathbf{b}}$ is the disease haplotype carried by H_i at $\mathbf{d}$.

The expected number of disease haplotypes $\mathbf{a_b} \in \mathscr{A}^{\mathbf{b}}$ carried by individual i at the disease SNPs $\mathbf{d}$ is given by

$$n_{i,\mathbf{a_b}} = \sum_{k_1=1}^{P}\sum_{k_2=1}^{P}(\Pr(Z_{i,x} = (k_1,k_2))(\mathrm{I}(H^{x,\mathbf{b}}_{k_1,\mathbf{d}} = \mathbf{a_b}) + \mathrm{I}(H^{x,\mathbf{b}}_{k_2,\mathbf{d}} = \mathbf{a_b})) \quad (7.3)$$

(7.4)

Our strategy using the k-mutation model is to sample the $N/(k+1)$ individuals carrying the greatest expected number of $\mathbf{a_b}$ alleles for each $\mathbf{a_b} \in \mathscr{A}^{\mathbf{b}}$. If $\mathbf{b}$ is a good approximation to the set of branches where the true disease mutations lie, then our strategy samples the maximum expected number of each disease haplotype at the true disease SNPs. For precisely the same reason as described for Strategy B, however, if $\mathbf{b}$ is not a good approximation for the set of branches with the true disease mutation, or if there are more than k disease SNPs, then this strategy can be underpowered compared to the random strategy.

7.3 Simulation study

We conduct our simulation study in two parts. The first part was conducted for the design stage of the WTCCC+ study. The aim was to assess how well segregation at disease SNPs can be captured under a single disease model. We assessed strategies A and B. In the second study, we simulated under a two disease SNP model to assess how the power of those two strategies are affected by allelic heterogeneity. In addition, we evaluated strategy C to see if using GENECLUSTER can provide a boost in power.

7.3.1 Calculating power

We define the power of each strategy as the frequency of selecting a subsample of individuals that segregates at the true disease SNP. Segregation at the disease SNP might not be observed if the minor allele at the disease SNP only appears once in the subsample, due to genotyping error. To control for this, it is preferable to capture

two instances of the minor allele. Therefore we will measure the power of capturing both alleles, at the disease SNP, at least once and at least twice in our results.

7.3.2 Sampling in a single mutation disease model

We generated our data sets using the same approach as in Section 5.2.3 except that we increased our sample size to 3000 control and 2000 case individuals, to match the WTCCC+ study. Briefly, we simulated the data sets using HAPGEN conditional on 5 ENCODE regions. For each SNP in an ENCODE region we generated a data set with the minor allele at that SNP as the disease allele. We simulated 4380 data sets with a single causal variant of relative risk 1.3, and again with relative risk 2.5. We presented to each strategy only data at SNPs on the the Affymetrix 500K chip.

We performed the single SNP log additive Trend Test using the software SNPTEST, which was used in the WTCCC study, at all Affymetrix SNPs. Since only regions that are detected and replicated are put forward to the WTCCC+ study, we discarded all data sets that did not have a P value below 5×10^{-7}, which is the significance threshold used by the WTCCC.

We tested the following selection strategies.

- **Random Control** Randomly sample 80 control individuals,
- **Random Case** Randomly sample 80 case individuals,
- **Random Mixed** Randomly sample 42 case and 38 control individuals,
- **SNPTEST Pure** Based on the Strategy B outlined in section 7.2.2, using the Trend Test P values, to select all 80 individuals.

All 4 selection strategies above have a stochastic element in their implementations, so to reduce the variance in our results we repeated each strategy 50 times on each data set.

As before, we stratify our results according to whether the disease SNP is tagged, with sample r^2 greater than 0.8, to an Affymetrix SNP or a MMP based on a pseudo HapMap panel (please refer back to Section 5.2.3 for more details), and whether the disease allele is common, with MAF greater than 5%.

For common variants all strategies have a power of 100% to capture each allele at least twice, both at a relative risk of 1.3 and 2.5. However, we do observe differences when the disease allele is rare.

At a relative risk of 1.3, there is very little power to detect a rare variant and only the following numbers of data sets exceed the P value threshold:

- 0 out of 749 data sets with a rare and untagged causal variant and
- 2 out of 426 data sets with a rare and tagged causal variant.

At a relative risk of 2.5, there is greater power to detect a rare variant and the numbers of data sets exceeding the P value threshold are:

- 174 out of 749 data sets with a rare and untagged causal variant and
- 378 out of 426 data sets with a rare and tagged causal variant.

Table 7.2 displays the power (in %) to capture both alleles at the disease SNP at least once in a subsample selected using each strategy; the power to capture both alleles at least twice are given in brackets.

In our simulation study the disease allele is set as the minor allele at the disease SNP and so the minor allele frequency is higher in the case sample. Therefore, Random Case is the most powerful random strategy since it samples only from the case sample, which maximises the chances of capturing the minor allele. For the same reason, Random Control has the least power and Random Mixed, which represents a compromise between Random Control and Random Case, has intermediate power.

The intelligent strategy, SNPTEST Pure, obtained a power of 100% to capture a minor allele at least twice. As explained before, the focal SNP (with maximum signal of association) is likely to be correlated with the disease SNP and explain the boost in power over the the random strategies.

Strategy	Rare and untagged (174/749)	Rare and tagged (378/426)
Random Control	96.9 (89.1)	93.6 (78.6)
Random Case	99.8 (99.2)	99.6 (97.9)
Random Mixed	99.7 (98.2)	98.9 (95.3)
SNPTEST Pure	100 (100)	100 (100)

Table 7.2: Power (%) of each strategy to capture both alleles at the disease SNP at least once (twice) when the true risk allele is rare with relative risk 2.5. We assessed only those data sets, which show a P value below 5×10^{-7}. The number of data sets satisfying this condition and the total number of data sets simulated are given for each disease SNP type. For example, a total of 749 data sets were generated with a rare and untagged disease allele, and 174 of those 749 data sets has a maximum P value below 5×10^{-7}.

7.3.3 Sampling in a two mutation disease model

To assess how the Random and SNPTEST strategies are affected by allelic heterogeneity, we repeated the simulation approach used in Section 5.2.4 but we increased the sample size to 3000 control and 2000 case individuals, to match the WTCCC+ study. Briefly, we simulated using HAPGEN under the two disease SNP model con-

ditional on 5 ENCODE regions. Another difference in our procedure from before is that, in order to increase the number of available data sets, we generated data sets for all 5 ENCODE regions even those with less than 50 suitable disease SNP pairs. We then pooled the data sets from all 5 regions for our power calculations. As before, the data sets were thinned to the Affymetrix SNPs before they were presented to each strategy.

We have chosen to simulate disease model 1 from Section 5.2.4, where each data set is assigned a rare disease allele (allele frequency less that 2%) and a common disease allele (allele frequency between 5% and 20%). We set the common disease allele to have a relative risk of 1.3 and varied the relative risk of the rare disease allele. We chose disease model 1 because it represents a model of allelic heterogeneity that we have observed before, at the *NOD2* locus for CD. Further, this disease model represents an example where the identification of both disease SNPs, in particular the rare SNP, might be challenging.

We applied GENECLUSTER with the 1-mutation model and 2-mutation model at SNPs every 5kb apart and the log additive Trend Test at all Affymetrix SNPs. We used a set of trees estimated using TREESIM and a penetrance prior of $\beta(20.0, 30.0)$, just as we did in Chapter 6.

We want to focus on regions that show a signal of association and allelic heterogeneity. Therefore, we only consider data sets with a GENECLUSTER Bayes Factor greater than 10^4 and the maximum 2-mutation Bayes Factor is at least 10 times greater than the 1-mutation model Bayes Factors, i.e. a posterior probability of at least 0.91 for the 2-mutation model assuming prior odds of 1:1.

As before, we stratify our analyses according to whether each allele is tagged. For

brevity, we abbreviate each type of disease allele pair as follows.

- **T1** rare untagged and common untagged
- **T2** rare tagged and common untagged
- **T3** rare untagged and common tagged
- **T4** rare tagged and common tagged

In addition to the Random and SNPTEST strategies, outlined previously, we assess the following additional strategies.

- **GENECLUSTER Pure**. Take equal numbers of individuals carrying the highest expected count of each disease haplotype defined by the best fitting 2 mutations identified by GENECLUSTER at the focal SNP. (Note that it is impossible to distribute the 80 individuals equally between the 3 haplotypes identified by the best fitting 2 mutations, so we distribute them as 27, 27 and 26).
- **GENECLUSTER Mixed**. Select only 40 individuals according to the GENECLUSTER Pure strategy (we distribute 13, 13 and 14 between the 3 distinct haplotypes identified by the best fitting 2 mutations) and select 20 control and 20 case individuals randomly.

To reduce the variance in power calculated for the stochastic strategies (Random, SNPTEST Pure and GENECLUSTER Mixed), we repeated each strategy 50 times on each data set and averaged the power.

We find that there is an insufficient number of suitable data sets to calculate power when the rare variant has a relative risk below 2.5. When the rare variant has a relative risk of 2.5 and 3.0, the power of each strategy to capture the alleles at both disease SNPs at least once are shown in Tables 7.3 and 7.4 respectively; the power to capture the alleles at both SNPs at least twice are shown in brackets. In those tables, we have also indicated how many data sets, out of the total number generated, satisfies the condition of having a 2-mutation model $\log_{10}$ Bayes Factor greater than 4 and also 1 greater than the 1-mutation model $\log_{10}$ Bayes Factors.

Strategy	T1 (54/429)	T2 (105/357)	T3 (70/291)	T4 (123/238)
Random Control	82.6 (54.7)	90.6 (70.6)	80.9 (51.9)	88.3 (67.1)
Random Case	98.3 (91.8)	99.5 (97.7)	97.9 (90.8)	99.3 (96.6)
Random Mixed	95.0 (81.8)	98.2 (92.0)	94.3 (80.7)	97.6 (89.7)
SNPTEST Pure	95.0 (87.3)	99.8 (99.4)	84.7 (71.4)	98.1 (95.8)
GENECLUSTER Pure	100 (100)	100 (100)	98.6 (98.6)	100 (100)
GENECLUSTER Mixed	100 (99.0)	100 (100)	99.4 (98.5)	99.8 (99.3)

Table 7.3: The power (%) of selection strategies when there are two disease SNPs. The first disease allele is rare (allele frequency less than 2%) with a relative risk of 2.5 and the second is common (allele frequency between 5% and 20%) with a relative risk of 1.3. The figures in the table give the frequency of capturing both alleles at the disease SNPs at least once, and at least twice in brackets. At the top, we have indicated how many data sets, out of the total number generated, satisfies the condition of having a 2-mutation model $\log_{10}$ Bayes Factor greater than 4 and also 1 greater than the 1-mutation model $\log_{10}$ Bayes Factors. For example, a total of 429 data sets were generated with a pair of untagged rare and untagged common disease alleles (type T1) and 54 out of those 429 data sets has a 2-mutation model $\log_{10}$ Bayes Factor greater than 4, which is also 1 greater than the 1-mutation model $\log_{10}$ Bayes Factor.

As before, sampling only the case sample (Random Case) is the most powerful random strategy since the minor alleles at the disease SNPs are the disease alleles and the minor allele frequencies are higher in the case samples.

SNPTEST Pure is not as powerful as in the single mutation disease model, especially if the rare variant is not tagged. The explanation for this is that the SNP with the

Strategy	T1 (99/429)	T2 (134/357)	T3 (133/291)	T4 (124/238)
Random Control	82.3 (53.6)	88.1 (65.5)	81.0 (52.4)	87.6 (65.2)
Random Case	98.9 (94.4)	99.4 (97.7)	99.1 (95.1)	99.5 (97.7)
Random Mixed	95.7 (84.9)	98.1 (92.2)	96.4 (85.3)	98.1 (91.5)
SNPTEST	94.2 (88.2)	99.6 (98.6)	88.3 (76.1)	98.5 (96.9)
GENECLUSTER	100 (99.0)	100 (100)	99.3 (99.3)	99.2 (98.4)
GENECLUSTER Mixed	100 (100.0)	100 (100)	99.9 (99.4)	100 (100)

Table 7.4: The power (%) of selection strategies when there are two disease SNPs. The first disease allele is rare (allele frequency less than 2%) with a relative risk of 3.0 and the second is common (allele frequency between 5% and 20%) with a relative risk of 1.3. The figures in the table give the frequency of capturing both alleles at the disease SNPs at least once, and at least twice in brackets. At the top, we have indicated how many data sets, out of the total number generated, satisfies the condition of having a 2-mutation model $\log_{10}$ Bayes Factor greater than 4 and also 1 greater than the 1-mutation model $\log_{10}$ Bayes Factors. For example, a total of 429 data sets were generated with a pair of untagged rare and untagged common disease alleles (type T1) and 99 out of those 429 data sets has a 2-mutation model $\log_{10}$ Bayes Factor greater than 4, which is also 1 greater than the 1-mutation model $\log_{10}$ Bayes Factor.

lowest P value is likely to be correlated with the common variant and efforts are concentrated on capturing that single variant rather than both variants.

GENECLUSTER Pure and GENECLUSTER Mixed are the most powerful strategies. This is evidence of the best fitting mutations on the genealogical tree, constructed by GENECLUSTER, providing a good approximation to the true causal mutations on the genealogical tree of a case-control sample. The mixed strategy is more powerful than the pure strategy, which suggests that a subsample of 40 individuals is sufficient to capture the signal identified by GENECLUSTER and extra power can be gained by using a different strategy (such as a random selection strategy) to capture any variants that are missed.

7.4 Discussion

In this chapter we have assessed the power to identify a disease SNP as segregating in a fine-mapping study. We have evaluated several selection strategies under simulation.

In the first instance we focused on the scenario where a region contains a single disease mutation that has been detected by the WTCCC study with a P value threshold of 5×10^{-7}. Under this scenario we find that the power to identify the causal SNP as a segregating site is high. The most powerful strategy is to select equal numbers of individuals carrying each allele at the focal SNP based on the single SNP association test. This strategy (SNPTEST Pure) has a power of 100% under our simulations and is so powerful because the focal SNP is likely to be correlated with the actual disease SNP.

We then moved on to a two disease SNP model. We considered regions where a signal of both association and allelic heterogeneity has been detected by GENECLUSTER. We find that in this scenario, SNPTEST Pure is underpowered. As we illustrated earlier with an example in Section 7.2.2, oversampling individuals that segregate at one of the disease SNPs might reduce the power to capture the alleles at the second disease SNP. However, our strategy based on the 2-mutation model of GENECLUSTER has a power of almost 100%. This strategy relies on using the best fitting mutations on our constructed genealogical tree for the case-control sample to identify the individuals that are likely to carry the real causal variants. The high power is further evidence that constructing the genealogy in the way that we described in Section 5.1.3 is a good approach for detecting association and allelic heterogeneity.

Based on these results, there appear to be value in applying GENECLUSTER to fine-mapping. We recommend the investigator to implement the 1-mutation and 2-mutation models, and if there is a signal of association and allelic heterogeneity (indicated by a large increase in the 2-mutation model Bayes Factor compared to the 1-mutation model Bayes Factor) then a strategy along the lines of GENECLUSTER Pure should be used. However, our strategy shares the same problems as SNPTEST Pure, which is that when the wrong sets of individuals are identified as carrying the disease allele, the power to capture the alleles at the real disease SNPs can be reduced. Therefore a mixed strategy similar to GENECLUSTER Mixed might be preferable, which can boost power when the GENECLUSTER Pure strategy fails. Although our results suggest that this strategy works well, it is only applicable to regions where a signal of association and allelic heterogeneity is detected. Tables 7.3 and 7.4 show that the power to detect association and allelic heterogeneity is at best 50% and can be as low as 13%, so our strategy looks to have limited use under our simulation parameters. On a related note, one might be interested to know how applicable our strategy is to the WTCCC+ study, which focuses on regions detected by a single SNP association test. Table 7.5 displays the total number of data sets we simulated for each type of disease SNPs, n; the total number of data sets, n_1, out of the n data sets, which has a P value below 5×10^{-7}; and the total number of data sets, n_2, out of the n_1 data sets, which has a maximum $\log_{10}$ Bayes Factor for the 2-mutation model greater than 4 and at least 1 greater than the $\log_{10}$ Bayes Factors for the 1-mutation model. We again see a limited applicability for our strategy – at most we can hope to apply our strategy 50% of the time.

In Section 5.3 we have highlighted some of the weaknesses in our simulation design. The main problem is that we do not know the models of allelic heterogeneity that

	T1	T2	T3	T4
n	429	357	291	238
n_1	76 (122)	120 (141)	236 (275)	201 (221)
n_2	26 (51)	40 (73)	82 (115)	98 (109)

Table 7.5: The table show total number of data sets we simulated for each type of disease SNPs under a two disease SNP model, n; the total number of data sets, n_1, out of the n data sets, which has a P value below 5×10^{-7}; and the total number of data sets, n_2, out of the n_1 data sets, which has a maximum $\log_{10}$ Bayes Factor for the 2-mutation model greater than 4 and at least 1 greater than the $\log_{10}$ Bayes Factor for the 1-mutation model. The numbers outside of the brackets refer to the model where the rare disease allele has a relative risk of 2.5, and the numbers in the brackets refer to the model where it has a relative risk of 3.0. The common disease allele has a relative risk of 1.3 in both models.

underlie common complex diseases. So even though the two disease SNP model we have simulated appear to be challenging for fine-mapping, we can not know how relevant this result is without knowing the prevalence of the disease model in the genome. On the other hand, we have only considered a two disease SNP model and if there are 3 (or more) disease SNPs then the power for all strategies will drop further. When there are 3 or more disease SNPs, using the 2-mutation model would suffer a loss of power for the same reasons as when SNPTEST Pure is applied to the two disease SNP model. A better implementation would be to use a strategy based on the k-mutation model (for some $k > 2$), when there are k disease SNPs. However, the application of this strategy would require us to detect k mutations but we do not know our power for that, and we are unlikely to be able to find out for large k due to the computation cost.

The WTCCC+ have chosen a mixed strategy comprising of SNPTEST Pure and a second strategy to sample haplotype diversity in an associated region. A tree-like representation of the genotype diversity in the region is created using a suitable clustering algorithm and after observing where the genotypes have been sampled by SNPTEST Pure, the remaining individuals are selected from the unsampled

groups of leaves under the tree. This strategy is also likely to be effective under scenarios of allelic heterogeneity and moreover makes weaker assumptions about where the disease mutations reside on the tree, which means that it can retain power over a larger set of disease models, compared to our strategy. However, if GENECLUSTER detects a signal of association and allelic heterogeneity in the associated region, GENECLUSTER Pure might outperform the WTCCC+ strategy since it oversamples the two specific groups of individuals under tree that are likely to the carry the protective and deleterious disease alleles respectively. Therefore, a mixed strategy incorporating GENECLUSTER Pure might be a more powerful strategy for the WTCCC+ study.

One final remark on fine-mapping with GENECLUSTER is that in Section 6.3.3 we showed that the best fitting mutations identify the panel haplotypes that carry the causal variants. So resequencing and analysing each haplotype group, in the panel, identified by the best fitting mutations might be an alternative and more efficient procedure to identify the causal variants.

Chapter 8

Summary and Future Work

8.1 Summary

The motivation for the project behind this thesis is to develop methods to detect association in genome-wide association studies and provide a boost in power over the SNP based approaches that are currently used. When we first set out with this goal, large scale population data were scarce and we therefore anticipated a need for simulation to evaluate the methods we wanted to develop. The simulation methods available at the time generally did not have the facility to incorporate the information that are available from population data. The approach in Chapter 3 offers us a fast and realistic method of simulation conditional on observed population data, which has facilitated the systematic evaluation of our methods in Chapters 4, 5 and 7.

After Chapter 3 we returned to our main focus of detecting association when SNP based analyses are underpowered. Imputation has greatly improved the power of

SNP based approaches to uncover causal variants that are not in strong LD with any single or a combination of typed SNPs. However, even if every SNP can be genotyped or correctly imputed, testing for association at each single SNP cannot guarantee to detect all causal variants, for example when the causal variant is a haplotype or when there are multiple disease SNPs at a locus. Chapter 4 represents our early attempts, where we looked at ways to use the model from Li and Stephens (2003) to detect association in haplotype data. One method in particular uses the idea of clustering haplotypes into multiple groups based on a tree, similar to Durrant et al. (2004), which inspired the GENECLUSTER method in Chapter 5. For GENECLUSTER we also have combined a number of ideas established in the field of association testing.

Our method is based on a very similar Hidden Markov Model to IMPUTE (Marchini et al., 2007), which allows us to analyse genetic data at all SNPs simultaneously, modulated by the local recombination rates. We therefore inherit the benefits of detecting rare and poorly tagged variants, which in general can not be detected by only a few surrogate SNPs.

In order to retain power under multiple mutations, we utilised the ideas developed by Zollner and Pritchard (2005) and Minichiello and Durbin (2006), who model association based on a genealogical tree for the case-control sample. Placing mutations on a tree provides a natural interpretation of the way a disease has propagated through a population and Minichiello and Durbin (2006) showed that this can yield additional information about the underlying disease model.

A novel feature of our approach is the way we have combined the ideas from IMPUTE, to model genetic variation, and the ideas of Zollner and Pritchard (2005) and

Minichiello and Durbin (2006), to model disease. We achieved this by constructing the genealogy for the case-control sample, which would be computationally prohibitive to estimate under the coalescent model, using the genealogy estimated for a smaller panel of reference haplotypes, which can be estimated with relative ease. We have related our approach to the coalescent model, however a discussion on how close this relationship is and how it can be improved is beyond the scope of this thesis. Our goal is to provide a powerful method for detecting association in genome-wide studies and the availability of the WTCCC data sets provided us a means to assess our approach on real population data in Chapter 6.

The results presented in Chapter 6 are encouraging. Comparisons at known disease loci show a clear correlation between the signals from SNP based approaches and our method. Moreover, we observe elevated signals in the presence of allelic heterogeneity and haplotypic association at confirmed disease loci, which provide compelling evidence that our method should be used for genome-wide association scans to boost the power of SNP based approaches. Further, we found that our analyses lead to accurate predictions about the underlying disease model, in particular, we were able to identify the haplotype backgrounds, which harbour a causal allele. There is a clear application for this method in fine-mapping, which we assessed in Chapter 7. There we found, through simulations under a specific model of allelic heterogeneity, that GENECLUSTER can provide an improvement for detecting the segregation of disease SNPs. The applicability of our strategy to practical settings is unclear, however those results provide further validation of our method as a way to detect association.

In this thesis we have taken a Bayesian approach because it offers us a coherent probabilistic framework to work from. The Bayes Factor offers a natural summary

for amount of evidence under a model of association compared to the null model of no association. There is growing support (Balding, 2006b; Servin and Stephens, 2007; Marchini et al., 2007) that this is a more preferable approach to perform inference. However, a discussion between the Bayesian and Frequentist paradigm should take place elsewhere. In any ways, our method can be easily adapted to perform Frequentist inferences, for example in the way described by Minichiello and Durbin (2006), should the reader or investigator be more comfortable with P values.

8.2 Future Work

We have now come to the end of this thesis. Due to constraints on time and resources, various approaches and analyses could not be implemented, and we list them here in the hope that our research will continue into the future.

In Chapter 3 we introduced HAPGEN. In order to evaluate how well HAPGEN does at creating appropriate levels of LD, we compared it to a coalescent simulator and found that it created slightly less variation. It would be important to know if the difference has any significant impact on simulation studies. One way to assess this is to simulate disease data sets using both methods, with HAPGEN conditioned on a subset of MS simulated haplotypes, and compare the performances of single SNP and haplotype association tests to see if there is a difference in power between the two types of data sets. Further, it would interesting to know which component of the HAPGEN model, recombination or mutation, is responsible for the reduced variation. Preliminary analyses showed that the fractions of tagged common variation for HAPGEN and MS match when HAPGEN's mutation rate is set to approximately twice the value given in Equation (3.11). We should therefore investigate the impact

of increasing the mutation rate on the simulated LD structure and more importantly whether it can lead to more realistic simulations.

On simulating allelic heterogeneity we have simulated a priori the recombination events between two disease SNPs. Conditioning on the observed data in the panel will result in a more realistic model. In addition, the current implementation does not allow an easy extension to a general n disease SNP model and it would be preferable if HAPGEN has this facility to aid further analyses on allelic heterogeneity.

HAPGEN has allowed us to perform extensive simulations in this thesis. A discussion on the weaknesses of our simulation design takes place in Section 5.3. Based on that, a more thorough assessment of our methods can be achieved by repeating simulations using an alternative simulator, such as MS, that is not based on the LS model and in doing so we remove any possible biases in the simulation that can favour our methods. Our understanding of disease models involving allelic heterogeneity is limited but this should change in the future as our field expands, hopefully aided by the use of GENECLUSTER. This will allow us to conduct more realistic simulation studies involving allelic heterogeneity. One final point on simulation is that we have, so far, exclusively simulated under the log additive disease model for genotype relative risk. The weak signal strength found at the BD locus in Chapter 6, which was identified by the WTCCC under a genotype model, suggests that we might be underpowered under alternative disease models. It is therefore important to assess the performance of our methods under a wider range of disease models.

The last point on disease models leads us to a weakness with GENECLUSTER. We have modelled association conditioned on the expected number of case and control haplotypes mapped to each leaf of the estimated tree for the panel. By summarising

the data in this way we can not model disease on a genotypic level, since we no longer have information about which leaves the pairs of haplotypes of each individual are mapped to. A simple solution would be to store at each leaf, corresponding to haplotype H_i in our panel, the expected number of control (and case) individuals that have one haplotype mapped to that leaf and the other haplotype to each of the other leaves. More precisely, we can store $m_{(i,j),x}$ and $n_{(i,j),x}$, which are the expected number of control and case individuals with first haplotype mapped to $H_i \in \mathbf{H}$ and the second to $H_j \in \mathbf{H}$:

$$m_{(i,j),x} = \sum_{k:\Phi_k=0} \Pr(Z_{k,x} = (i,j)), \tag{8.1}$$

$$n_{(i,j),x} = \sum_{k:\Phi_k=1} \Pr(Z_{k,x} = (i,j)). \tag{8.2}$$

Based on this summary statistic, we have the information to model a general genotype disease model. For example, the number of homozygote controls and cases carrying the putative disease allele A_b, created by a mutation on branch b of our constructed genealogical tree, would then be

$$m_{(A_b,A_b)} = \sum_{i=1}^{P}\sum_{j=1}^{P} m_{(i,j),x} \mathrm{I}(H_{i,d}^{x,b} = A_b, H_{j,d}^{x,b} = A_b), \tag{8.3}$$

$$n_{(A_b,A_b)} = \sum_{i=1}^{P}\sum_{j=1}^{P} n_{(i,j),x} \mathrm{I}(H_{i,d}^{x,b} = A_b, H_{j,d}^{x,b} = A_b). \tag{8.4}$$

This implementation would require an increase in storage space from $O(P)$ to $O(P^2)$. In taking the expected number of homozygote cases and controls we have ignored the uncertainty in our estimation of the copying states (as explained in Section 5.1.4). We should therefore perform inferences based on a set of copying states drawn from its posterior distribution and average over the resulting set of Bayes Factors. This

will improve accuracy but increases the computation cost.

For inference we also rely on the estimated tree for the panel haplotypes. There are two elements of a tree, the topology and the branch lengths, and we have chosen to estimate them separately. With regards to the topology, we currently only perform inference based on a single estimate at each position of interest. This underestimates the uncertainty in our estimation process. Our analysis from the WTCCC data in Chapter 6, at the *IL23R* and *NOD2* loci for CD, reveals that possible imperfection in the estimated topology can prevent the precise identification of the different haplotypes defined by the causal variants, which otherwise would lead to a likely increase in signal. Although we can not be sure that the true genealogical tree will partition the distinct risk haplotypes into separate groups, it does illustrate how power can potentially be lost. One way to minimise this effect, other than to find more accurate approaches to estimate marginal trees, is to perform analysis using a sample of trees and then average over the resulting set of Bayes Factors. The TREESIM method, which estimates genealogies based on the coalescent model with recombination, can output a set of trees at a given locus with its corresponding posterior probability conditioned on the panel. LSTREE can estimate a set of trees by using stochastic clustering, instead of the deterministic clustering approach that it currently employs.

With respect to the branch lengths, we have currently set the prior probability of a branch containing a causal mutation (under the mutation model, Model A) to be proportional to its expected length under the coalescent model. As with the topology, sampling tree lengths can improve power. However, in estimating the tree conditional on the panel, we can obtain a distribution on the number of mutations that each branch contains. Therefore conditioning on this information can improve

our estimations of each branch length, for example branches with a large number of expected mutations are likely to be longer (Cardin, N., private communications).

The improvements suggested for GENECLUSTER so far will vastly increase the computation cost and it is not clear whether they can be feasibly implemented. One way to reduce the computation cost of the mutation model is to employ a greedy algorithm, which places mutations successively on the tree, so that at each stage conditional on the previous mutations a new mutation is placed on the best fitting branch, i.e. one that contributes most according to Equation (5.31). This allows an approximate, but very fast, inference based on Model A with many mutations. This can have applications in fine-mapping. For example, it can facilitate estimations on the number of causal mutations present at a locus and identify the haplotype backgrounds in the panel that causal variants are likely to reside on.

A major problem for the basis cluster model (Model B) is that the number of basis clusters to use is a model choice that the investigator has to make. As yet we can not provide information on exactly how many basis clusters should be chosen. One way to improve this is to incorporate the number of basis clusters as a model parameter with a prior to reflect our belief on the number to use. What is an appropriate prior is very much an open question.

Data on numerous diseases, with large numbers of individuals, are becoming available at an ever increasing rate, so it seems that genome-wide association studies will play an important role in the future. We hope that the work presented in this thesis, combined with the refinements outlined here, will help progress the understanding genetic diseases.

Bibliography

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

D. J. Balding. A tutorial on statistical methods for population association studies (supplementary note). *Nature Reviews Genetics*, 165:781–791, 2006b.

Barrett et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, 2008. In press.

B. L. Browning and S. R. Browning. Efficient multilocus association testing for whole genome association studies using localised haplotype clustering. *Genetic Epidemiology*, 31:365–375, 2007.

N. Cardin. *Approximating the coalescent with recombination*. PhD in Statistics, Department of Statistics, 1 South Parks Road, Oxford, UK, 2007.

J. M. Chapman, J. D. Cooper, J. A. Todd, and Clayton D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human Heredity*, 56:18–31, 2003.

D. Clayton. Population association. In *Handbook of Statistical Genetics*, pages 519–540. 2001.

P. I. W. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nature Genetics*, 37:1217–1223, 2005.

de Bakker et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics*, 38:1166–1172, 2006.

D. G. T. Denison and C. C. Holmes. *Bayesian Methods for Nonlinear Classification and Regression*, chapter 2. John Wiley and Sons, Chichester, West Sussex, U.K., 2005.

Duerr et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, 314:1461–1463, 2006.

C. Durrant, K. T. Zondervan, L. R. Cardon, S. Hunt, P. Deloukas, and A. P. Morris. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *American Journal of Human Genetics*, 75:35–43, 2004.

Frazer et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851–861, 2007.

W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

R.C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3:479–502, 1996a.

R.C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. (Eds.) Tavar, editors, *IMA Volumes in Mathematics and its Applications*, volume 87, pages 257–270. Springer Verlag, Berlin, 1996b.

M. Guedj, G. Nuel, and B. Prum. A note on allelic tests in case-control association studies. *Annals of Human Genetics*, 72:407–409, 2008.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, New York, 2001.

J. N. Hirschhorn and J. M. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6:95–108, 2005.

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

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

R. R. Hudson. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44, 1990.

Hugot et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411:599–603, 2001.

B.S. Kerem, J. Rommens, J. M. Buchanan, and T. K. Markiewicz, Cox. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245:1073–1080, 1989.

J. F. C. Kingman. On the genealogy of large populations. *Appl Probability Trust.*, 82:27–43, 1982.

A. Konstas, L. M. Shearwin-Whyatt, A. B. Fotia, B. Degger, D. Riccardi, D. I. Cook, C. Korbmacher, and S. Kumar. Regulation of the epithelial sodium channel by N4WBP5A, a novel Nedd4/Nedd4-2-interacting protein. *Journal of Biological Chemistry*, 33:29406–29416, 2002.

T. Koski. *Hidden Markov Models for Bioinformatics*. Springer, New York, 2001.

J. C. Lam, K. Roeder, and B. Devlin. Haplotype Fine Mapping by Evolutionary Trees. *Am J. Hum. Genet.*, 66:659–673, 2000.

F. Larribe and S. Lessard. Gene mapping via the ancestral recombination graph. *Theoretical Population Biology*, 62:215–229, 2002.

N. Li and M. Stephens. Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics*, 165(4):2213–2233, 2003.

Y Li, J Ding, and G R Abecasis. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference [abstract program number 2290]. Presented at the annual meeting of The American Society of Human Genetics, 12 October 2006, New Orleans, Louisiana, October 2006. Available from `http://www.ashg.org/genetics/ashg06s/index.shtml`.

Libioulle et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLOS Genetics*, 3(4):538–543, 2007.

J. Marchini, B. Howie, G. Myers, S.and McVean, and P. Donnelly. Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Nature Genetics*, 39:906–913, 2007.

G. A. T. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304:581–584, 2004.

M. J. Minichiello and R. Durbin. Mapping trait loci using infrerred ancestral recombination graphs. *American Journal of Human Genetics*, 79:910–922, 2006.

M. Missbach, B Jagher, I. Sigg, S. Nayeri, C. Carlberg, and I. Wiesenberg. Thiazolidine Diones, specific ligands of the nuclear receptor Retinoid Z Receptor/Retinoid Acid Receptor-related orphan receptor with potent antiarthritic activity. *Journal of Biological Chemistry*, 271:13515–13522, 1996.

J. Molitor, P. Marjoram, and D. Thomas. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *American Journal of Human Genetics*, 73:1368–1384, 2003a.

J. Molitor, P. Marjoram, and D. Thomas. Application of bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. *Genetic Epidemiology*, 25:95–105, 2003b.

A. P. Morris. A flexible bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *American Journal of Human Genetics*, 79:679–694, 2006.

A. P. Morris, J. C. Whittaker, and D. J. Balding. Bayesian fine-scale mapping of disease loci, by hidden markov models. *American Journal of Human Genetics*, 67:155–169, 2000.

Ogura, Y. et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, 411:603–606, 2001.

R. L. Prentice and R. Pyke. Linkage disequilibrium in humans: models and data. *Biometrika*, 66:403–411, 1979.

J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases. *American Journal of Human Genetics*, 69:124–137, 2001.

J. K. Pritchard and N. J. Cox. Linkage disequilibrium in humans: models and data. *Human Molecular Genetics*, 11(20):2417–2423, 2002.

J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, 69:1–14, 2001.

N. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405: 847–856, 2000.

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

P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Genetic Epidemiology*, 18:143–156, 2006.

Scott, L. J. et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316:1341–1345, 2007.

B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLOS Genetics*, 3(7):1296–1308, 2007.

L. M. Shearwin-Whyatt, D. L. Brown, F. G. Wylie, J. L. Stow, and S. Kumar. N4WBP5A (Ndfip2), a Nedd4-interacting protein, localizes to multivesicular bodies and the Golgi, and has a potential role in protein trafficking . *Journal of Cell Science*, 117:3679–3689, 2004.

M. Stephens and P. Donnelly. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, 73:1162–1169, 2003.

M. Stephens, N.J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68: 978–989, 2001.

A. R. Templeton, E. Boerwinkle, and C. F. C. F. Sing. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. *Genetics*, 117:343–351, 1987.

A. R. Templeton, T. Maxwell, D. Posada, J. H. Stengaard, E. Boerwinkle, and C. F. Sing. Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics*, 169:441–453, 2005.

The Diabetes Genetics Initiative. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316:1331–1336, 2007.

The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, 306:636–639, 2004.

The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

The Wellcome Trust Case Control Consortium. Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447: 661–678, 2007.

D. C. Thomas. *Statistical methods in epidemiology*. Oxford University Press, New York, 2004.

J. Tzeng, B. Devlin, L. Wasserman, and K. Roeder. On the Identification of Disease Mutations by the Analysis of the Haplotype Similarity and Goodness of Fit. *American Journal of Human Genetics*, 72:891–902, 2003.

D. E. Undliem, B. A. Lie, and E. Thorsby. HLA complex genes in type 1 diabetes and other autoimmune diseases. Which genes are involved? *Trends in Genetics*, 17:93–100, 2001.

J.. Wakefield. Bayes factors for genome-wide association studies: comparison with P values. *Genetic Epidemiology*, 2008. In press.

W. Y. S. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6:109–118, 2005.

C. M. Weyand, T. G. McCarthy, and J. J. Goronzy. Correlation between disease phenotype and genetic heterogeneity in rheumatoid arthritis. *The Journal of Clinical Investigation*, 95:2120–2126, 1995.

Zeggini et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316:1336–1341, 2007.

Zeggini et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, 40:638–645, 2008.

Y. Zhu, S. McAvoy, R. Kuhn, and D. I. Smith. RORA, a large common fragile site gene, is involved in cellular stress response. *Oncogene*, 25:2901–2908, 2006.

S. Zollner and J. K. Pritchard. Coalescent-Based Association Mapping and Fine Mapping of Complex Trait Loci. *Genetics*, 169:1071–1092, 2005.

Appendix A

Supplementary Figures For Chapter 4

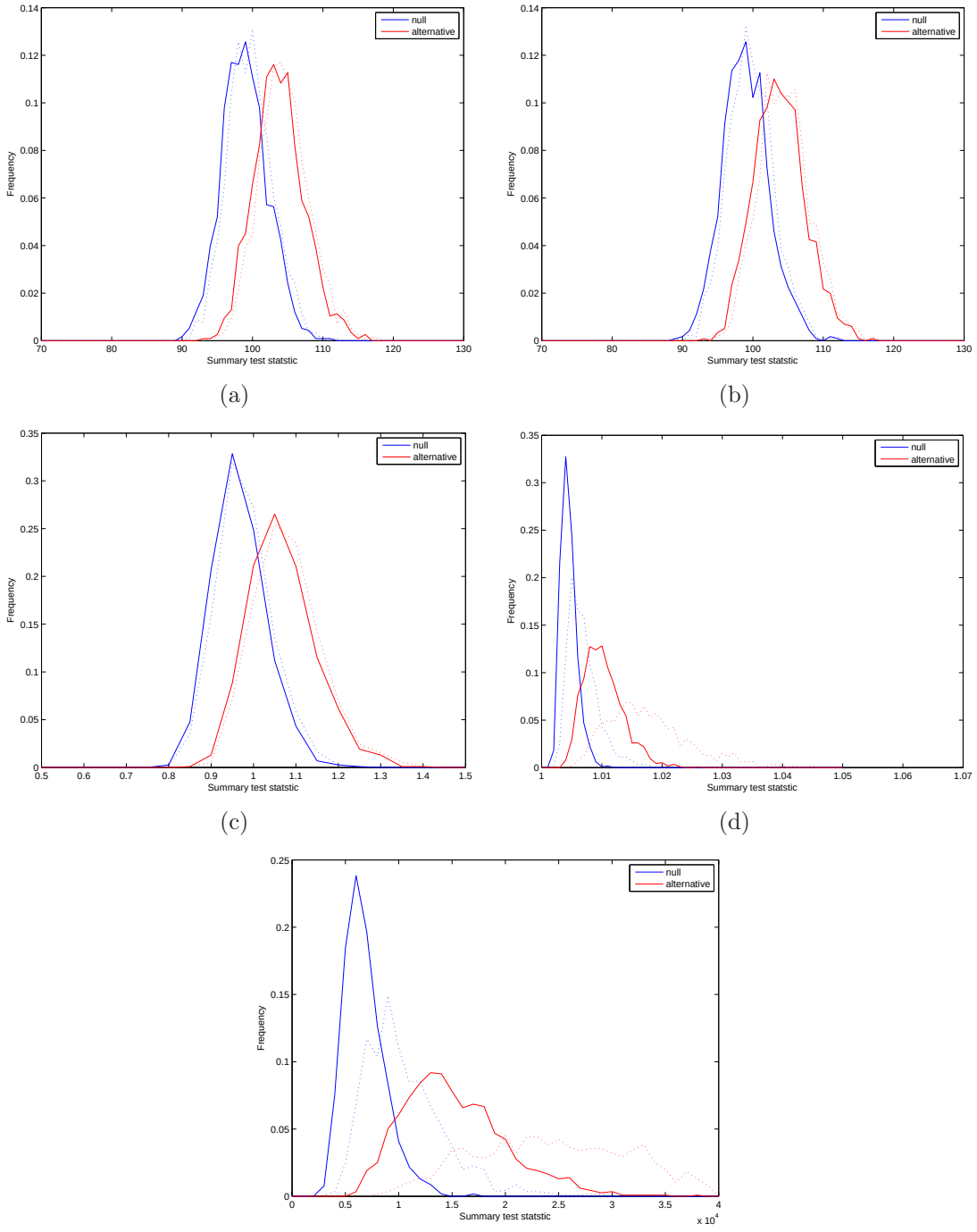


Figure A.1: Empirical null and alternative distributions for HAPSEARCH summary test statistics, **a** T_1, **b** T_2, **c** T_3, **d** T_4 and **e** T_5, from data sets simulated from ENr123. The solid lines indicate the mean test statistics and the dotted lines indicate the mode test statistics.

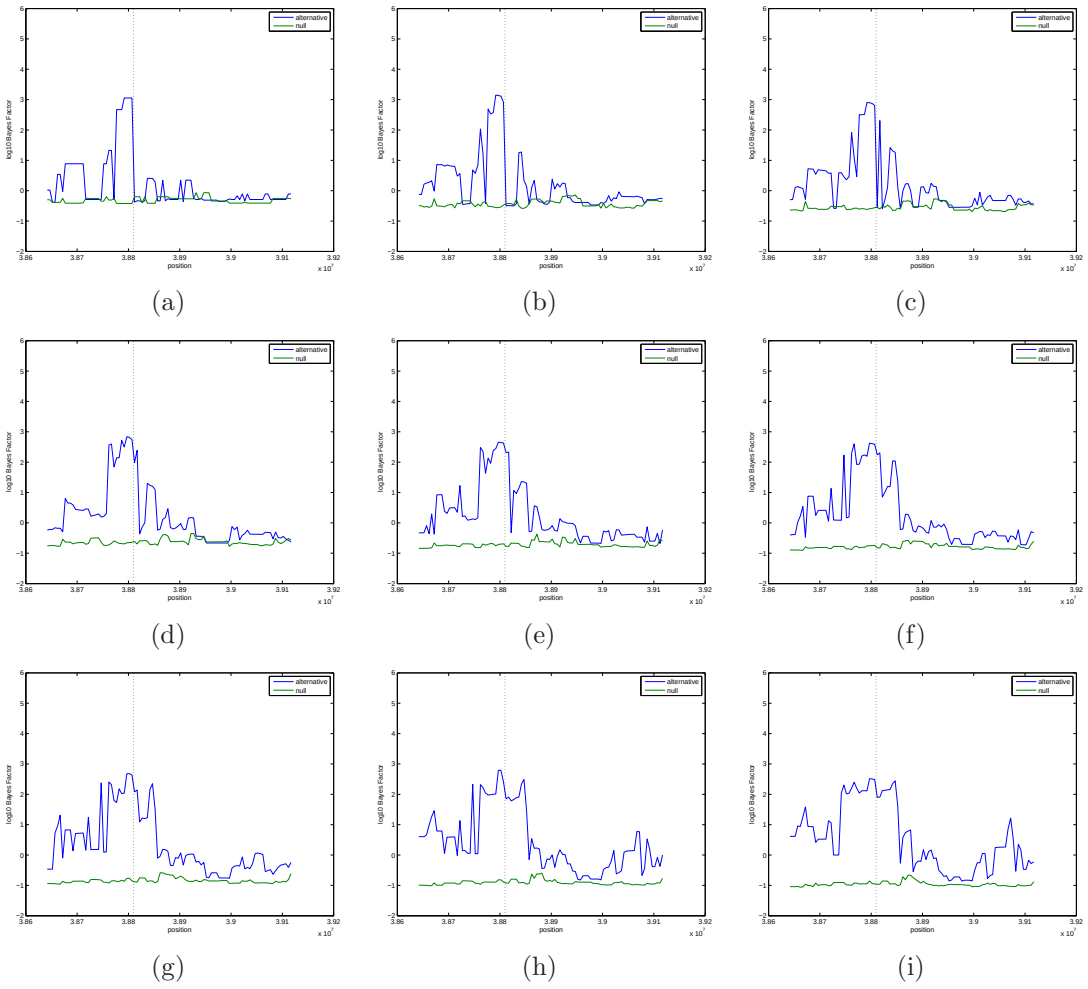


Figure A.2: HAPCLUSTER Bayes Factors for a single data set simulated from the ENCODE region ENr123. The x-axis is physical location and the y-axis is the $\log_{10}$ Bayes Factor. The blue lines indicate the test statistics for the disease data set and the green lines indicate the test statistics for the corresponding null data set. The number of basis clusters are **a** 2, **b** 3, **c** 4, **d** 5, **e** 6, **f** 7, **g** 8, **h** 9 and **i** 10. The dotted vertical line indicates the position of the disease SNP.

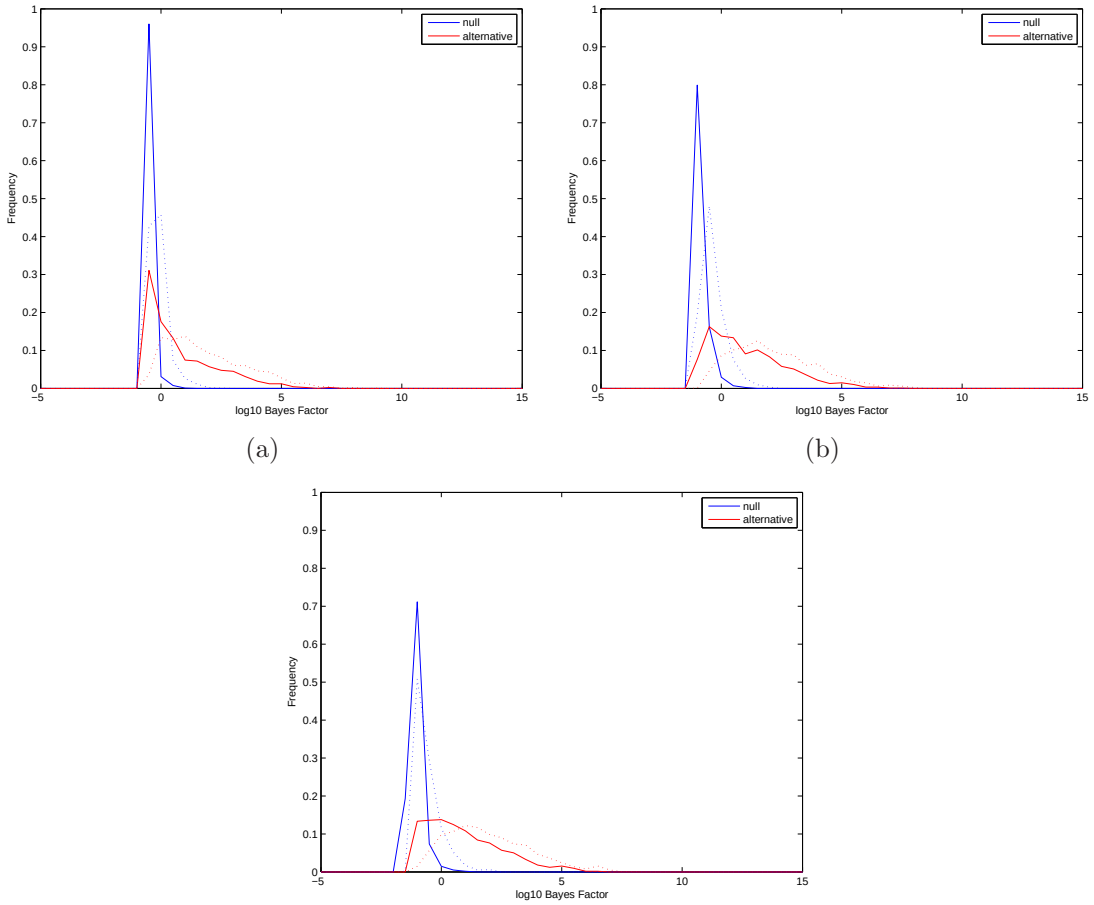
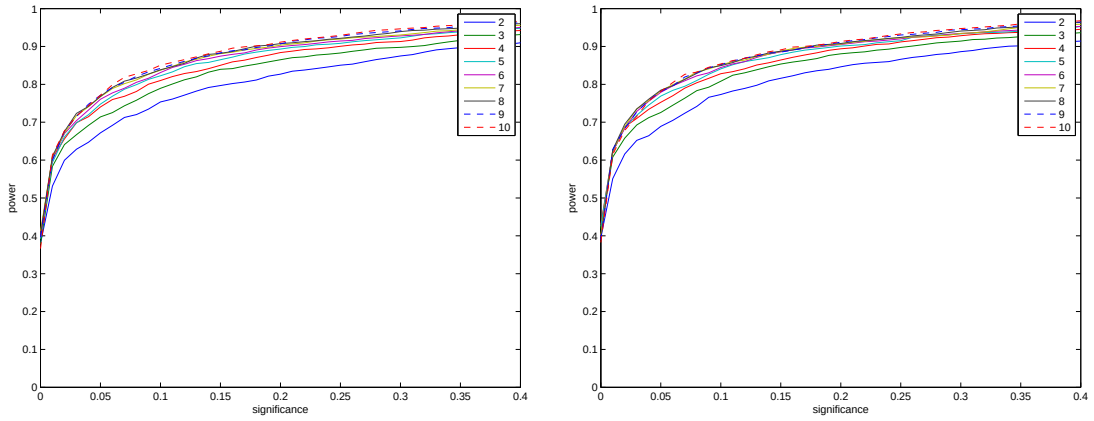
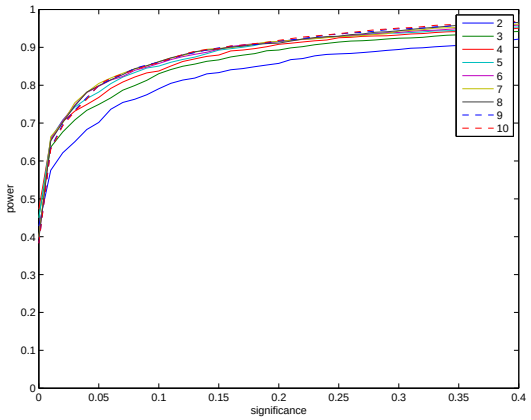


Figure A.3: The empirical null and alternative distributions of HAPCLUSTER summary test statistics, using **a** 2, **b** 6 and **c** 10, basis clusters for data sets generated from the ENCODE region ENr123. The solid lines indicate the mean Bayes Factors and the dotted lines indicate the mode Bayes Factors.



(a)

(b)



(c)

Figure A.4: The power of HAPCLUSTER using penetrance priors **a** $\beta(0.6, 0.6)$, **b** $\beta(1.0, 1.0)$, **c** $\beta(2.0, 2.0)$ and the mean Bayes Factor as the summary test statistic. The results are averaged over the 3 ENCODE regions that the data sets were generated from.

Appendix B

Supplementary Figures For Chapter 5

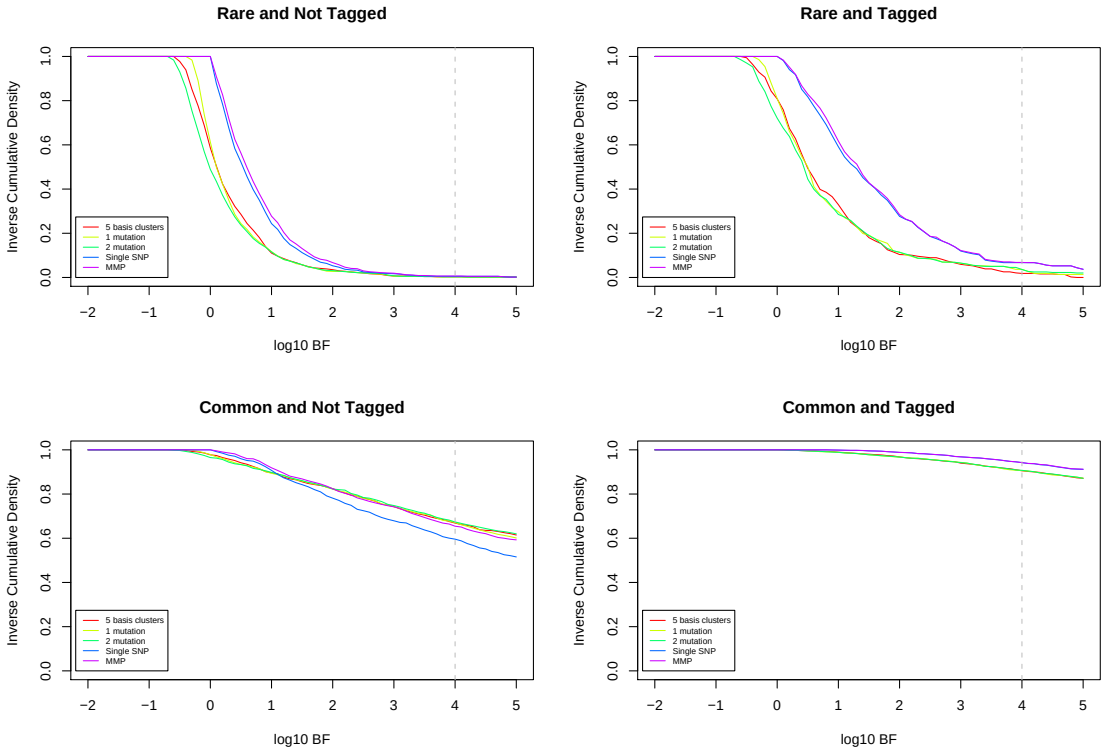


Figure B.1: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation model, 2-mutation model from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with a single causal variant of relative risk 1.5. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results are averaged over the 5 regions, which we simulated the data sets from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

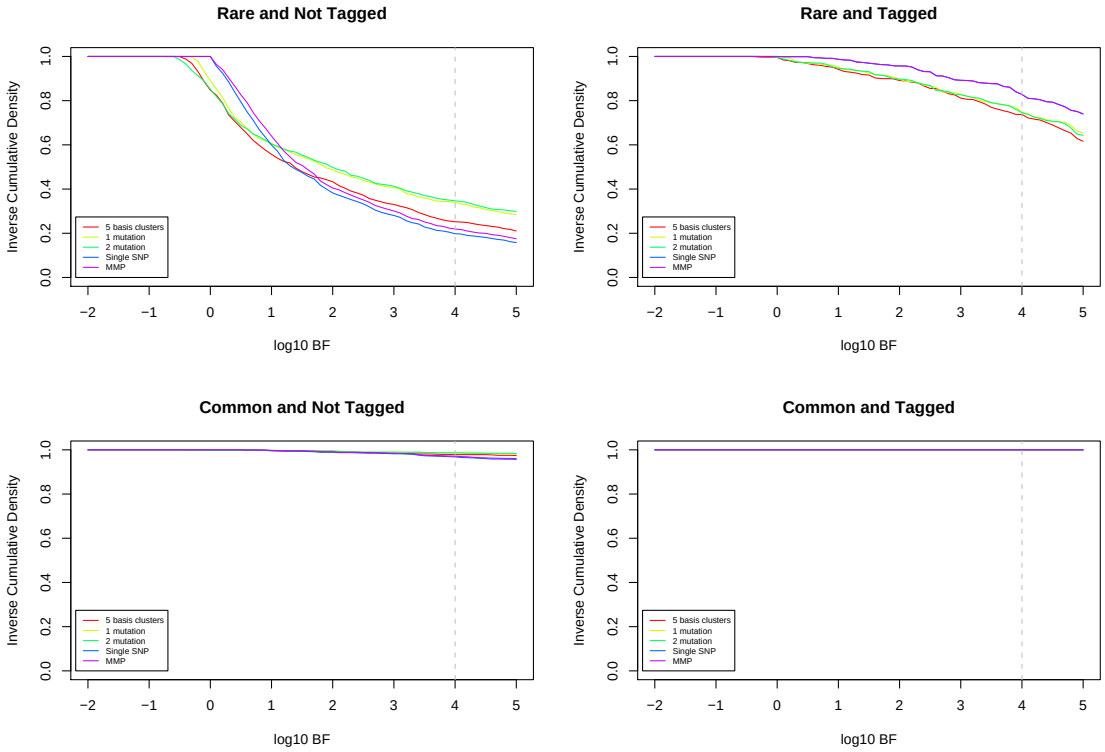


Figure B.2: The empirical inverse cumulative density function of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with a single causal variant of relative risk 2.5. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates the proportion of data sets with a maximum Bayes Factor exceeding the threshold specified on the x-axis, which is on the $\log_{10}$ scale. The results are averaged over the 5 regions, which we simulated the data sets from. The vertical dotted line indicates the 10^4 Bayes Factor threshold.

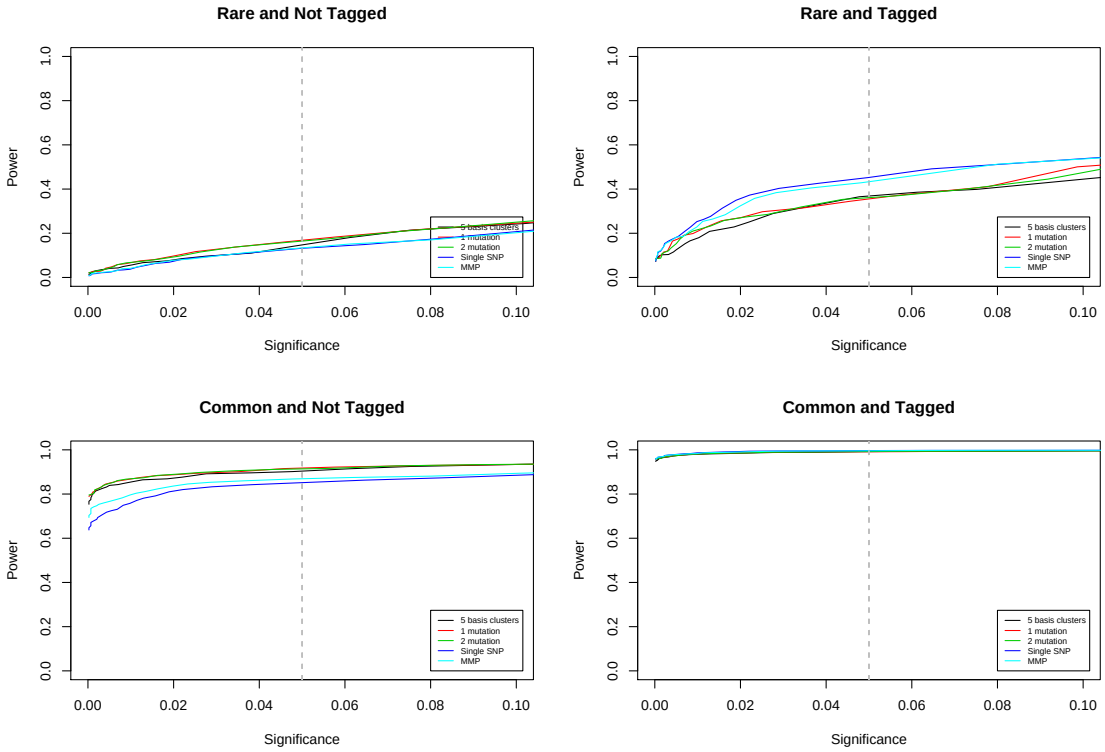


Figure B.3: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with a single causal variant of relative risk 1.5. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity). The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 5% significance threshold.

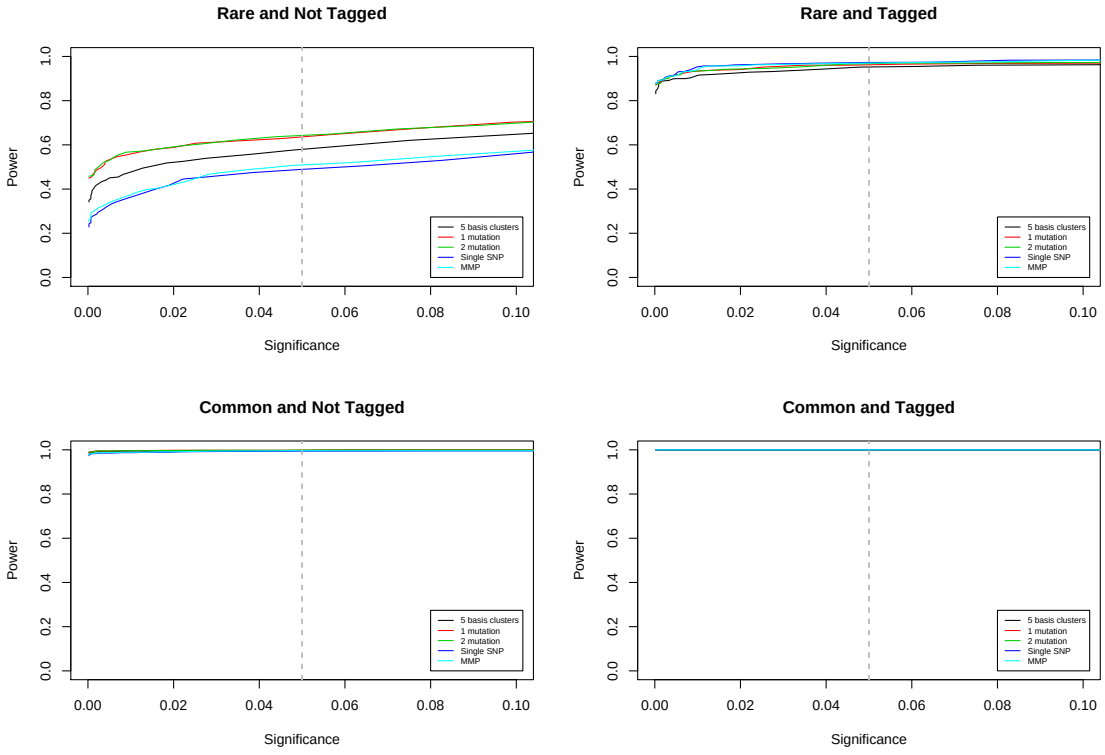


Figure B.4: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with a single causal variant of relative risk 2.5. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity). The results are averaged over the 5 ENCODE regions, which we simulated the data sets from. The vertical dotted line indicates the 5% significance threshold.

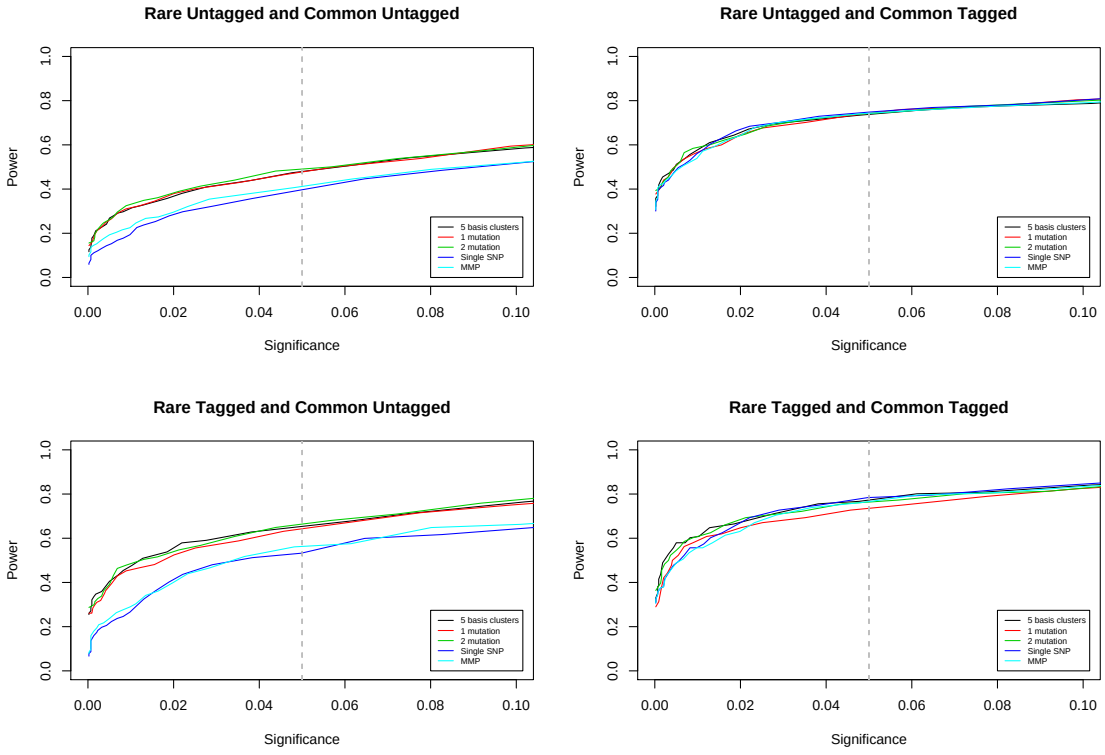


Figure B.5: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with 2 causal variants, one rare and one common, of relative risks 1.5 and 1.3, respectively. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity). The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 5% significance threshold.

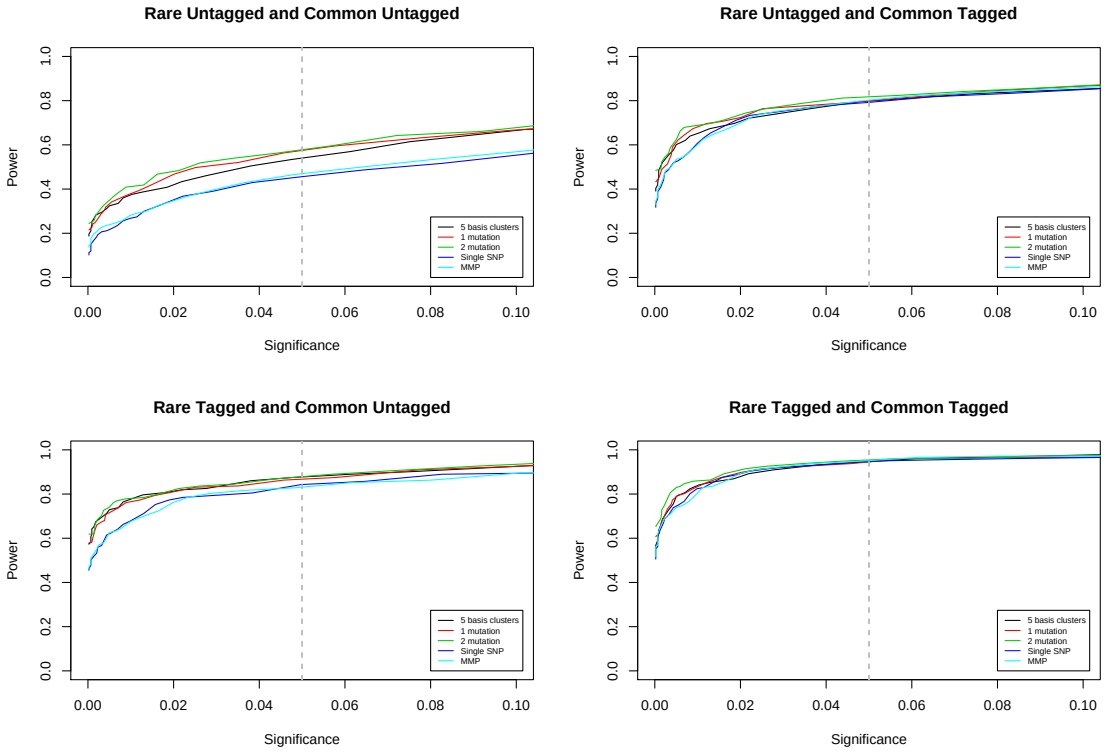


Figure B.6: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with 2 causal variants, one rare and one common, of relative risks 2.0 and 1.3, respectively. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity). The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 5% significance threshold.

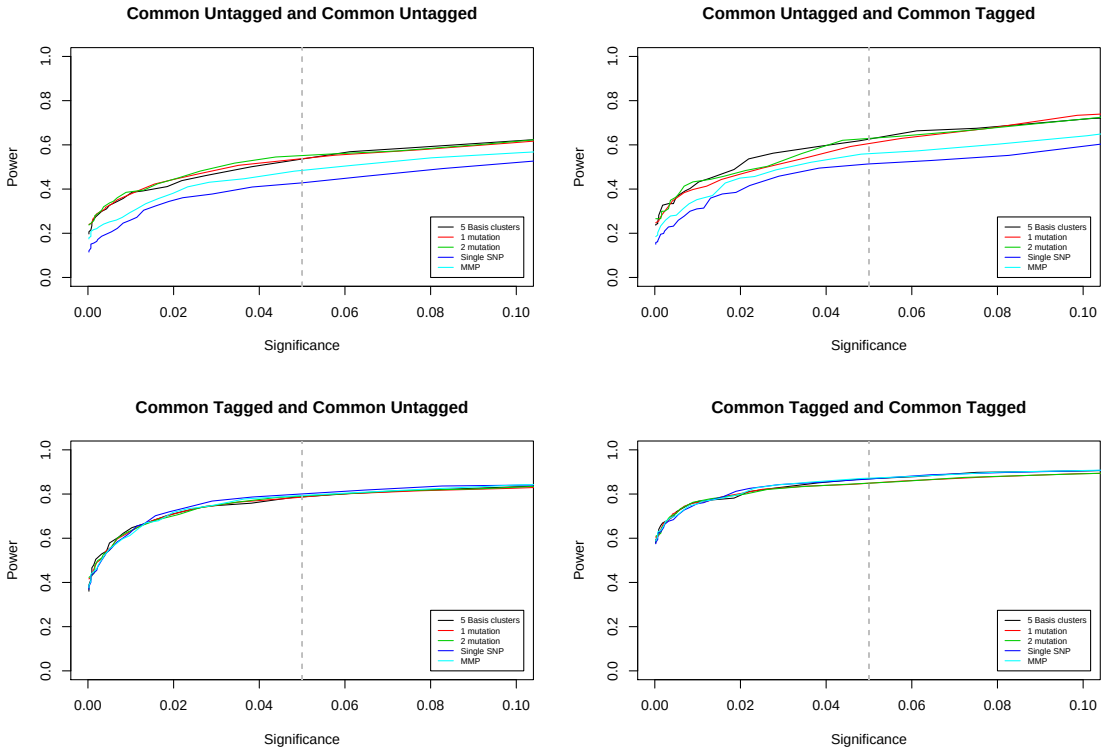


Figure B.7: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with 2 causal variants, both common, of relative risks 1.3 and 1.1. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity). The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 5% significance threshold.

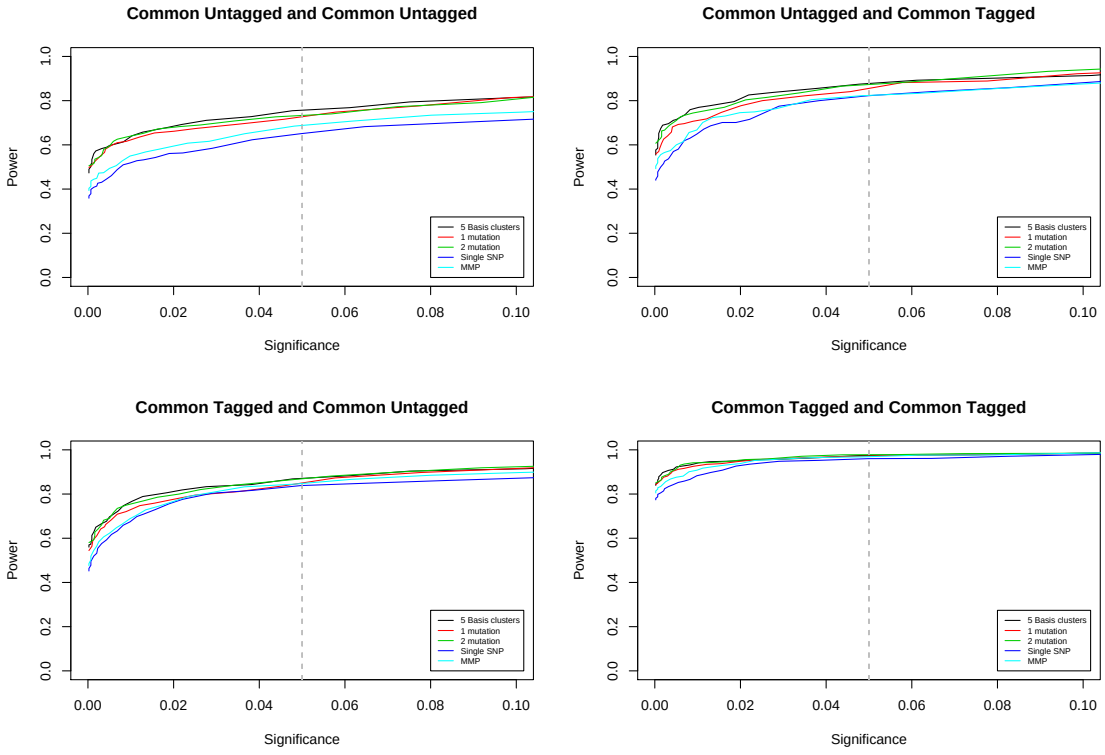


Figure B.8: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with 2 causal variants, both common, of relative risks 1.3 and 1.3. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity). The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 5% significance threshold.

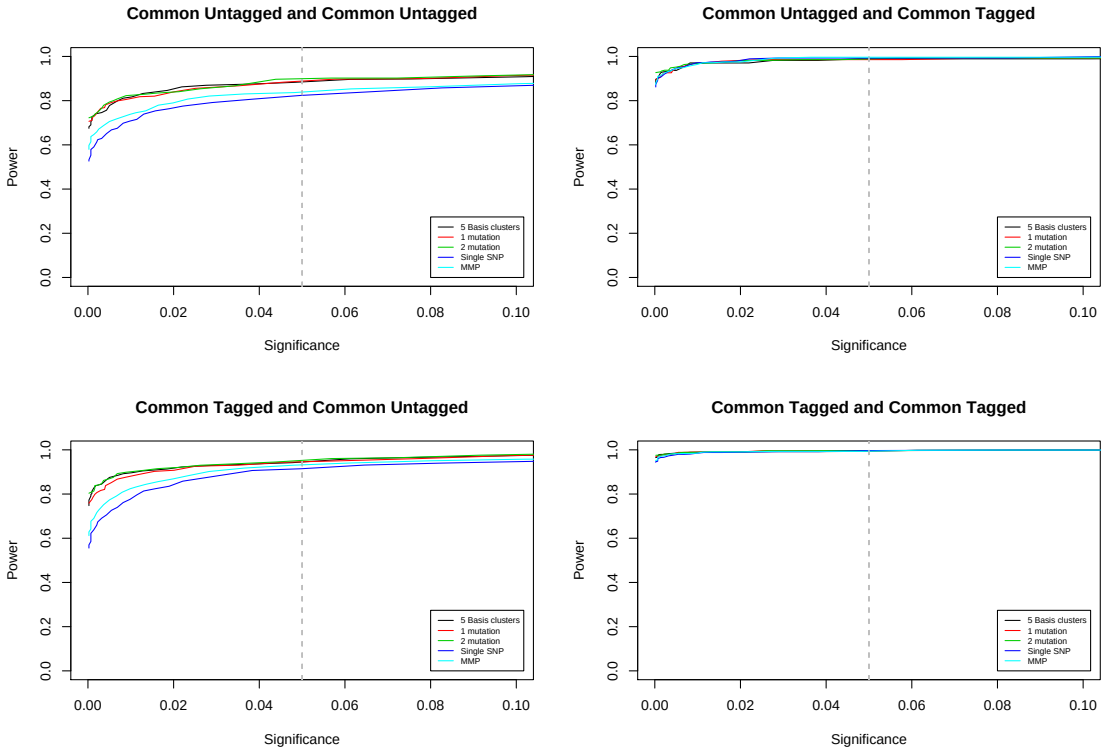


Figure B.9: The Receiver Operating Characteristics (ROC) of the maximum Bayes Factor for GENECLUSTER and SNP based approaches. For GENECLUSTER, we used the 1-mutation and 2-mutation models from model A and 5 basis clusters from model B; for the SNP based approaches, we performed a Bayesian single SNP test at each Affymetrix SNP and also at SNPs predicted by a MMP. We have analysed data sets with 2 causal variants, both common, of relative risks 1.3 and 1.5. Each panel of the figure refers to the 4 types of causal variants that we analysed. The y-axis indicates power (sensitivity) and the x-axis indicates significance (1-specificity). The results for each type of causal variants are averaged over the ENCODE regions, which the data sets were simulated from. The vertical dotted line indicates the 5% significance threshold.

Appendix C

TREESIM

Below is a description of how a tree, T, for a set of haplotypes $\mathbf{H}$ is constructed by the TREESIM method (Cardin, 2007).

The trees are built using the coalescent with recombination and approximate the posterior modal tree given the haplotypes. To do this it is useful to be able consider $\Pr(T|\mathbf{H})$ under the coalescent model. Using Bayes Formula it is possible to rewrite this as: $\Pr(\mathbf{H}|T)\Pr(T)/\Pr(\mathbf{H})$. Although it is simple to calculate these values under the coalescent with simple mutation models it is not known how to simulate directly from this distribution, or how to produce trees which maximise this expression.

To make this task simpler it is useful to factorise this expression into the individual events that make up the tree. It is useful to note that trees augmented with mutation track the haplotypes backwards in time, and after each event occurs these haplotypes change. Also, note that $\Pr(T) = \prod_i \Pr(E_i)$ where i indexes the events backwards in time and E_i is the ith event. Also $\Pr(T|\mathbf{H}) = \prod_i \Pr(E_i|\mathbf{H}_i)$ where $\mathbf{H}_i$ denotes the haplotypes as changed by the first i events. Then, note that $\Pr(E_i|\mathbf{H}_i) =$

$\Pr(\mathbf{H}_i|E_i)\Pr(E_i)/\Pr(\mathbf{H}_i)$.

It is not known how to calculate $\Pr(\mathbf{H}_i)$ directly. However, as the coalescent is Markov backwards in time $\Pr(\mathbf{H}_i|E_i)$ (the probability of the haplotypes given that the next event backwards in time is E_i) is equal to $\Pr(\mathbf{H}_{i+1})$ (the probability of the haplotypes as changed by the event E_i). So to calculate $\Pr(E_i|\mathbf{H}_i)$ it is only necessary to calculate $(\Pr(\mathbf{H}_{i+1})/\Pr(\mathbf{H}_i))\Pr(E_i)$. For all types of events (coalescence, recombination or mutation) the quotients $\Pr(\mathbf{H}_{i+1})/\Pr(\mathbf{H}_i)$ simplify to give terms of the form $\Pr(h_{n+1}|h_1,\ldots,h_n)$ where h_j denotes a single haplotype. These terms still cannot be calculated efficiently under the coalescent, however they are amenable to approximation using Hidden Markov Models, such as the Li and Stephens (2003) Hidden Markov Model.

Once these values can be approximated it is possible to generate a tree which approximates the modal posterior tree as follows:

1 Initialise: Decide on mutation model, recombination rates, and initialise the haplotypes $\mathbf{H}_0$ as the set of known haplotypes input to the method.

2 Recursion (Steps 2 through 6): Enumerate all possible events that may be the next event backwards in time.

3 For each of these events calculate $\Pr(E_i|\mathbf{H}_i)$, the posterior probability of each event, as described above.

4 Choose the event with the highest posterior probability.

5 Generate haplotypes $\mathbf{H}_{i+1}$ by applying the chosen event to haplotypes $\mathbf{H}_i$.

6 Stop: When each locus has reached its common ancestor the process termi-

nates.

We used the recombination rates estimated from the HapMap, and an infinite sites mutation model for the analyses in this thesis.