University of Oxford

Extensions of the Case–Control Design in Genome-wide Association Studies

by

Charalambos Loizides

St. Hugh's College

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford.

October 2011

Department of Statistics, 1 South Parks Road, Oxford

Abstract

The case-control design is one of the most commonly used designs in genome-wide association studies. When we increase the sample size of either the controls or, more importantly, the cases, the power of whatever test we use will certainly increase. However increasing the sample size, means that additional individuals need to be genotyped and this implies extra financial costs. However, nowadays with the emergence of genetic studies, a large number of genetic data are available at low or no extra cost. Even though those data may not be completely relevant to the current study, they can still be used to increase the probability to identify true associations. Furthermore, additional information, non-necessarily genetic, can also be used to improve the power of a method.

In this thesis we extend the case-control design in order to take advantage of such types of additional data and/or information. We discuss three designs; the case-cohort-control, the kin-cohort and the super-case–case–control–super-control designs. For each of these, we present methods that are adjusted or modified versions of standard case-control methods but we also propose novel ones developed with those extended designs in mind. Ultimately, we describe how those methods can be used in order to increase the power of association tests, especially compared to similar methods of the case-control design.

Acknowledgements

I would like to thank my supervisor Chris Holmes for his help and guidance. I would also like to record my gratitude to Jean-Baptiste Cazier for his valuable guidance and advice. Many thanks to Ian Tomlinson for providing the genetic data I used throughout my studies; credit for this should also go to Jean-Baptiste. The majority of those data are used in this thesis. Special thanks Christopher Yau and George Nicholson for their help with Chapters 4 and 5 respectively.

Last but not least I would like to thank my family and friends in Cyprus for their moral support and encouragement, especially during the most difficult periods of my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Genetic association studies aim to detect associations between one or more genetic polymorphisms and a trait, which may be as simple as a binary trait, for example a disease; or more complex quantitative traits (Cordell and Clayton, 2005). In the last few years it has become evident that genetic susceptibility to many common disorders involves many genes, most of which have small effects. Thus in order to identify all the possible genetic risk effects associated with a certain disease it is necessary to test for association across the whole genome. Such studies are called genome-wide association studies (GWAS). The genetic data that are typically used in GWAS are markers from across the genome; the number of markers varies from study to study. Advancements in genotyping allow us to use more markers now than few years ago but also financial costs is another important factor that affects the number of markers in a study.

The most common type of genetic markers that are currently used in

GWAS are the single nucleotide polymorphisms (SNPs) which are the type of polymorphism involving variation of a single base pair. (Collins, url). Usually a subset of all known SNPS is chosen to be genotyped and/or used in a study. These chosen SNPs are considered representative of a region in the genome and they are called tag SNPs. There are many studies regarding the optimal selection of tag SNPs (e.g. Stram, 2004) however these are beyond the scope of this thesis. In most occasions GWAS use single-marker (single-SNP particularly) statistical tests in order to identify direct or even indirect associations. Since the tag-SNPs represent loci or other regions in the genome, they might not be directly associated with the disease. The use of independent tests for potentially non-independent markers, mainly due to linkage disequilibrium (LD), does not cause any problem. On the other hand it can be used to verify the results since we expect markers that are in strong LD to provide consistent results. Additionally multistage analyses (Thomas et al., 2009) can use an even smaller subset of "promising" tag-SNPs for the second or subsequent phases, where additional individuals are genotyped, in order to replicate disease associations from the first phase.

Traditionally GWAS use retrospective designs although there are few studies that have departed from that with the use of prospective designs using, for example, a cohort model (Manolio, 2009). When the trait of interest is a binary trait, in most cases a disease, the most prominent retrospective design is the case-control design. This design involves a sample of unrelated affected individuals (cases) and a sample of unrelated unaffected individuals (controls), although that is not always true. In many occasions, especially

for relatively rare diseases, the controls are selected from the overall population without knowledge of their phenotype. The inference is based on the genotypic differences of these two samples. Family-based studies (e.g. family trios, case-sibling) is another example of retrospective design where individuals, usually affected, and members of their families are selected in the sample.

In this thesis we develop methods that can be used in designs that extend the traditional case-control design. The continuously decreasing genotyping costs, the emergence of biobanks and the already available genetic data from the numerous association studies that took place in the last decade, enable us to obtain additional data samples with little or no extra cost. The information provided by the data may vary. For example a sample from the population that is typically used as a control group does not actually provide any information about the phenotype of the individuals in this sample. On the other hand, a sample of, say, cases obtained from a family-based study will provide information not only about the individuals but their families as well. Essentially, these different types of data give rise to new designs that do not simply incorporate cases and controls. The purpose of this thesis is to investigate and develop methods that can be used under three of those designs, by incorporating the information provided from the additional data, in order to increase the probability of identifying true associations.

1.1 Organisation of the Thesis

In chapter 2 we provide a general overview of GWAS and introduce the major variables, models and other quantities that we use throughout this thesis. In chapter 3 we describe the case-cohort-control design. In that chapter we mostly focus on power analyses with respect to the parameters commonly involved in GWAS. In chapter 4 we present a novel bayesian method that can be used under the kin-cohort design. In chapter 5 we introduce the super-case–case–control–super-control (SCCS) design. We describe a number of methods, both bayesian and frequentist that can be used under this design. Finally, in chapter 6 we provide a brief summary of this thesis and discuss our plans for relevant future work.

Chapter 2

Overview of GWAS

Although GWAS are by no means confined to the case-control design, it is by far the most common. In this chapter we focus on this design and specifically in case-control studies with SNP genotyped subjects. Specifically, we consider the class of GWAS that single-SNP tests are applied across the genome. We present a brief outline of the theory behind case-control studies, the common parameters and variables that are involved and some examples of GWAS.

2.1 Parameters and Variables

For the rest of this thesis we only consider biallelic SNPs which comprise the vast majority of common SNPs. Also the use of the rare allele in many of our definitions is not mandatory but rather out of convenience. We may as well use the common allele instead, with no need for any change. All the

methods we present are symmetric with respect to the allele we choose to code 0 or 1.

2.1.1 Genotype, Allele Frequencies and the Phenotype

For a given SNP let $x \in 0, 1, 2$ denote the number of copies of the rare allele. We call x the genotype of that individual. Obviously for $x = 0$ and $x = 2$ the individual is homozygous to the common and rare allele respectively and it is heterozygous for $x = 1$. With no further assumptions the distribution of the genotype is trinomial with two degrees of freedom,

$$x \sim \text{Multi}(1, g_0, g_1, g_2) \quad \text{where} \quad g_i = P(x = i). \tag{2.1}$$

However it is common to assume that Hardy-Weinberg equilibrium (HWE) holds in the population (Hardy, 1908; Mayo, 2008). Essentially the HWE assumption implies that the alleles in each chromosome have independent and identically distributed Bernoulli distributions. Thus the genotypic distribution, now depends in only one parameter, θ; the probability to have the rare allele in the haplotype.

$$g_x \equiv g_x(\theta) = \begin{cases} (1-\theta)^2 & , \quad x = 0 \\ 2\theta(1-\theta) & , \quad x = 1 \\ \theta^2 & , \quad x = 2 \end{cases} \tag{2.2}$$

We call the parameter θ, the minor allele frequency (MAF) of the population. Throughout this thesis, unless explicitly stated, we will always assume that HWE holds.

In GWAS and especially in case-control studies the trait of interest is almost always a disease. Thus the phenotype, Y, is usually binary with $Y = 1$ denoting the presence of the disease in an individual and $Y = 0$ its absence. For a given population we denote $K = P(Y = 1)$ the disease prevalence in that population. In our methods, when necessary, we assume that we have a precise enough estimate of the population disease prevalence, safe to assume that it is known.

Let $f_x = P(Y = 1|x)$ denote the probability of an individual having the disease conditional on its genotype. Alternatively we can think of f_x as the risk for an individual to have or develop the disease, given that all we know about that individual is his/her genotype. We usually assume a deterministic model for f_x which we call disease model. We describe these models in more detail later in this chapter. Using simple probability theory we can calculate the conditional allele frequencies, p_x and q_x, of the sub-populations of affected and unaffected individuals respectively, in terms of g_x, K and f_x.

$$p_x = P(x|Y = 1) = \frac{g_x f_x}{K}, \quad q_x = P(x|Y = 0) = \frac{g_x(1 - f_x)}{1 - K} \tag{2.3}$$

2.1.2 Disease Models and the Penetrance Parameter

As we mentioned earlier, the probability of disease, conditional on the genotype, is usually modelled as a function of the genotype., $f_x = f(x, \eta)$. We call these functions disease models. In $f(x, \eta)$, η denotes a number of parameters which are used to determine a model. Although the number and type of parameters may vary from model to model, there is one parameter, β, which we call the penetrance parameter, that is necessary for all the models. It is the parameter that determines the type and magnitude of the genotype's effect on f_x. In the following paragraphs we present five of the most widely used disease models.

The **logistic disease model** uses the logistic function in the same fashion as the logistic regression does.

$$f_x = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}, \quad \beta \in \Re \tag{2.4}$$

When this model is used in a retrospective design; a case-control study for example, then it is indeed logistic regression. However when it is used prospectively, for example in simulations, then it is a deterministic model and the output is simply the risk of disease.

The **additive disease model**,

$$2f_1 = f_0 + f_2 \quad \text{and} \quad f_1 = \beta f_0 \quad (\Leftrightarrow f_2 = (2\beta - 1)f_0), \quad \beta > 0, \tag{2.5}$$

and the **multiplicative disease model**,

$$f_x = \beta^x f_0 \quad \beta > 0, \tag{2.6}$$

are, as is the case of the logistic model, strictly monotonic when there is true association.

The **recessive disease model**,

$$f_0 = f_1 \quad \text{and} \quad f_2 = \beta f_1, \quad \beta > 0, \tag{2.7}$$

and the **dominant disease model**,

$$f_1 = f_2 \quad \text{and} \quad f_1 = \beta f_0, \quad \beta > 0, \tag{2.8}$$

we use are probably the simplest models that can be defined to take into account dominant or recessive effects. More complex such models exist. For example if we set $x^* = I_{[x=2]}$ or $x^{**} = I_{[x>0]}$ and use the logistic model with x^* or x^{**} instead of x we obtain a more complex recessive or dominant model respectively.

The parameter α in (2.4) determines the baseline probability or risk of disease. This baseline risk is a consequence of all the other factors; genetic, environmental or other, that we do not consider in the model. Similarly f_0 in (2.5)-(2.8) is exactly this baseline risk. We can think of f_0 as the expected disease prevalence in the population if the SNP we study did not exist.

Clearly the equality $K = \sum_{x=0,1,2} g_x f_x$ must hold for all the models we

presented. Using that we can express the baseline probabilities of disease of the additive , multiplicative, recessive and dominant models in terms of K, θ and β (Table 2.1).

Model	f_0
Additive	$\frac{K}{1+2(\beta-1)\theta}$
Multiplicative	$\frac{K}{(1+(\beta-1)\theta)^2}$
Recessive	$\frac{K}{1+(\beta-1)\theta^2}$
Dominant	$\frac{K}{1+2(\beta-1)\theta+(\beta-1)\theta^2}$

Table 2.1: Baseline disease probabilities of the additive, multiplicative, recessive and dominant disease models.

For the logistic model $f_0 = (1+\exp(\alpha)^{-1})^{-1}$. A closed form solution of α (or $\exp(\alpha)$) in terms of the other three parameters does not exist. Instead, numerical methods can be used to calculate α and consequently the baseline risk of the logistic model.

In GWAS and especially in prospective analysis one of the aforementioned disease models is usually adopted. However, methods that do not require an explicit knowledge of the model that generated the data have also been developed. For example Freidlin et al. (2002) study robust test statistics in the case where the disease model is unknown.

2.1.3 Conditional Allele Frequencies

As we indicated in (2.2), under HWE, the population MAF cannot only be interpreted as the proportion of rare alleles in the population but also as the single parameter that fully determines the distribution of the genotypes for that particular SNP. In most of our applications, however, we will not use the

population MAF but rather the allele frequencies of subpopulations which can be expressed as conditional allele frequencies. The conditions involve either the phenotype or more generally, some sort of disease history in the individual's family. Henceforth whenever we refer to a subpopulation we will imply one that it is determined by one of those conditions. An example of this, is the allele frequency of the subpopulation of affected individuals in a population. Staying firm on what we discussed in section 2.1.1, we consider the conditional rare allele frequency, that is the conditional MAF.

In general, the conditional MAF does not determine the distribution of the genotypes in the subpopulation but rather these have a trinomial distribution. The conditional MAF can be interpreted only as the proportion of rare alleles in the subpopulation. Conditional MAFs will be still denoted by θ albeit with a subscript.

$$\theta_C = P(x = 2|C) + \frac{1}{2}P(x = 1|C), \quad \text{for a condition } C \tag{2.9}$$

If $X = (x_1, \ldots, x_N)^T$ is the vector of the genotypes of a random sample from the entire population then the likelihood of θ is that of a binomial parameter.

$$L(\theta|X) = \binom{2N}{\sum x_i} \theta^{\sum x_i} (1-\theta)^{2N - \sum x_i} \tag{2.10}$$

This is because the HWE implies that the two haplotypes of a genotype (always regarding a single marker) are independent. This is not the case when we have a conditional minor allele frequency. For a subpopulation

$X_c = (x_{c,1}, \dots x_{c,N_c})^T$ the haplotypes are no longer independent and θ_c is not a parameter of the genotypic distribution. Instead, we have a trinomial likelihood of genotype frequencies.

$$L(g_{1|c}, g_{2|c}|X_c) = \binom{N}{\sum I_{[x_i=0]}, \sum I_{[x_i=1]}, \sum I_{[x_i=2]}} (1 - g_{1|c} - g_{2|c})^{\sum I_{[x_i=0]}} g_{1|c}^{\sum I_{[x_i=1]}} g_{2|c}^{\sum I_{[x_i=2]}}, \quad (2.11)$$

where $g_{k|c} = P(x = k|C)$.

Nevertheless, in most of our methods we will use binomial likelihoods for the conditional MAFs. The truth is that, usually, the distribution of the genotypes in a subpopulation deviates very little from a single parameter distribution in the style of (2.2). Valid inference can still be drawn when we "approximate" the trinomial distribution in (2.11) with a binomial. The advantages of using a simple likelihood, reducing the number of unknown parameters, are more considerable than the potentially small improvement in the accuracy of our model. An improvement which is not guaranteed since the need to estimate an additional parameter for each subpopulation adds more uncertainty to the results.

2.2 Testing for Association

In the majority of GWAS the data are collected retrospectively. That is, from the sub-population of the unaffected individuals we randomly choose n_0 controls and from the sub-population of the affected we choose n_1 cases.

Normally this sampling design would lead us to adopt a retrospective likelihood, $P(x|Y)$, the obvious choice here being the binomial. Hence the test hypotheses in this retrospective design involve the conditional MAFs θ_0 and θ_1, of the unaffected and affected sub-populations respectively.

$$H_0 : \theta_0 = \theta_1 \qquad H_1 : \theta_0 \neq \theta_1 \tag{2.12}$$

As a side note, it is worth mentioning that in accordance to what we discussed in the end of the previous section (2.1.3), we can use the "more appropriate" trinomial likelihood. The statistical hypothesis in this case will involve four parameteres in total instead of two. This practice is discouraged for this and the other reasons we discussed previously.

There are various test statistics that can be used to test the hypotheses in (2.12). These can be likelihood-based, for example a likelihood ratio test or non-likelihood-based. All Bayesian tests fall, by default, in the likelihood-based methods since the likelihood of the data is a requisite for all Bayesian methods. Non-likelihood-based tests are usually based on a 2×2 allele table of counts where each individual contributes two haplotypes and thus the sum of counts is twice the sample size. This design is commonly used to estimate a measure of the effect size such as the odds ratio (OR) or relative risk (RR).

When studying dominant or recessive effects all the tests can be easily adjusted by using ϑ_i, $i = 0, 1$ instead of θ_i, where $\vartheta_i = P(x > 0|Y = i)$ for dominant effect and $\vartheta_i = P(x = 2|Y = i)$ for recessive effect. In this case the binomial likelihood is exact and not approximate. Also non-likelihood-based

methods use a 2×2 genotype table of counts (although now the genotypes are coded $\{0, 1\}$, see section 2.1.2) and thus the total sum is equal to the sample size.

Nevertheless, despite the retrospective nature of the sampling scheme, methods based on a prospective likelihood $P(Y|x)$ can also be valid. Prentice and Pyke (1979) showed that the log-odds ratio estimators and their respective variances can be obtained by applying the original logistic regression model (a model that uses prospective likelihood) to the case-control study as if the data had been obtained in a prospective study. Seaman and Richardson (2004) proved a similar result for the posterior distribution of the log odds ratios in a Bayesian analysis, again suggesting the simpler logistic model that treats the data as as though generated prospectively and which does not involve nuisance parameters for the exposure distribution. Carroll et al. (1995) generalise this prospective formulation of case-control studies further to include multiplicative models, stratification, missing data, measurement error and robustness.

Similar to the retrospective methods, prospective methods can also be likelihood or non-likelihood-based. Likelihood-based methods typically assume one of the disease models we describe in section 2.1.2. The statistical hypotheses and consequently the statistical inference are based on the penetrance parameter.

$$H_0 : \beta = \beta_0 \qquad H_1 : \beta \neq \beta_0 \tag{2.13}$$

In (2.13), β_0 denotes the value of the penetrance parameter when the marker has no effect. For example $\beta_0 = 0$ for the logistic disease model and $\beta_0 = 1$ for the other four models we describe. In practice the logistic model is the one that is almost always used since estimates of β and its variance can be straightforwardly obtained. It is a good proxy for additive and multiplicative models while with simple adjustments, it can be used to study dominant or recessive effects (section 2.1.2).

On one hand we saw that the results obtained from the logistic model are equivalent to those obtained from retrospective methods such as the odds-ratio (OR) and relative risk (RR). Specifically it can be easily proven (Appendix A) that

$$\mathrm{OR} = \exp(\beta), \tag{2.14}$$

$$\mathrm{RR} = \frac{\exp(\beta) + \exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)} \tag{2.15}$$

where obviously OR=RR=1 if and only if $\beta = 0$. On the other hand using the logistic model allow us to consider additional variables in our analysis, whether they be genetic markers (multiple - marker test) or environmental, physical or behavioural characteristics of the individuals. This flexibility favors the use of the logistic model in many association studies.

Trend tests are the most common non-likelihood-based prospective methods. The most prominent of those is the Cochran-Armitage test (Cochran, 1954; Armitage, 1955). It is a score test that tests the hypothesis of a zero slope that fits the three genotypic risk estimates best. In other words, the

hypotheses of the test, now, involve f_x, $x = 0, 1, 2$.

$$H_0 : f_0 = f_1 = f_2 \qquad H_1 : f_i \neq f_j \quad \text{for at least one pair } i \neq j \qquad (2.16)$$

The choice of scores is important when it comes to optimising the power of the test. The choice heavily depends on what kind of genetic effects we are looking for; additive, dominant, recessive or other more complex effects. Strictly speaking, the Cochran-Armitage test cannot be considered as a pure non-likelihood-based method since its connection with the ordinary linear regression is obvious. However it falls in this category because it is based on a 2×3 genotype table of counts, without the need of further distributional assumptions. In fact, even the HWE assuption is not required.

The main objective of all those statistical tests is to identify whether there is an association between a certain disease and a marker or a locus more generally. In other words they simply seek to answer and question with a positive or negative anwser, always in the context of statistics. However some of these tests can help us go deeper than that by providing estimates of a certain genetic effect size. Common effect sizes in GWAS are the OR, RR or the penetrance parameter β. Note that not all tests necesarily return an estimate of the effect size. Having a good estimate of the effect size is very important when we are interested in prediction, i.e. provided an individual's genotype we want to be able to estimate his/her risk of developing the disease. Estimating the effect size is a much more complex and difficult task than simply identifying an association. There are a lot of reported problems, both in estimating the effect size itself (Siegmund, 2002) and its use in prediction

(Jakobsdottir et al., 2009).

2.3 Notable Works

There is a plethora of methodologolical papers in the literature, describing both frequentist and Bayesian methods used in GWAS. Although frequentist methods (Balding, 2006) tend to be preferred nowadays due to their relative simplicity and more straightforward approach, Bayesian methods are still popular (Stephens and Balding, 2009; Beaumont and Rannala, 2004; Shoemaker et al., 1999).

As for actual genome-wide association studies, the most notable example is the work by the Wellcome Trust Case-Control Consortium (WTCCC) (Consortium, 2007). Other notable works, many of them by WTCCC include, Consortium et al. (2009), Samani et al. (2007), Parkes et al. (2007), Zeggini et al. (2007) and Todd et al. (2007). Worth mentioning here, is also a series of works concerning colo-rectal cancer, (Tomlinson et al., 2007; Broderick et al., 2007; Jaeger et al., 2008; Tomlinson et al., 2008; Pittman et al., 2008). We talk in more detail about those studies later in this thesis, where we actually use some of the data and results that the authors used.

Chapter 3

The Case-Cohort-Control Design

In case-control association studies the researcher needs to obtain data that are relevant with the disease or diseases that his/her study is concerned. This means that he or she will either need to have a group of cases and a group of controls genotyped, paying the relevant cost and use those as the case-control samples or he/she can obtain those samples from a previous study concerning the same disease through collaborators or other sources. In either case, one can obtain as many data as it is financially possible or as many data as the others are willing or can offer. However there is a large amount of data that can be easily accessible and at no extra cost, that at first sight may appear to have no use in the curent study. Examples of this include data from previous studies, blood donors or biobanks. In this chapter we study methods where we use this type of additional data in order to increase

the power of the association tests. The main assumption we make is that the additional data are not relevant to the disease we study.

We extend the traditional case-control design into what we call the case-cohort-control design. Although it may be conidered a misnomer we provide a clear definition of the three sample groups and thus avoid any confusion and conventionally adopt that name. The controls in our design are assumed to be a sample from the unaffected population. The individuals have been properly diagnosed and identified as unaffected. At this stage we do not consider any misdiagnosis errors. This type of sample is, sometimes, referred to as super-controls in many GWAS. The cases on the other hand are individuals that have positively been diagnosed as diseased. We consider them as a random sample from the sub-population of the affected individuals. It is the same type of sample that is used as cases in the vast majority of GWAS.

We do not attempt to create an entirely new framework to analyse case-cohort-control data. Instead we focus on the power analysis of various tests that are commonly used in case-control studies, only in our setting we merge the cohort and control sample groups together and treat this merged sample as if it was a the control sample in a case-control study. Since the majority of GWAS deal with relatively rare diseases or at least not very common ones we expect the cohort sample group to comprise mostly of unaffected individuals. Iniatially we use simulations to compare the power of three common association tests under the case-cohort-control design. Apart from comparing the tests to see which is the most powerful, we also investigate whether using the cohort sample always leads to an increase in power. We

show that in some occasions actually using the cohort sample leads to a decrease in power and we propose a hybrid test statistic which can be used whether inclusion of the cohort is advised or not.

Subsequently we study in more detail how the various parameters affect the power of the Cochran-Armitage (CA) test, which is the one that proved to be more powerful than the other tests we used in our simulations. We study how the values of these parameters affect whether the test is more powerful under the case-cohort-control than the case-control design. In order to achieve that we use the concepts of sensitivity and specificity and use an asymptotic formula for the power of the CA test.

An issue with the case-cohort-control design is population stratification. Specifically the additional data (cohort) may not come from the same exact population as the original data (cases and controls). This will lead to many false associations that are due to population heterogeneity rather than the disease. In order to deal with this issue we use a data set concerning colo-rectal cancer and take into account markers that have been successfully replicated and found to be associated with the disease. Based on our anlysis of this data we propose a two-stage association test that takes into account potential population heterogeneity.

3.1 Power Comparisons Using Simulations

In this section we use simulations in order to calculate the power of three tests, commonly used in association studies, under the case-control (CC) and

case-cohort-control (CCC) designs. As we mentioned earlier, a test under the CCC design is nothing more than the application of the case-control test, where the control sample group comprises both the original controls and the cohort.

There is a number of assumptions we need to make and parameters we need to specify in order to simulate our data. The first assumption concerns the disease model in the population. We consider four of the models discussed in section (2.1.2); the logistic, additive, recessive and dominant models. Additionally we assume HWE in the population and thus the distribution of the genotypes is fully determined by the MAF (θ). Additional parameters we have to specify are the disease prevalence (K), the penetrance parameter (β) and the sample sizes R, S and M of cases, controls and cohort respectively. In our study we assume that the cases and controls have the same sample size ($R = S = n$) and use $M = \lambda n$ for the cohort sample size.

For the purposes of our simulations, we first choose the values of θ, K, n and λ and then choose an equally spaced sequence of penetrance parameter values $\beta_0, \beta_1, \ldots, \beta_h$. The values β_i are chosen such that β_0 is the parameter value under the null hypothesis of no association and β_h is a value that results in power which is approximately equal to 1 for a pre-determined value of Type-I error, α_{T-I}. The latter means that even if we choose a β that is much larger than β_h, the power of the test for this new β will be roughly equal to the power of the same test for $\beta = \beta_h$ and almost equal to 1. Without loss of generality we choose the values of β_is such that the minor allele has a causal effect. Due to symmetry the same conclusions can

be reached if we consider protective markers.

Thus for our analysis we consider the power π as a function of β conditional on the rest of the parameters,

$$\pi(\beta) \equiv \pi(\beta|\theta, K, n, \lambda, \alpha_{T-I}). \tag{3.1}$$

There are three main objectives in our study. First we want to compare the power of the three tests we use, under both CC and CCC designs, and see whether one is always more powerful than the others. Second we want to see how the power of the tests changes when we adopt the CC or the CCC design. We want to see whether a test under the CCC design is always more powerful than its CC counterpart. The third objective is a direct result of the second one. We see that for certain parameter values the tests under the CCC design are more powerful while for others CC tests are actually more powerful. For this reason we seek a test statistic that will act as a compromise between the two designs, a statistic that can be used without having to à priori choose between the two designs and one that is still reasonably powerful.

3.1.1 The Test Statistics

Cochran-Armitage test

The Cochran-Armitage (CA) test (Sasieni, 1997) is a score test that tests the hypothesis of a zero slope for a line that fits the three genotypic risk estimates best. Being a non-parametric test, no parameter estimates are

used. Instead for the traditional case-control test we use the genotype counts. The tabulated data in Table 3.1 follow the notation of Sasieni (1997).

	Genotype			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Table 3.1: 2×3 table showing the generic genotype distribution in a case-control study.

The derivation of the CA test is based on the asyptotic distribution of U under the null hypothesis (Zheng et al., 2003),

$$U = \frac{1}{N} \sum_{x=0,1,2} z_x(Sr_x - Rs_x), \tag{3.2}$$

where $\boldsymbol{z} = (z_0, z_1, z_2)$ are increasing scores assigned to the three genotypes. Here we adopt the formula used by Sasieni (1997),

$$CA = \frac{N(N\sum_{x=0,1,2} r_x z_x - R\sum_{x=0,1,2} n_x z_x)^2}{RS\left\{N\sum_{x=0,1,2} n_x z_x^2 - \left(\sum_{x=0,1,2} n_x z_x\right)^2\right\}}. \tag{3.3}$$

Under the null hypothesis of no association, the CA statistic has an asymptotic χ_1^2 distribution. The exact choice of scores $\boldsymbol{z} = (z_0, z_1, z_2)$ is in general subjective but it is guided by the assumed disease model. In our analysis we use $z_x = x$ for the logistic and additive models, $\boldsymbol{z} = (0, 0, 1)$ for the recessive model and $\boldsymbol{z} = (0, 1, 1)$ for the dominant model. Zheng et al. (2003) showed that the scores we use are optimal in the case of logistic and additive model while the ones we use for the recessive and dominant models are

locally optimal.

When we extend the case-control to the case-cohort-control design (Table 3.2), the CA test is easily adjusted by substituting S in (3.3) with $S + M$ while the other quantities are obtained from Table 3.2.

	Genotype			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Cohort	m_0	m_1	m_2	M
Total	n_0	n_1	n_2	N

Table 3.2: The generic distribution of the genotypes in a case-cohort-control study.

Welch's t-test

Welch's t-test is an adaptation of Students t-test for the difference in means of two populations with different variance. The general form of Welch's t-statistic is

$$\mathcal{T} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.4}$$

where $\bar{X}_i$, s_i^2 and n_i are the i^{th} population's sample mean, sample variance and sample size respectively. Like the common t-test Welch's test statistic follows a t_ν distribution under the null hypothesis of equal population means. Unlike the common t-test the degrees of freedom (ν) of this distribution are

not necessarily a natural number and are approximated by the formula

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}. \tag{3.5}$$

Considering the hypotheses in (2.12) the Welch's t-statistic can be adjusted for the standard case-control association test.

$$\mathcal{T} = \frac{\hat{\theta}_p - \hat{\theta}_q}{\sqrt{\hat{V}(\hat{\theta}_p) + \hat{V}(\hat{\theta}_q)}} \tag{3.6}$$

This can be further adjusted for the CCC design with merged controls and cohort groups by substituting $\hat{\theta}_q$ with $\hat{\theta}_{q,g}$ in (3.6).

$$\begin{aligned} \hat{\theta}_{q,g} &= \frac{S\hat{\theta}_q + M\hat{\theta}}{S+M} \\ \hat{V}(\hat{\theta}_{q,g}) &= \frac{1}{(S+M)^2}\left(S^2\hat{V}(\hat{\theta}_q) + M^2V(\hat{\theta})\right) \end{aligned} \tag{3.7}$$

When the logistic or additive disease model is assumed, $\hat{\theta}$, $\hat{\theta}_p$ and $\hat{\theta}_q$ in (3.6) and (3.7) are the sample estimates of the population minor allele frequency $\theta = \frac{1}{2}g_1 + g_2$ and the conditional minor allele frequencies $\theta_p = \frac{1}{2}p_1 + p_2$ and $\theta_q = \frac{1}{2}q_1 + q_2$ of the affected and unaffected sub-populations respectively where g_x, p_x and q_x are genotype frequencies.

$$\begin{aligned} g_x &= P(x) \\ p_x &= P(x|Y=1) \\ q_x &= P(x|Y=0), \qquad x = 0,1,2 \end{aligned} \tag{3.8}$$

However when we assume the recessive or dominant model the parameter of interest is not the MAF. Although we keep the same notation, when we deal with these models we consider that $\theta = g_2$, $\theta_p = p_2$ and $\theta_q = q_2$ for the recessive model and $\theta = g_1 + g_2$, $\theta_p = p_1 + p_2$ and $\theta_q = q_1 + q_2$ for the dominant model. In any case, the estimates we use in the test statistic can be calculated using the maximum likelihood estimators of the genotype frequencies which in turn are easily obtained using the information in Table 3.2.

$$\hat{p}_x = \frac{r_x}{R} \quad \hat{q}_x = \frac{s_x}{S} \quad \hat{g}_x = \frac{m_x}{M} \tag{3.9}$$

For the logistic and additive models the sample variance estimates are calculated using the properties of the multinomial distribution, with the exception of $\hat{V}(\hat{\theta})$ where we use the binomial distrubution.

$$\begin{aligned} \hat{V}(\hat{\theta}_p) &= \frac{1}{4R}\left\{\hat{p}_1(1-\hat{p}_1) + 4\hat{p}_2(1-\hat{p}_1-\hat{p}_2)\right\} \\ \hat{V}(\hat{\theta}_q) &= \frac{1}{4S}\left\{\hat{q}_1(1-\hat{q}_1) + 4\hat{q}_2(1-\hat{q}_1-\hat{q}_2)\right\} \\ \hat{V}(\hat{\theta}) &= \frac{\hat{\theta}(1-\hat{\theta})}{2M} \end{aligned} \tag{3.10}$$

Similarly but much simpler, the sample variance estimates for the recessive (3.11) and dominant (3.12) models are calculated using properties of the

binomial distribution.

$$\begin{aligned} \hat{V}(\hat{\theta}_p) &= \frac{\hat{p}_2(1-\hat{p}_2)}{R} \\ \hat{V}(\hat{\theta}_q) &= \frac{\hat{q}_2(1-\hat{q}_2)}{S} \\ \hat{V}(\hat{\theta}) &= \frac{\hat{g}_2(1-\hat{g}_2)}{M} \end{aligned} \tag{3.11}$$

$$\begin{aligned} \hat{V}(\hat{\theta}_p) &= \frac{(\hat{p}_1+\hat{p}_2)(1-\hat{p}_1-\hat{p}_2)}{R} \\ \hat{V}(\hat{\theta}_q) &= \frac{(\hat{q}_1+\hat{q}_2)(1-\hat{q}_1-\hat{q}_2)}{S} \\ \hat{V}(\hat{\theta}) &= \frac{(\hat{g}_1+\hat{g}_2)(1-\hat{g}_1-\hat{g}_2)}{M} \end{aligned} \tag{3.12}$$

Although, as we mentioned, the test statistic $\mathcal{T}$ has t_ν distribution under H_0, with v given by (3.5); in our study we assume a standard normal distribution. The sample size in GWAS is large enough to guarantee that the normal approximation of the t distribution is extremely accurate.

Logistic Regression - Deviance test

The deviance test is used in generalised linear models for model selection. Specifically it is used as a decision criterion when we compare nested models. For a model M with fitted coefficients $\hat{\boldsymbol{\beta}}_M$, the deviance is calculated as (Hardin et al., 2007),

$$Dev(M) = 2\phi(l(y,y) - l(y,\hat{\boldsymbol{\beta}}_M)), \tag{3.13}$$

where ϕ is a scale parameter and depends on the link function used ($\phi = 1$ when the logistic link function is used), $l(y, y)$ is the log-likelihood of the full model and $l(y, \hat{\boldsymbol{\beta}}_M)$ is the log-likelihood of the fitted model. If we have two nested models M_0 and M_1 with $\hat{\boldsymbol{\beta}}_{M_0} \subset \hat{\boldsymbol{\beta}}_{M_1}$, then the test statistic we use to compare the two models is the difference of their respective deviances.

$$\mathcal{D} = Dev(M_1) - Dev(M_0) \tag{3.14}$$

In other words we test the hypothesis H_0 : choose M_0 against H_1 : choose M_1. Under the null hypothesis $\mathcal{D}$ has a χ^2 distribution with $d_1 - d_0$ degrees of freedom, where $d_i = \dim(\hat{\boldsymbol{\beta}}_{M_i})$, $i = 0, 1$.

In a genetic case-control association study the two competing models are

$$M_0 : \text{logit}\big(P(Y = 1|x)\big) = \alpha \qquad M_1 : \text{logit}\big(P(Y = 1|x)\big) = \alpha + \beta x. \tag{3.15}$$

Obviously under this setting the asymptotic distribution of $\mathcal{D}$ is χ^2_1. When we adopt the CCC design we simply set all the otherwise unknown phenotypes (Y) of the cohort sample group to zero. When we fit the model, the genotypes, x, keep their observed values $\{0, 1, 2\}$ if we assumed that the underlying disease model is either the logistic or the additive. However when one of the recessive or dominant models is assumed the genotypes need to be adjusted in the manner we discussed in section 2.1.2 - also the same rationale applies to the scores we use in the CA test. So when we assume the recessive disease model we use $x^* = I_{[x=2]}$ instead of x and when we assume the dominant disease model we use $x^{**} = I_{[x>0]}$. It should also be noted that under

the CC design this test performs better when the underlying disease model is the logistic since either M_0 or M_1 is the model that actually generated the data. Nevertheless it can still detect any potential association even when the data were generated by any of the other three models.

3.1.2 The Bias – Variance trade-off

Using a different perspective we can think of the conditional minor allele frequencies of cases and controls as location parameters and essentially any genetic association test is used to determine whether these two parameters are equal. This is obvious for retrospective tests like the Welch's t-test, odds-ratio and relative risk tests but also under the assumption of HWE, it becomes clear for the Cochran-Armitage test as well. This is probably less obvious for prospective tests but especially in the case of the logistic regression when we consider the connection we discussed earlier in this thesis between its slope coefficient (β) estimate and the odds ratio statistic the location parameter interpretation becomes more clear. So essentially, directly or indirectly, the power of a test depends in the distance between the estimates, usually ML estimates, of the two conditional MAFs.

Considering all the above we examine what happens, in theory, to the power of a test when we change the design from the case-control to the case-cohort-control or in other words what happens when we add and merge the cohort and controls. We consider the maximum likelihood estimators $\hat{\theta}$, $\hat{\theta}_p$ $\hat{\theta}_q$ of θ, θ_p and θ_q respectively. Since we deal with proportions the MLE is an unbiased estimator, $E(\hat{\theta}_*) = \theta_*$, where θ_* denotes any of the three allele

frequencies. Additionally it can be easily proven (Appendix A) that

$$\begin{aligned} \theta_q \leq \theta \leq \theta_p \quad &, \text{ if the minor allele is causal} \\ \theta_p \leq \theta \leq \theta_q \quad &, \text{ if the minor allele is protective} \end{aligned} \tag{3.16}$$

and the equality holds only when there is no association. Also considering that the disease prevalence of most GWAS is significantly less than 50% it is safe to assume that θ is much closer to θ_q than θ_p. Hence merging cohort and controls and thus using the joint MLE $\hat{\theta}_{q,g}$ (3.7) instead of $\hat{\theta}_q$ results in shorter distance between the expected values of the two estimators. This alone would suggest that a test under the CC design is always more powerful than the same test under the CCC design. However we cannot ignore the variance of the ML estimator.

In general and assuming regularity conditions are satisfied, all ML estimators converge in distribution to a normal distribution,

$$\hat{\phi} \overset{d}{\sim} N\left(\phi, I^{-1}(\phi)\right) \tag{3.17}$$

where $\hat{\phi}$ is a generic MLE and $I(\phi)$ is the Fisher information. In the case of the binomial distribution as is the case with allele frequencies the variance calculated from the Fisher information is

$$I^{-1}(\theta_*) = \frac{\theta_*(1-\theta_*)}{2N_*}, \tag{3.18}$$

where N_* is the sample size; R,S and M for the cases, controls and cohort

respectively. So this asymptotic variance depends not only on the true parameter value but most importantly on the sample size. The larger the sample size the larger the precision of the asymptotic normal distribution. In general, considering the sources from which it can be obtained, we expect that the cohort sample size will be at least as large as the cases or controls size. Thus we expect the distribution of $\hat{\theta}$ to be tighter than the other two. Even if that is not true the asymptotic distribution of $\hat{\theta}_{q,g}$ - its variance is given in (3.7) - will always be tighter than the distribution of $\hat{\theta}_q$. After all when we adopt the CCC design we consider $\theta_{q,g}$ as the conditional MAF of the unaffected sub-population instead of θ_q.

So, overall, the power of the test does not depend only on the location parameter, which is the expected value of the conditional MAF estimate, but it depends on its variance as well. And since the asymptotic distribution of the ML estimator is normal, the expected value and variance fully determine the distribution of the estimate. Thus the power of a test depends on the overlap of the distributions $N(\theta_p, Var(\hat{\theta}_p))$ and $N(\theta_q, Var(\hat{\theta}_q))$ when we adopt the case-control design and the overlap of the distributions $N(\theta_p, Var(\hat{\theta}_p))$ and $N(\theta_{q,g}, Var(\hat{\theta}_{q,g}))$ when we adopt the case-cohort-control design. For a given sample the design with the smaller overlap will give more powerful tests (Figure 3.1).

Therefore when using the cohort as controls there is a loss in terms of bias and a gain in term of variance, a bias–variance trade-off. Whether the drawbacks of the former overwhelm the benefits of the latter depends on the variables - MAF, disease prevalence sample size, effect size - of the problem.

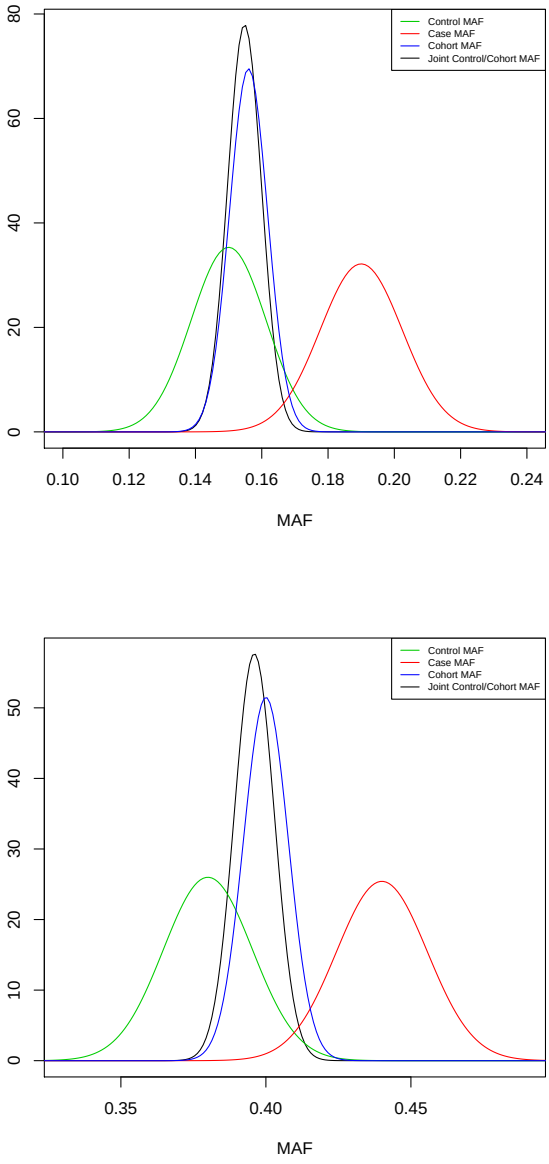


Figure 3.1: Graphs showing the normal approximation of conditional MAF estimates. The graph on the top shows a possible scenario where using the merged cohort/control sample (CCC design) will increase the power of a test. The graph on the bottom shows the opposite scenario where adopting the CCC design will result in power loss.

3.1.3 Simulation Results

We run simulations using a large number of different values and combinations of the parameters $(\theta, K, n, \lambda, \beta)$. We use the the logistic, additive, recessive and dominant disease models. The simulations are used to calculate and compare the power of the three tests we discuss in section 3.1.1. Four examples of the simulation based power-curves are shown in figures 3.2 - 3.5. It should be noted here that simulation results were not deemed reliable and were extremely noisy when we adopted the recessive disease model using small values of MAF together with small sample sizes. This was somewhat expected since the recessive model increases the risk of disease only to those homozygote individuals with two copies of the rare allele. For low MAF those individuals are very rare. Therefore this setting was exluded from the comparisons.

The first thing we do, is to compare the power of the three tests. Under the CC design all three appear equally powerful. However, under the CCC design, the Cochran-Armitage test is, in most cases, more powerful than the other two. In the cases when it is not more powerful, it is equally powerful as the other two tests. In summary, under the CCC design, neither the Welch's t nor the deviance test were ever more powerful than the CA test. As far as the other two tests are concerned, the deviance test appears to fare slightly better than the Welch's t-test in terms of power.

The second thing we do is to compare the two designs. Specifically we want to see whether the individual tests have more power under one design

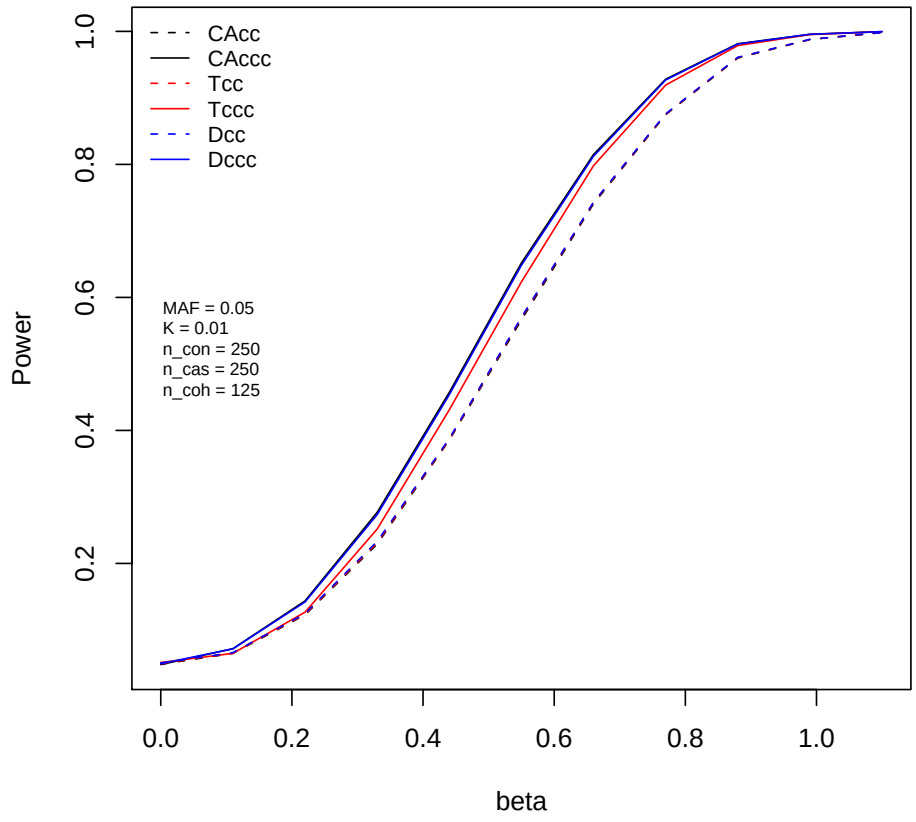


Figure 3.2: An example of the power-curves used for power comparisons. Here the data were simulated using the logistic disease model. In this plot all the tests under the CCC design are more powerful than the respective tests under the CC design. Here the CA and Deviance tests are tied as the most powerful CCC tests (their power-curves overlap). The Welch's t-test is slightly less powerful while the three tests under the CC design give the same power between them.

compared to their counterparts under the second design for all the parameter values. The results we obtain are in accordance with the bias-variance trade-off we discussed in section 3.1.2. Neither design results in more powerful tests for all parameter values. In the majority of the cases the three tests

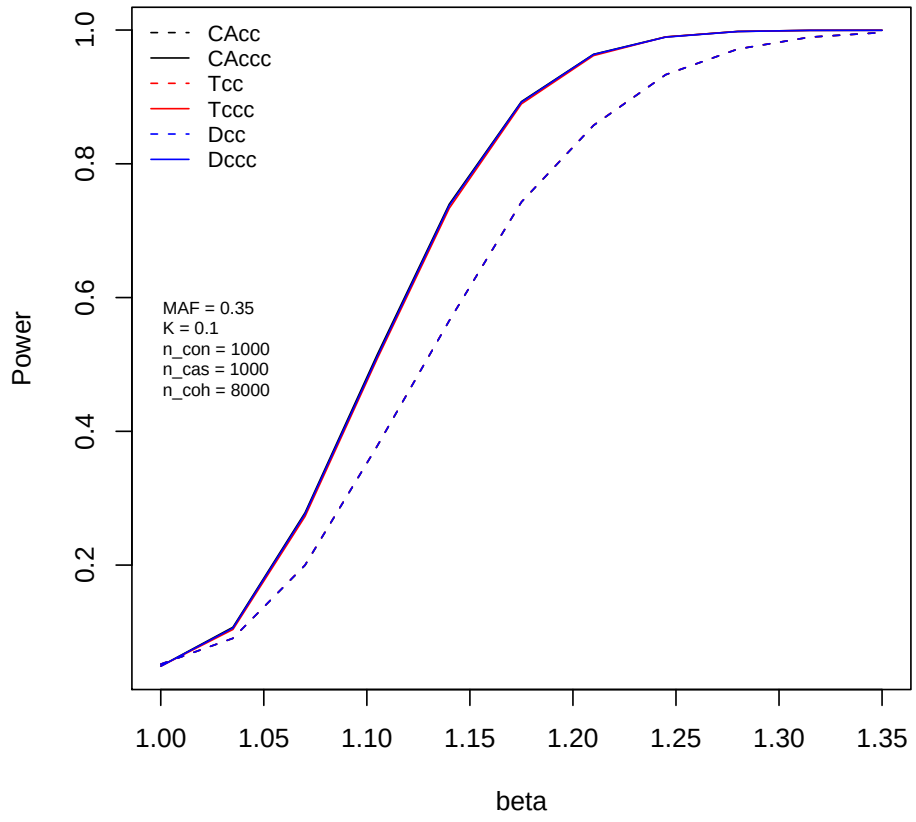


Figure 3.3: An example of the power-curves used for power comparisons. Here the data were simulated using the additive disease model. The three tests under the CCC design appear to be roughly equal in terms of power with the Deviance test being the one marginally less powerful. The tests under the CC design are also equally powerful with respect to each other but significantly less powerful than the CCC tests.

are more powerful under CCC design but there are many occasions where the CC tests outperform their CCC counterparts. Although the simulation results are not sufficient to provide rigid conclusions as to when each design should be used in order to increase the power of our tests, we can still get

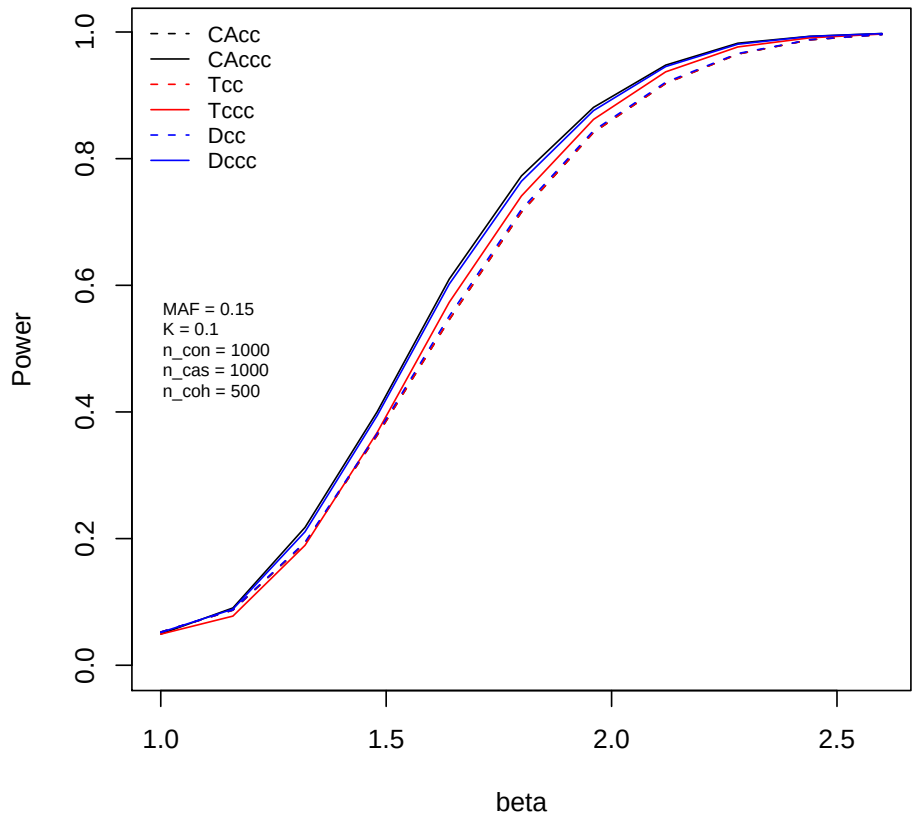


Figure 3.4: An example of the power-curves used for power comparisons. In this particular example the data were simulated using the recessive disease model. The CA test under the CCC design is the most powerful followed by the slightly less powerful Deviance test under the same design. The CC tests are give less power than any CCC test and the Welch's t-test is marginally the least powerful among them.

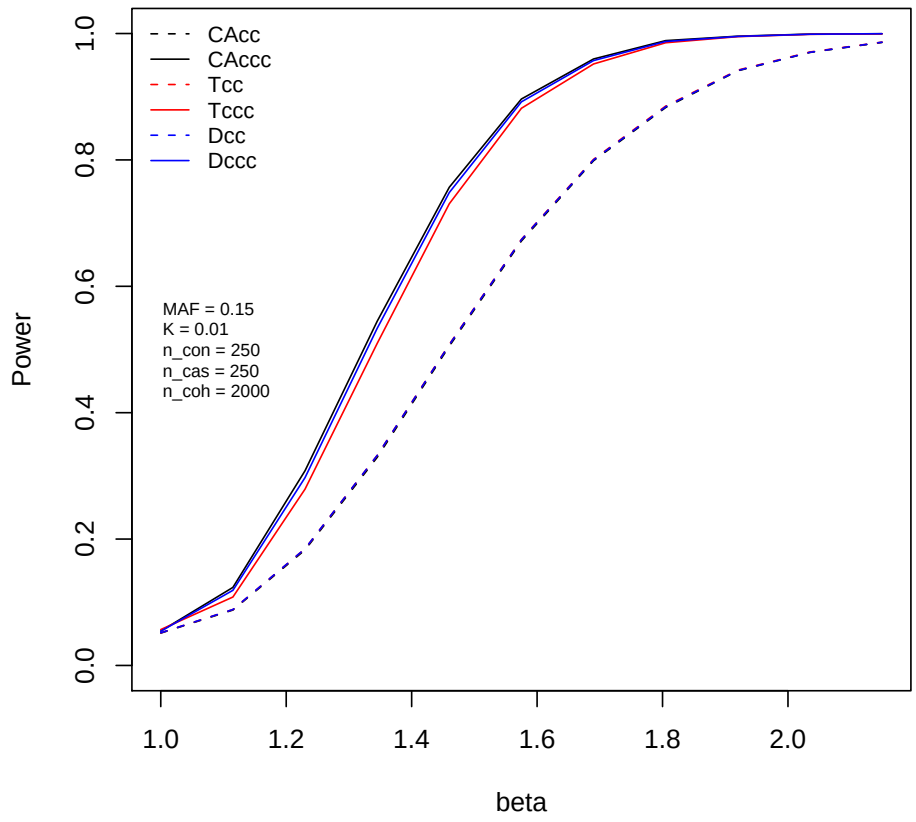


Figure 3.5: An example of the power-curves used for power comparisons. Here the data were simulated using the dominant disease model. The tests under the CCC design are significantly more powerful than their CC counterparts, with the most powerful being the CA test, followed by the Deviance test and then the Welch's t-test. All three CC tests give the same power.

a rough picture of the parameters' role on this decision (CC or CCC). In other words, we can draw some empirical conclusions based on the simultion results.

The parameter that appears to have the greatest effect on whether a test is more powerful under the CCC or CC design, is the disease prevalence K. For small to moderate values of K all tests under CCC design are always more powerful than the ones under the CC design. The opposite happens for large values of K. Those observations may appear vague but keep in mind that the power depends on all the parameters. For example, for a given value of K, a test may have more power under the, say, CCC design. Then it is possible, for the same K but different values of the other parameters, the same test to be, this time, more powerful under the CC design. The effect that K has on the design choice can be explained intuitively. Since we treat the cohort as controls, a large K will result in a large number of affected inviduals in the joint control-cohort sample. Therefore there will be not as much differrence between this sample and the cases as it would have been, had we used just the genuine controls.

The effect of the MAF, θ, is much more subtle and it is better observed in conjunction with the effect of K, specifically for relatively large values of K where there is a possibility for a test to be more powerful under either design. In general a low MAF favours the CCC design. For example, we can have a setting where for a given value of K, a low MAF will result in the tests being more powerful under the CCC design, while a high MAF, close to 0.5 will result in the tests having more power under the CC design.

We remind that for our simulations we assume equal sample size n for cases and controls, while the cohort sample size is determined by the parameter λ and it is equal to λn. The value of n does not appear to affect the decision between CC and CCC. Increasing n causes the tests under both designs to gain more power. The parameter λ, obviously does not affect the power of the tests under the CC design. It does not affect the decision between CC or CCC either. What it does affect is the relative power of a test under the two designs. If a test is already more powerful under the CCC design, increasing λ will cause its power to increase too. Since the power under the CC design is not affected the relative power, $\pi(\text{Test}_{CCC})/\pi(\text{Test}_{CC})$, increases. The opposite happens if a test is more powerful under the CC design.

We should also note that all the empirical conclusions we describe in this section hold regardless of the choice of disease model. Since results that are obtained from data generated from different models are not comparable, we cannot comment on whether the effect of a parameter is more significant for a certain model.

3.1.4 A Hybrid Test Statistic

Having seen that neither design guarantees most powerful tests, we propose a hybrid test statistic that renders the need to *à priori* choose which design we should use, redundant; or even when we decide to use both designs, it can help us avoid the ambiguity over which results we should trust. For each test statistic we use, its asymptotic distribution under the null hypothesis is the

same whether we adopt the CCC or CC design. Therefore the design that results in more powerful tests, is the one where the test statistic has a more extreme value. So, a natural choice is to calculate the test statisics under both designs and then trust the most extreme. Let $CA_{,CC}$, $\mathcal{T}_{CC}$ and $\mathcal{D}_{CC}$ denote the three test statistics under the CC design and $CA_{,CCC}$, $\mathcal{T}_{CCC}$ and $\mathcal{D}_{CCC}$ the three test statistics under the CCC design. Then we define the hybrid test statistics:

$$\begin{aligned} CA_{\max} &= \max\{CA_{CC}, CA_{CCC}\} \\ \mathcal{T}_{\max} &= \operatorname{sgn}(\mathcal{T}) \max\{|\mathcal{T}_{CC}|, |\mathcal{T}_{CCC}|\} \\ \mathcal{D}_{\max} &= \max\{\mathcal{D}_{CC}, \mathcal{D}_{CCC}\} \end{aligned} \tag{3.19}$$

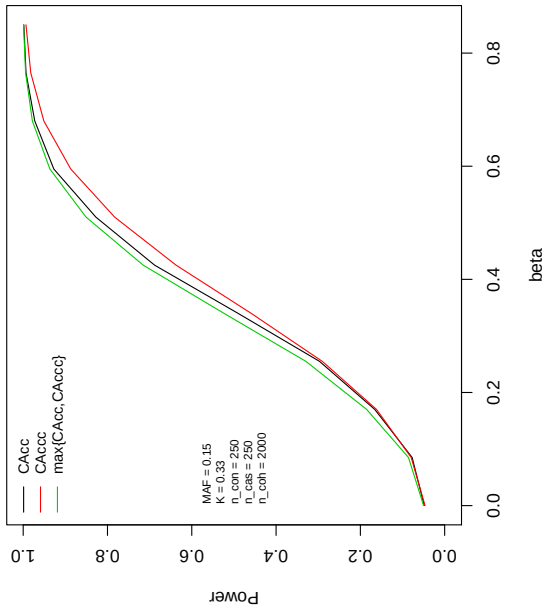
Two issues arise when we decide to test for association using those hyrid statistics. First, is the question whether they are really more powerful than the respective CC and CCC test. The second and most important issue concerns their distribution under the null hypothesis. Obviously, it is not χ^2_1 any more.

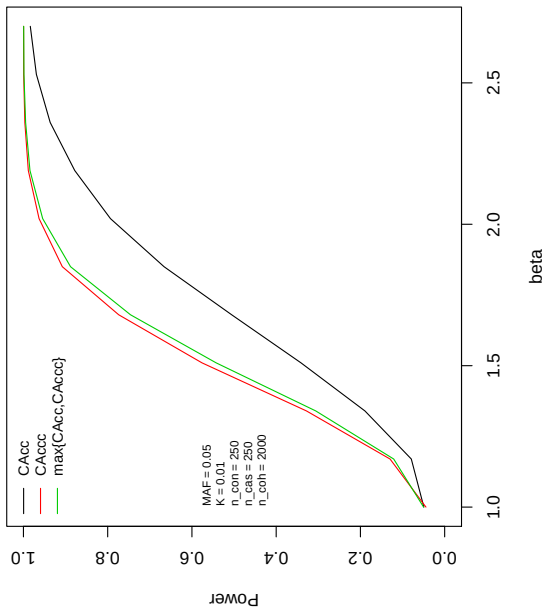
We study the first issue by doing another simulation-based power-analysis. However, in order to do that we still have to deal with the second issue; we need the distribution or more precisely the quantiles of the distribution of the hybrid test statistics, under the null hypothesis. In our study we deal with this problem by doing additional Monte-Carlo simulations. We generate a large sample of realisations of those hybrid test statistics under H_0. Thus based on this Monte-Carlo sample we can obtain the empirical quantiles

we need in our power calculations. The simulation results show that the test based on the hybrid test statistic is not always as powerful as the most powerful of the CC or CCC tests. In some cases the hybrid test statistics give, indeed, more powerful results than the other two but in most cases they are slightly less powerful than the most powerful test among the other two. However the loss in power is really small (Figure 3.6).

The difference between using the hybrid test statistic on one hand and simply trusting the test statistic with the most significant value on the other, is that we discard additional information, thus inserting bias in our decision to accept or reject the hypothesis of association. This issue manifests in the distribution of the hybrid test statistics under H_0. If we simply choose the most significant test statistic, having seen both of them, then the distribution under H_0 is no longer χ^2_1. The added bias in this case favours the alternative hypothesis. Therefore when we apply a test using both designs, we either have to report both results separately or if we want an joint result we have to change the distibution under H_0.

So, now, our problem is to find the distribution of the hybrid test statistics under the null hypothesis. In our power analysis we dealt with this problem by using Monte-Carlo methods. Unfortunately, in practice, this is not possible since the parameters we used to generate the Monte-Carlo samples are unknown. The one scenario where we can still use Monte-Carlo methods to approximate that distribution, is when it does not depend on the parameters we use in the simulations or at least when the distribution is not as sensitive in small changes of those parameters, something that will





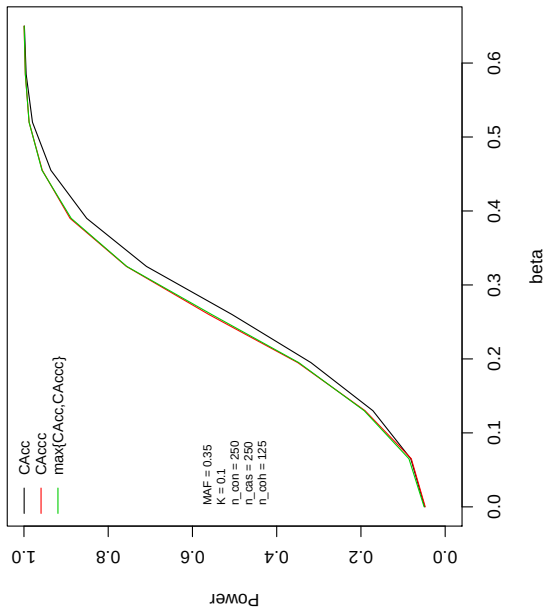


Figure 3.6: The power of the test based on the CA_{max} test statistic, compared with the power of CA test under CC and CCC designs. Top left: CA_{max} is slightly less powerful than CA_{CCC}; Top right: CA_{max} is the most powerful; Bottom: CA_{max} is equally as powerful as CA_{CCC} while both are more powerful than CA_{CC}.

allow us to use their sample estimates. We study the effect that the individual parameters have on on the distribution under H_0, by generating samples from this distribution using various parameter values. The results show that, other than the expected Monte-Carlo error, the distributions of the hybrid test statistics under the null hypothesis do not depend on the MAF or the sample size (Figure 3.7). Therefore in an real life association study there is no need to generate samples from the distribution under the null for each marker we study. Since we assume that the disease prevalence is known, as are the sample sizes, we can obtain the empirical distribution using a fixed value of MAF and use this distribution to calculate the empirical p-values of the hybrid tests for all markers.

In general, as a rule of thumb, we suggest that, when we deal with a relatively rare disease, the CCC is preferable, when there are available data. If we deal with a very common disease or trait then the hybrid test should be used instead.

3.2 Asymptotic Power Analysis of the Cochran-Armitage Test

It is evident from our simulation-based power analysis that the CA test is more powerful than the other two tests we used. For this reason we proceed to a more detailed power analyses of this test. Specifically, we study how the values of the parameters, θ, K, n and λ, affect the power of the CA test under the two designs, namely, whether CA_{CC} or CA_{CCC} is more powerful.

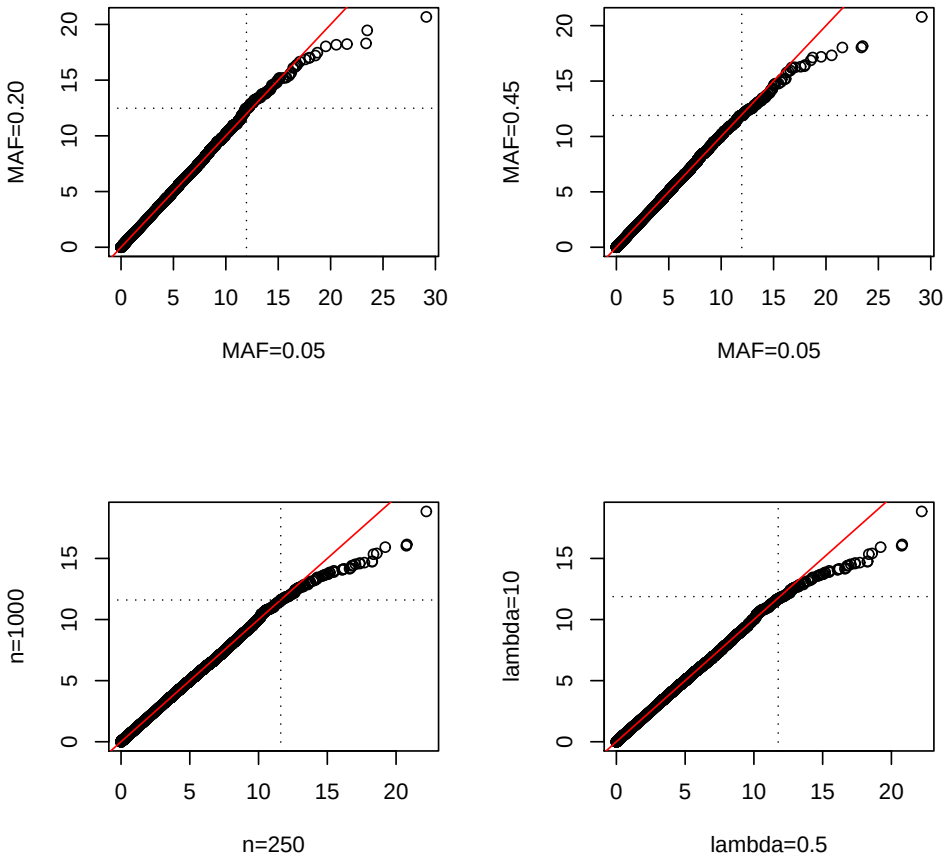


Figure 3.7: Comparison of the empirical distribuion of CA_{max} obtained from data generated under the null hypothesis. In the two plots on the top we use different values of the MAF on each axis. In the bottom left plot we use different sample size on each axis and in the bottom right plot different values of λ. The red line has gradient 1 and passes through the origin. Dotted lines show the 0.99 sample quantiles of each distribution.

We make use of an asymptotic formula for the power of the CA test. This helps us avoid time consuming simulations and thus being able to use a much larger array of parameter values.

3.2.1 Asymptotic Power and a Phenotype Misclassification Problem

The formula for the asymptotic power of the CA test statistic (Slager and Schaid, 2001) is given by:

$$\text{Power} \approx 1 + \Phi\left(\frac{-\zeta_{1-\alpha/2}\sqrt{\tilde{\sigma}_a^2 + \mu_a^2} - N\mu_a}{\sigma_a}\right) - \Phi\left(\frac{\zeta_{1-\alpha/2}\sqrt{\tilde{\sigma}_a^2 + \mu_a^2} - N\mu_a}{\sigma_a}\right), \quad (3.20)$$

where α is the significance level (Type I error), Φ is the cumulative distribution function and $\zeta_{1-\alpha/2}$ is the upper $1-\alpha/2$ percentile of the standard normal distribution. Furthermore,

$$\begin{aligned}
\mu_a &= \frac{1}{N^2}\sum_x z_x(p_x - q_x) \quad (3.21)\\
\sigma_a^2 &= \frac{1}{N^3}RS^2\left\{\sum_x z_x^2 p_x - \left(\sum_x z_x p_x\right)^2\right\} + \frac{1}{N^3}R^2S\left\{\sum_x z_x^2 q_x - \left(\sum_x z_x q_x\right)^2\right\}\\
\tilde{\sigma}_a^2 &= \frac{1}{N^3}R^2S\left\{\sum_x z_x^2 p_x - \left(\sum_x z_x p_x\right)^2\right\} + \frac{1}{N^3}RS^2\left\{\sum_x z_x^2 q_x - \left(\sum_x z_x q_x\right)^2\right\}
\end{aligned}$$

The notation used in (3.20) and (3.21) is in accordance to the one introduced in equations (2.3) and (3.3) and the table 3.1.

The asymptotic formula was derived and it is meant to be used under the premise of a standard case-control design. Therefore we need to make some appropriate adjustements in order to be able to use it to calculate the power

of the CA_{CCC} test. To do this, we contemplate an alternative interpretation of our testing process under the CCC design. We can cast aside the whole notion of the design and instead consider it as a case-control test on data, where there may be errors in some of the controls' phenotypes.

Zheng and Tian (2005) use the same asymptotic formula while Edwards et al. (2005) use a similar asymptotic formula for the more general Pearson chi-square test, for power analysis and sample size determination while at the same time taking into consideration potential phenotype misclassification or alternatively diagnostic error. Additional issues arise when genotype coding errors are also assumed (Gordon et al., 2000). The authors use the concepts of sensitivity and specificity in order to revise the frequencies p_x and q_x and then proceed to use the revised frequencies p_x^* and q_x^* in the formula for the asymptotic power. Sensitivity is the probability of correct diagnosis for an affected individual and specificity is the probability of correct diagnosis for an unaffected individual. More formally, if $D \in \{0, 1\}$ denotes the assigned phenotype based on a diagnosis,

$$Se = P(D = 1|Y = 1), \qquad Sp = P(D = 0|Y = 0) \tag{3.22}$$

With these in consideration, we revise the distribution of the genotypes we present in Table 3.2 and reshape the table, so that now it corresponds to a case-control design with phenotype misclassification errors. In Table 3.3 m_x^0 denote the unaffected individuals in the cohort while m_x^1, the affected ones.

Since the cohort is a random sample from the general population, the ex-

		Genotype			
		0	1	2	Total
Case		r_0	r_1	r_2	R
Control	True Control	$s_0 + m_0^0$	$s_1 + m_1^0$	$s_2 + m_2^0$	$S + M^0$
	False Control	m_0^1	m_1^1	m_2^1	M^1
Total		n_0	n_1	n_2	N

Table 3.3: Revised distribution of the genotypes under the case-cohort-control design. The format, now, is that of a case-control design with phenotype misclassification errors for the cohort samples.

pected proportion of affected individuals in the sample is equal to the disease prevalence, K. Consequently the expected number of affected individuals is $K \cdot M$. In our data the specificity is always 1, since we do not assume that any true cases were misclassified. Therefore there is no need to revise the genoype frequencies, p_x of the cases. The sensitivity on the other hand can be expressed as the ratio of the identified cases to the expected number of total cases in our sample, i.e. $Se = R/(KM + R)$. However, when we revise the genotype frequencies of the controls we also have to take into account that only a subset of the sample use as controls, namely the individuals from the cohort, is susceptible to misclassification. As we already mentioned the sensitivity of the cohort is K. Therefore the expected frequency of the genotypes in the cohort is $Kp_x + (1 - K)q_x$. The revised control frequency q_x^*, depends on the sample sizes and can be expressed as a weighted average of the frequencies from the constituent samples that form the joint control group.

$$q_x^* = \frac{(Kp_x + (1-K)q_x)M + q_x S}{M + S} = \frac{KMp_x + ((1-K)M + S)q_x}{M + S} \quad (3.23)$$

Overall, in order to be able to use the asymptotic formula for power calculations under the CCC design we have to substitute, in equations (3.20) and (3.21), q_x with q_x^* and also change the controls sample size from S to $S + M$.

3.2.1.1 Accuracy of the Asymptotic Power

In order to assess the accuracy of formula for the asymptotic power, we compare it with the simulation-based power. We apply the formula using the same values of the parameters we choose to generate the data for the power simulations. Our main concern is the accuracy of the formula under the CCC design, where we made the adjustment regarding the genotype frequencies of the controls. The results show that there is no difference in power, whether it is calculated based on simulation or using the asymptotic formula. In fact the power-curves we draw for comparisons, completely overlap each other regardless of the parameters we choose (Figure 3.8).

3.2.2 Results

As we already stated our main interest is to study the effect that the four parameters (θ, K, n, λ), have on the power of the Cochran-Armitage tests under both the CC and CCC design. In order to do that, initially we study each parameter individually. That is, we keep the other three parameters fixed and calculate the power for a sequence of values of the parameter we study. Once we study the individual parameter effects we do some cross-section comparisons of the initial results in order to study potential joint

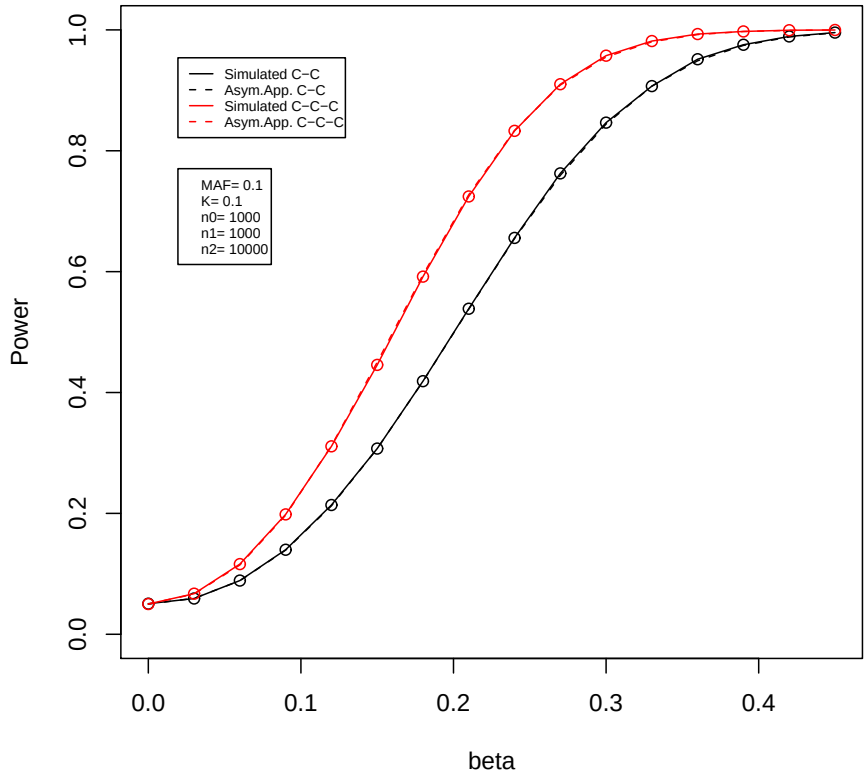


Figure 3.8: Comparison of asymptotic and simulations-based power of the CA test, under the CC and CCC design.

parameter effects.

Let $\pi(\beta|\theta, K, n, \lambda; CA_{CC})$ denote the power of the CA test under the CC design and similarly $\pi(\beta|\theta, K, n, \lambda; CA_{CCC})$ the power of the CA test under the CCC design. The power is calculated on a sequence of values of the effect size parameter β and it is conditional on the choice of the other parameter values. When we focus on the effect of a certain parameter, say for example θ, that means we will compare the power for various values of

θ in the conditional part of the power. The conventional notation we use in this cases is $\pi(\beta|\theta^{(i)}, \cdot)$, where the dot indicates that all the other parameters and the design are the same. In general, if a parameter is not included in the conditional part, that means that the observation does not concern the effect of that parameter.

Also we define the function γ; the relative change in power,

$$\gamma(\beta|\theta_g, K, n, \lambda) = \frac{\pi(\beta|\theta, K, n, \lambda; CA_{CCC}) - \pi(\beta|\theta, K, n, \lambda; CA_{CC})}{\pi(\beta|\theta, K, n, \lambda; CA_{CC})}. \quad (3.24)$$

This indicates the relative increase or decrease in power when we adopt the CCC design instead of the CC design. The function γ plays an import role in our power comparisons and we use the same conventional notation as the one we use for the power.

3.2.2.1 Empirical Observations on Individual Parameter Effect

Here we list our empirical conclusions on how the individual parameters affect the behaviour of the power as a function of β (Figure 3.9 [1]). In our study we include all four disease models; the logistic, additive, recessive and dominant. Sometimes the effect of a parameter depend on the choice of disease model. Some irregularities occur when we use the dominant disease model but we explain this in more detail in a subsequent section.

Minor Allele Frequency: As the value of θ increases, the power increases at a higher rate as a function of β. More formally, if $\theta^{(1)} < \theta^{(2)}$, then

[1] The complete set of figures is available in Appendix B

$\pi(\beta|\theta^{(1)},\cdot) \leq \pi(\beta|\theta^{(2)},\cdot)\ \forall\beta$. The equality holds only asymptotically as $\beta \to \infty$ and thus $\pi(\beta|\cdot) \to 1$. This is true regardless of the choice of design.

Disease Prevalence: When the logistic disease model is used to generate the data, as the value of K increases, the power increases at a lower rate as a function of β that is if $K^{(1)} < K^{(2)}$ then $\pi(\beta|K^{(1)},\cdot) \geq \pi(\beta|K^{(2)},\cdot)$. Changing the value of K, mostly affects the power of CA_{CCC} while the effect on CA_{CC} is minimal. On the other hand, when we use one of the other three models to generate the data the effect of K changes. Now, as K increases, the power also increases at a higher rate as a function of β. So, in this instance, if $K^{(1)} < K^{(2)}$ then $\pi(\beta|K^{(1)},\cdot) \leq \pi(\beta|K^{(2)},\cdot)$. Additionally now it is the power of CA_{CC} that is mostly affected by the change in the value of K. In any case, increasing K causes the difference $\pi(\beta|\cdot;CA_{CCC}) - \pi(\beta|\cdot;CA_{CC})$ to decrease.

Cases/controls sample size: When we increase sample size n the rate that power increases as a function of β increases as well. The effect of n on the behaviour of the power is actually similar to that of θ. In general, if $n^{(1)} < n^{(2)}$ then $\pi(\beta|n^{(1)},\cdot) \leq \pi(\beta|n^{(2)},\cdot)$ and this holds under both designs.

Cohort sample size: The parameter λ, obviously, does not affect the power of CA_{CC} in any way. Even so, its effect on the power of CA_{CCC} is much more complex than that of the other parameters. In some occasions the power of CA_{CCC} increases at a higher rate as we increase λ while others, it increases at a lower rate as λ increases. The effects of λ is

better studied in conjunction with the other parameters, mainly the disease prevalence K.

3.2.2.2 The Maximum Rare Disease Prevalence

We saw that the parameter that has the largest effect on whether CA_{CC} or CA_{CCC} is more powerful is the disease prevalence, K. For this reason we study the effect of this parameter in more detail. We saw that for a rare disease CA_{CCC} is more powerful than CA_{CC} while the opposite generally holds when the disease is very common. We introduce the notion of the Maximum Rare Disease Prevalence (K_{MR}). This is the maximum value of disease prevalence where the power of CA_{CCC} is greater the power of CA_{CC} for all values of β. More formally:

Definition 1. *For any given values of the parameters (θ, n, λ), there exists a value of $K_{MR} = K_{MR}(\theta, n, \lambda)$ where, $K_{MR} = \max\{K : K \in (0,1)\}$ such that for any $K \leq K_{MR}$:*

$$\begin{aligned} \pi(\beta|K,\cdot;CA_{CC}) &< \pi(\beta|K,\cdot;CA_{CCC}) \quad \forall\beta \quad \textit{and} \\ \pi(\beta|K,\cdot;CA_{CC}) &\to \pi(\beta|K,\cdot;CA_{CCC}) \to 1 \quad \textit{only as} \quad \beta \to \infty \end{aligned}$$

Also, for any $K > K_{MR}$, $\exists \quad \beta^* < \infty$, such that:

$$1 > \pi(\beta^*|K,\cdot;CA_{CC}) \geq \pi(\beta^*|K,\cdot;CA_{CCC})$$

We call K_{MR} the ***Maximum Rare Disease Prevalence***

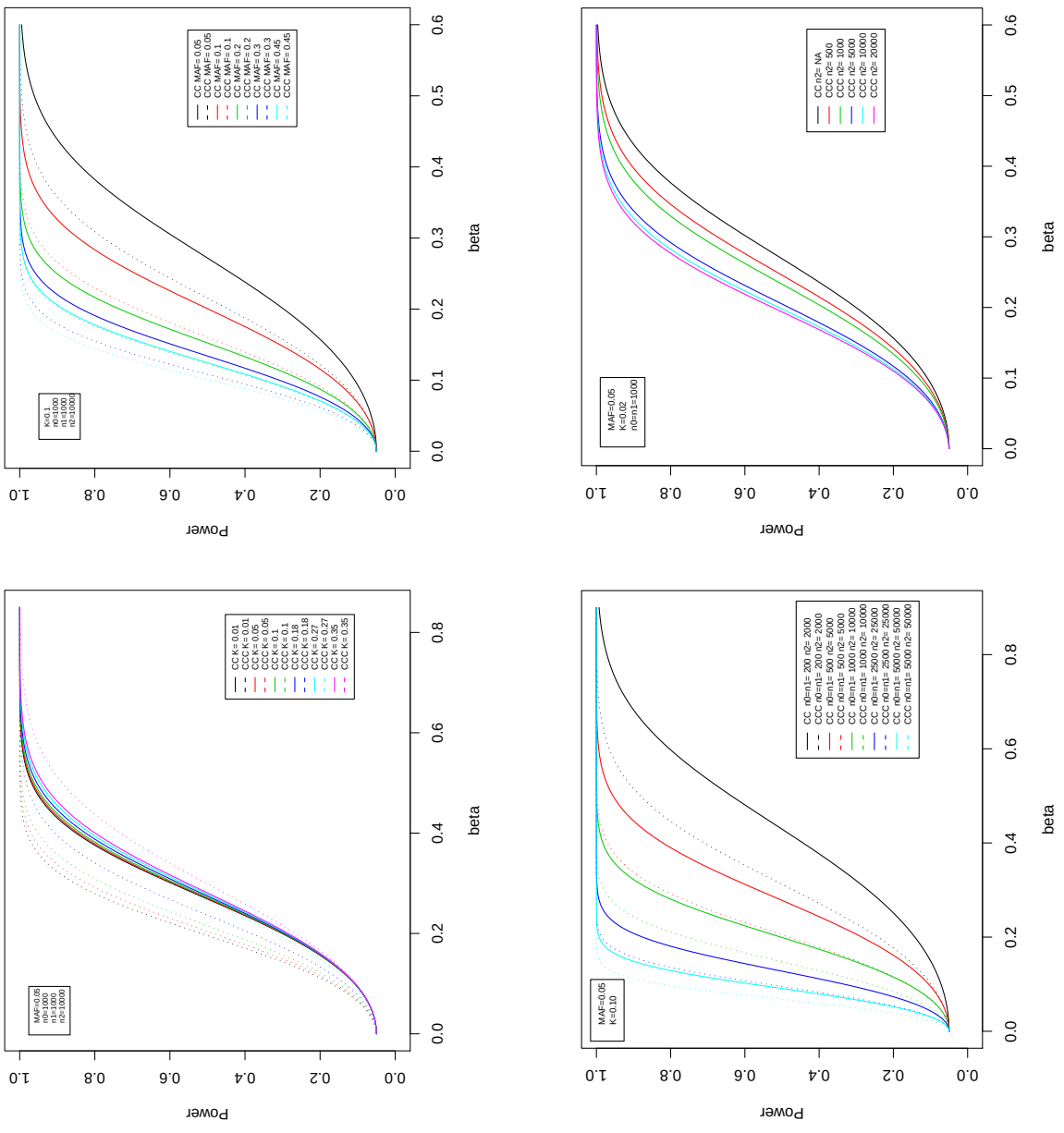


Figure 3.9: Examples of the power-curves we used in our study. We use different values of K (top left), θ (top right), n (bottom left) and λ (bottom right)

Our inference here, is based less on the power-curves and more on the plots of the relative increase in power, $\gamma(\beta|\cdot)$. Now we are more interested on how the other three parameters affect the value of K_{MR}. Before we comment on the effect of θ, n and λ we need to add some more general remarks.

- By definition, if $K < K_{MR}$ the power of CA_{CCC} is always greater than the power of CA_{CC}. In this case $\gamma(\beta|K, \cdot)$ is always positive and it is a concave function of β. That is, it increases, reaches a maximum and then decreases asymptotically to 0^+ (Figure 3.10[2]).

- When $K > K_{MR}$ the power of CA_{CC} is not always greater than the power of CA_{CCC}, that is, not for all the values of β. For small values of β, CA_{CCC} is still more powerful. Then for some $\beta^* = \beta^*(K_{MR})$, the two tests have equal power. For $\beta > \beta^*$, CA_{CC} is always more powerful. In other words, two power-curves intersect only once. In this case, $\gamma(\beta|K, \cdot)$ resemples a cubic function with initially a local maximum, which is positive and then a local minimum, which is negative. After that the function, asymptotically increases to 0^-.

When we study the effect of θ, n and λ on the value of K_{MR}, we consider the latter as a function of the three parameters ($K_{MR}(\theta, n, \lambda)$). Similar to our process when we were studying the effect of the individuals parameters on the power, here as well, we keep two of the parameters fixed and see how K_{MR} changes as we change the value of the third parameter.

MAF: $K_{MR}(\theta, \cdot)$ is an increasing function of θ. That is, if $\theta^{(1)} < \theta^{(2)}$ then $K_{MR}(\theta^{(1)}, \cdot) < K_{MR}(\theta^{(2)}, \cdot)$.

[2]The complete set of figures is available in Appendix B

Furthermore as we increase θ, the absolute value of $\gamma(\beta|\theta,\cdot)$ increases as well. In other words if we have a gain in power by using the cohort, a larger MAF will result in larger gain. On the other hand, if we lose power by using the cohort, the loss will be greater if the MAF is large. So, if $\theta^{(1)} < \theta^{(2)}$ and $K < K_{MR}(\theta^{(1)},\cdot)$, then $0 < \gamma(\beta|\theta^{(1)},\cdot) < \gamma(\beta|\theta^{(2)},\cdot)$. If $K > K_{MR}(\theta^{(2)},\cdot)$ then $0 > \gamma(\beta|\theta^{(1)},\cdot) > \gamma(\beta|\theta^{(2)},\cdot)$ for the values of β where CAS_{CC} is more powerful than CA_{CCC} while it is still $0 < \gamma(\beta|\theta^{(1)},\cdot) < \gamma(\beta|\theta^{(2)},\cdot)$ for those values of β where CA_{CCC} is more powerful.

Sample sizes Similar conclusions hold for n as well as λ. If $n^{(1)} < n^{(2)}$ then $K_{MR}(n^{(1)},\cdot) < K_{MR}(n^{(2)},\cdot)$. If $K < K_{MR}(n^{(1)},\cdot)$ then $0 < \gamma(\beta|n^{(1)},\cdot) < \gamma(\beta|n^{(2)},\cdot)$ and if $K > K_{MR}(n^{(2)},\cdot)$ then $0 > \gamma(\beta|n^{(1)},\cdot) > \gamma(\beta|n^{(2)},\cdot)$.

If $\lambda^{(1)} < \lambda^{(2)}$ then $K_{MR}(\lambda^{(1)},\cdot) < K_{MR}(\lambda^{(2)},\cdot)$. If $K < K_{MR}(\lambda^{(1)},\cdot)$ then $0 < \gamma(\beta|\lambda^{(1)},\cdot) < \gamma(\beta|\lambda^{(2)},\cdot)$ and if $K > K_{MR}(\lambda^{(2)},\cdot)$ then $0 > \gamma(\beta|\lambda^{(1)},\cdot) > \gamma(\beta|\lambda^{(2)},\cdot)$.

It is worth noting here, that during our study K_{MR} was never that small so that the disease may be considered rare. In the majority of the cases K_{MR} was between 0.3 and 0.4. Thus if we deal with a rare disease and we are able to use an additional sample from the population, the CCC design will certainly result in increased power.

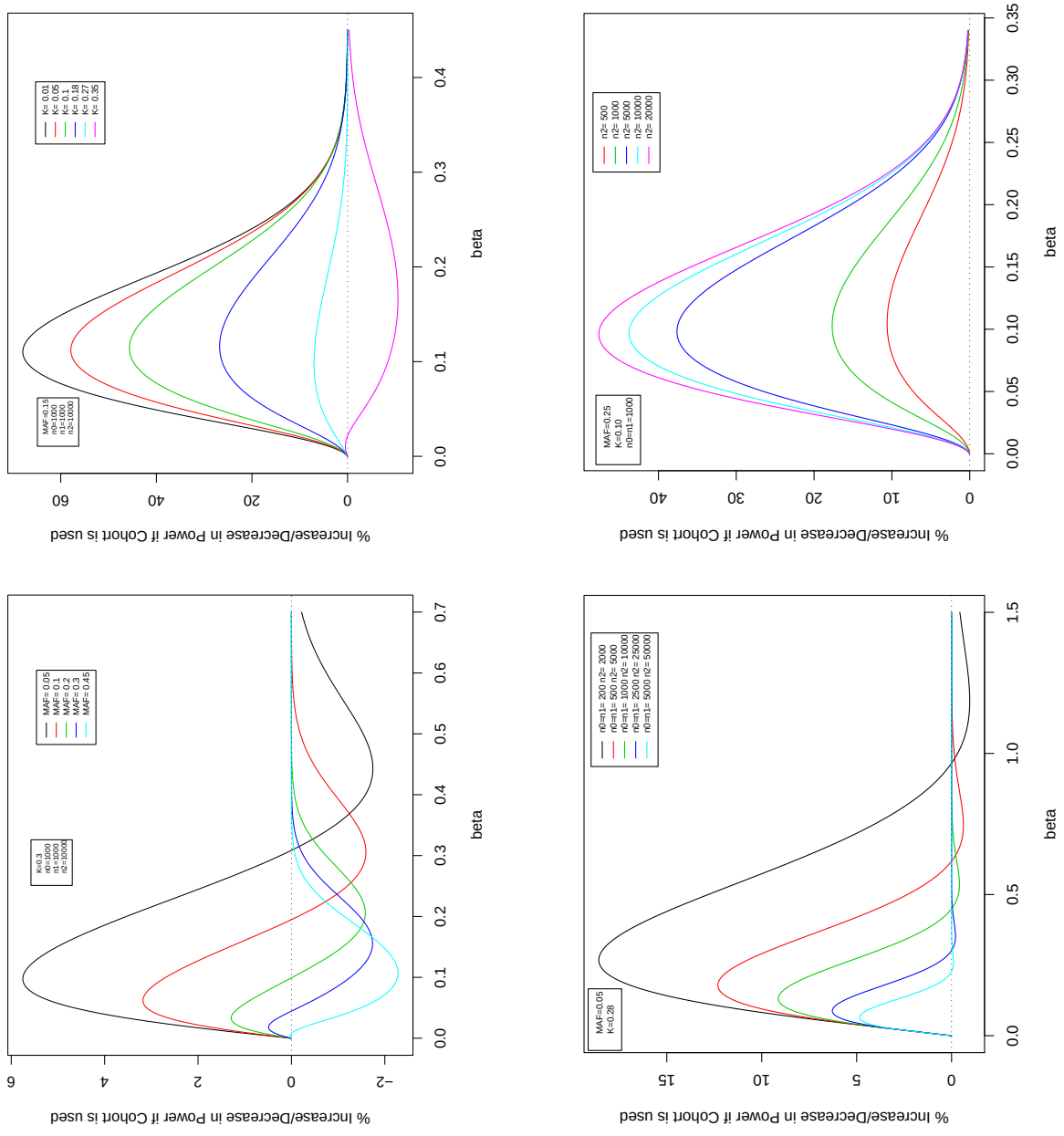


Figure 3.10: Plots of the relative increase in power, $\gamma(\beta|\cdot)$. We use different values of θ (top left), K (top right), n (bottom left) and λ (bottom right)

3.2.2.3 Some notes on the Dominant Disease Model

When it comes to the dominant disease model, there are some departures from our empirical observations. This is mostly true when it comes to the study of the MAF. For large values of θ the inequalities we present are not true any more and they are actually reversed. To understand this we go back to the derivation of the Cochran-Armitage test statistic. The CA test statistic can be written as $U^2/\text{Var}(U)$, where,

$$U \propto \sum_{x=0}^{2} z_x(p_x - q_x) \tag{3.25}$$

$$\text{Var}(U) \propto \left(\sum_{x=0}^{2} z_x p_x(1 - \sum_{x=0}^{2} z_x p_x) + \sum_{x=0}^{2} z_x q_x(1 - \sum_{x=0}^{2} z_x q_x)\right) \tag{3.26}$$

Now, considering the scores $\boldsymbol{z}$ we use for the different disease models express (3.25) and (3.26) as,

$$U \propto (\phi_p^{DM} - \phi_q^{DM}) \tag{3.27}$$

$$\text{Var}(U) \propto \left(\phi_p^{DM}(1 - \phi_p^{DM}) + \phi_q^{DM}(1 - \phi_q^{DM})\right), \tag{3.28}$$

where DM stands for the disease model. Specifically we have that for the logistic and additive disease models $\phi_p^{L-A} = 0.5p_1 + p_2 := \theta_p$ and $\phi_q^{L-A} = 0.5q_1 + q_2 : \theta_q$ (The scores we use, are $\boldsymbol{z} = (0, 1, 2)$ but since for those disease models we consider haplotypes the sample size doubles. For this reason we do not use $\phi_p = p_1 + 2p_2$ but we divide by 2). For the recessive disease model $\phi_p^R = p_2$ and $\phi_q^R = q_2$ while for the dominant we have, $\phi_p^D = p_1 + p_2$ and $\phi_q^D = q_1 + q_2$. Of course, when we apply the test, estimates, $\hat{\phi}^{DM}$, based on the data counts are used.

In our study we always assume that $0.01 \leq \theta \leq 0.5$. It is interesting to see how the components of the variance of U change as the value of θ changes.

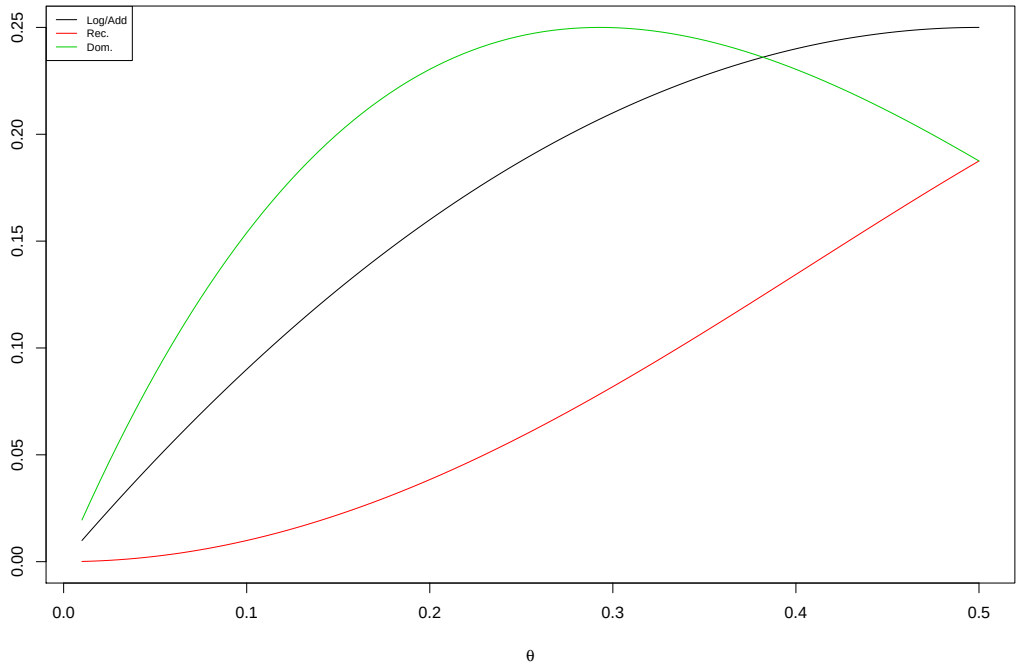


Figure 3.11: Demonstration of how the components of Var(U) change for different values of θ. This example concern the unconditional (population) genotype frequencies. Black line: $\phi^{L-A}(1-\phi^{L-A})$; Red line: $\phi^{R}(1-\phi^{R})$; Green line: $\phi^{D}(1-\phi^{D})$

Figure 3.11 demonstrates that for the unconditional genotype frequencies (i.e. g_x). It is clear that when we assume the logistic, additive or recessive disease models the variance of $\hat{\phi}^{DM}$ is a strictly increasing function of θ. This is not the case when we assume the dominant disease model, where the variance is a concave function of θ. Thus when we assume any of the three models, other than the dominant, the precision of the estimate $\hat{\phi}^{DM}$ decreases as θ increases. When the dominant model is assumed the precision

initially decreases, hits a minimum and then starts increasing again. These facts cause the irregularities we observe in our analyses under the dominant disease model.

We should also note that, when we assume that the rare allele is causal, like we assumed in our study, since $\theta_p > \theta$ the precision of $\hat{\phi}_p^{L-A}$ may also start to increase, even for $0.01 \leq \theta \leq 0.5$. However the impact in this case is minimal and almost imperceptible, since it happens for values of θ close to 0.5 and the variance in that area is almost constant.

Of course all the above depends on whether we decide to code the alleles in our analyses. As we mentioned, it is common to use 0 for the common allele and 1 for the rare allele. Had we swapped the allele coding, all the issues would have been observed in the recessive model. After all a dominant model where the rare allele is causal can be equivalently considered as a recessive model where the common allele is protective.

3.3 Potential Population Heterogeneity Between the Samples.

An issue that we have not addressed in our previous power analyses, is the possibility of population heterogeneity. Specifically, our main concern is whether the cohort sample will be compatible with the case/control sample. Since it is very likely the source that we obtain the cohort is different than the source we obtain the case/control sample, there is a danger that

those two samples will not come from the same exact population. This will lead to false discoveries when we test for association under the CCC design, since identified associations may be due to population heterogeneity and not due to actual association of the disease with the genotype.

In order to address this issue we develop a two-stage Likelihood Ratio Test (LRT). The idea is simple. In the first stage we test for population homogeneity. If we reject the hypothesis of population homogeneity in the first stage, then in the second stage we test for genetic association using only the cases and controls; therefore discarding the cohort. If the hypothesis of population homogeneity is not rejected, then in the second stage we test for genetic association using all three samples. One key aspect of our method, is the threshold, c, we use in the first stage. In order to calibrate our method we use data that have already been analysed as well as some of the results of those analyses.

3.3.1 Methods

The first thing we should note is that we assume that the disease prevalence, K, is known or we have a fairly accurate estimate of it. The key concept in our method, is that the population MAF, $\theta.$ can be expressed as a linear combination of the conditional MAFs of the affected and unaffected sub-populations, θ_p and θ_q respectively (Appendix A).

$$\theta = (1 - K)\theta_q + K\theta_p \tag{3.29}$$

Therefore if the cohort and cases/control samples were drawn from the same general population (homogeneous), we expect the (3.29) to hold for their respective allele frequencies. That is the first stage of our method.

Stage 1: Test for Homogeneity

$$H_0 : \theta = K\theta_p + (1-K)\theta_q \qquad H_1 : \theta \neq K\theta_p + (1-K)\theta_q \tag{3.30}$$

Consider that we have case-cohort-control data $\mathcal{G}$. For our data we follow the notation of Table 3.2. Then the likelihood of $(\theta_p, \theta_q, \theta)$ is given by:

$$L(\theta_p, \theta_q, \theta|\mathcal{G}) \propto \theta_p^{2r_2+r_1}(1-\theta_p)^{2r_0+r_1}\theta_q^{2s_2+s_1}(1-\theta_q)^{2s_0+s_1} \\ \theta^{2m_2+m_1}(1-\theta)^{2m_0+m_1} \tag{3.31}$$

In order to test the null hypothesis of homomogeneity we use a likelihood ratio (LR) test statistic.

$$LR_1 = \frac{\max L_{H_1}(\theta_p, \theta_q, \theta|\mathcal{G})}{\max L_{H_0}(\theta_p, \theta_q, \theta|\mathcal{G})} = \frac{\max_{\{\theta_p,\theta_q,\theta\}} L(\theta_p, \theta_q, \theta|\mathcal{G})}{\max_{\{\theta_p,\theta_q,\theta:\theta=K\theta_p+(1-K)\theta_q\}} L(\theta_p, \theta_q, \theta|\mathcal{G})} \tag{3.32}$$

The likelihood under H_1 is maximised for the usual unconstrained MLEs $\hat{\theta}_p = \frac{2r_2+r_1}{2R}$, $\hat{\theta}_q = \frac{2s_2+s_1}{2S}$, $\hat{\theta} = \frac{2m_2+m_1}{2M}$ of θ_p, θ_q and θ respectively. Under the null hypothesis the estimates $\tilde{\theta}_p$, $\tilde{\theta}_q$, $\tilde{\theta}$ that maximise the likelihood should also satisfy the constraint $\tilde{\theta} = K\tilde{\theta}_p + (1-K)\tilde{\theta}_q$. Even with Laplace multipliers we were unable to find a closed form solution for this problem. For this reason we work numerically. We consider sequences with 0.001 increments for θ_p and θ. That is $\theta_p^{i+1} = \theta_p^i + 0.001$, $\theta_p^0 = 0$, $i = 0, 1, 2, \ldots 10000$. Similarly we have θ^j,

$j = 0, 1, 2, \ldots 10000$ where $\theta_p^i = \theta^i$. The sequence of θ_q is obtained by setting $\theta_q^k = \frac{\theta^k - K\theta_p^k}{1-K}$. Then we calculate the likelihood $L_{H_0}(\theta_p^i, \theta_q^j, \theta^k|\mathcal{G})_{i,j,k=0,10000}$ for all $(10001)^3$ points determined by the two-dimensional grid of θ_p and θ values and the values of θ_q that result from each point in the grid. We set $(\tilde{\theta}_p, \tilde{\theta}_q, \tilde{\theta})$ to be the point $(\theta_p^i, \theta_q^j, \theta^k)$ that returns the maximum likelihood value.

Since the difference in degrees of freedom between the null and alternative hypothesis is 1 (2 and 3 respectively), the asymptotic distribution of $2\log(LR_1)$ under H_0 is χ_1^2 (Casella and Berger, 2002, pp.490–491). We assume that we have pre-determined a threshold c. If the p-value of the LRT is less than c we reject the null hypothesis and proceed to stage 2a. If the p-value is greater than c we do not reject H_0 and move to stage 2b.

Stage 2a: Case-Control test for Genetic Association

Here, assuming that there was evidence for population heterogeneity in stage 1, we discard the cohort and test for association using only the cases and controls.

$$H_0 : \theta_p = \theta_q := \theta^* \qquad H_1 : \theta_p \neq \theta_q \tag{3.33}$$

If $\mathcal{G}^-$ is the data without the cohort, then the likelihood of (θ_p, θ_q) is given by:

$$L(\theta_p, \theta_q|\mathcal{G}^-) \propto \theta_p^{2r_2+r_1}(1-\theta_p)^{2r_0+r_1}\theta_q^{2s_2+s_1}(1-\theta_q)^{2s_0+s_1} \tag{3.34}$$

A LRT is used to test for association in this stage.

$$LR_{2a} = \frac{\max L_{H_1}(\theta_p, \theta_q|\mathcal{G}^-)}{\max L_{H_0}(\theta_p, \theta_q|\mathcal{G}^-)} = \frac{\max_{\{\theta_p,\theta_q\}} L(\theta_p, \theta_q|\mathcal{G}^-)}{\max_{\{\theta^*\}} L(\theta^*, \theta^*|\mathcal{G}^-)} \tag{3.35}$$

The values that maximise the likelihood under H_1 are the same unconstrained MLEs we used in stage 1; $\hat{\theta}_p = \frac{2r_2+r_1}{2R}$ and $\hat{\theta}_q = \frac{2s_2+s_1}{2S}$. Under H_0 the value of θ^* that maximises the likelihood is the MLE of the merged case-control sample, $\hat{\theta}^* = \frac{2(r_2+s_2)+r_1+s_1}{2(S+R)}$. Under the null hypothesis, $2\log(LR_{2a})$ has an asymptotic χ^2_1 distribution since there are 2 degrees of freedom under the alterative and degree of freedom under the null hypothesis.

Stage 2b: Case-Cohort-Control test for Genetic Association

If no evidence in stage 1 suggests that the cohort is not compatible with the case/control data we test for genetic association using all three samples.

$$H_0 : \theta_p = \theta_q = \theta := \theta^* \qquad H_1 : \theta = K\theta_p + (1-K)\theta_q \tag{3.36}$$

The likelihood we consider here, is the one in (3.31) and the LR test statistic we use in this stage is:

$$LR_{2b} = \frac{\max L_{H_1}(\theta_p, \theta_q, \theta|\mathcal{G})}{\max L_{H_0}(\theta_p, \theta_q, \theta|\mathcal{G})} = \frac{\max_{\{\theta_p,\theta_q,\theta:\theta=K\theta_p+(1-K)\theta_q\}} L(\theta_p, \theta_q, \theta|\mathcal{G})}{\max_{\{\theta^*\}} L(\theta^*, \theta^*, \theta^*|\mathcal{G})} \tag{3.37}$$

Note that the alternative hypothesis in this stage is essentially the null hypothesis of the first stage. Thus the values of $(\theta_p, \theta_q, \theta)$ that maximise the likelihood under H_1 are the same as the ones we used to maximise the likelihood under H_0 in stage 1. In this stage the value that maximises the likelihood under the null hypothesis is the pooled MLE, $\hat{\theta}^* = \frac{2(m_2+r_2+s_2)+m_1+r_1+s_1}{2N}$. The asymptotic distribution of LR_{2b} under the null is again χ^2_1 (2 degrees of freedom under H_1 and 1 df under H_0).

3.3.2 The Data

British CORGI 550K data comprises 928 cases (437 males, 491 females) with colo-rectal cancer (CRC) or high-risk colorectal adenomas and 931 controls (423 males, 508 females). For these data 555,352 tag SNPs were genotyped using Illumina Hap550 BeadChip Array. Our cohort group comprises genotypes from 1437 individuals (628 males, 809 females) from the 1958 British Birth Cohort (BC58). This was generated by the Wellcome Trust Sanger Institute in collaboration with the 1958 BC as part of the Wellcome Trust Case-Control Consortium (WTCCC).

The CORGI data were used in a number of published studies (Tomlinson et al., 2007; Broderick et al., 2007; Jaeger et al., 2008; Tomlinson et al., 2008; Pittman et al., 2008) regarding genetic association of CRC. Overall between all those studies four phases of meta-analyses were done. In each phase only a subset of the markers used in the previous phase is used - the ones deemed most significant - while additional individuals are genotyped. Of special interest in our study is Phase II described in Tomlinson et al. (2008). There, after an initial case-control analysis, 43,037 markers were deemed most significant and an additional 2873 controls and 2870 cases were genotyped (English Phase II data) as part of the National Study of Colorectal Cancer Genetics NSCCG. Although we do not have the data from Phase II we do have the results (p-values) which we use as reference in our study. Overall, as of the time of our study, there were 15 SNPs (Table 3.4) that were succesfully replicated for associataion with CRC. We take that into account in our analyses as well.

We do not study all 550K markers in our study but rather just the ones used in Tomlinson's Phase II. Of those 43,037 markers we further remove the ones from the sex chromosomes since initial analysis did not show any evidence for association for those markers. Addditionally we remove mitochondrial DNA markers plus markers with sample MAF less that 0.01. Overall we end up with 42,665 markers in our sample.

An estimate of the prevalence of CRC in the UK was obtained from the International Agency for Research on Cancer [3] (IARC) and according to 2002 studies this was 5 in 10,000 inividuals ($K = 5 \times 10^{-4}$).

An indication for the differences between the CORGI and BC58 samples is shown in figure 3.12. In it, we plot the ratios $\frac{\hat{\theta}^j}{\hat{\theta}^j_q}$ and $\frac{\hat{\theta}^j_p}{\hat{\theta}^j_q}$ for all $j = 1, \ldots 42,665$ SNPs in our data. Specifically $\hat{\theta}^j$ is the MAF sample estimate for SNP j in the cohort (BC58), $\hat{\theta}^j_p$ the MAF sample estimate for the same SNP in the cases (CORGI cases) and $\hat{\theta}^j_q$ the respective estimate in the controls (CORGI controls). Normally one would expect to see some extreme spikes in the $\frac{\hat{\theta}^j_p}{\hat{\theta}^j_q}$ line while the ones in the $\frac{\hat{\theta}^j}{\hat{\theta}^j_q}$ line should be more modest. However we can see that the majority and the most extreme of the spikes appear in the latter. That means that if we do not take into account potential heterogeneity in our samples we will most likely end up with falsely identifying the markers where the spikes in the $\frac{\hat{\theta}^j}{\hat{\theta}^j_q}$ line appear, as significant. Moreover for those SNPs where both lines have spikes - and there are a few clearly visible - we may reduce the potential of identifying true associations.

[3] www.iarc.fr

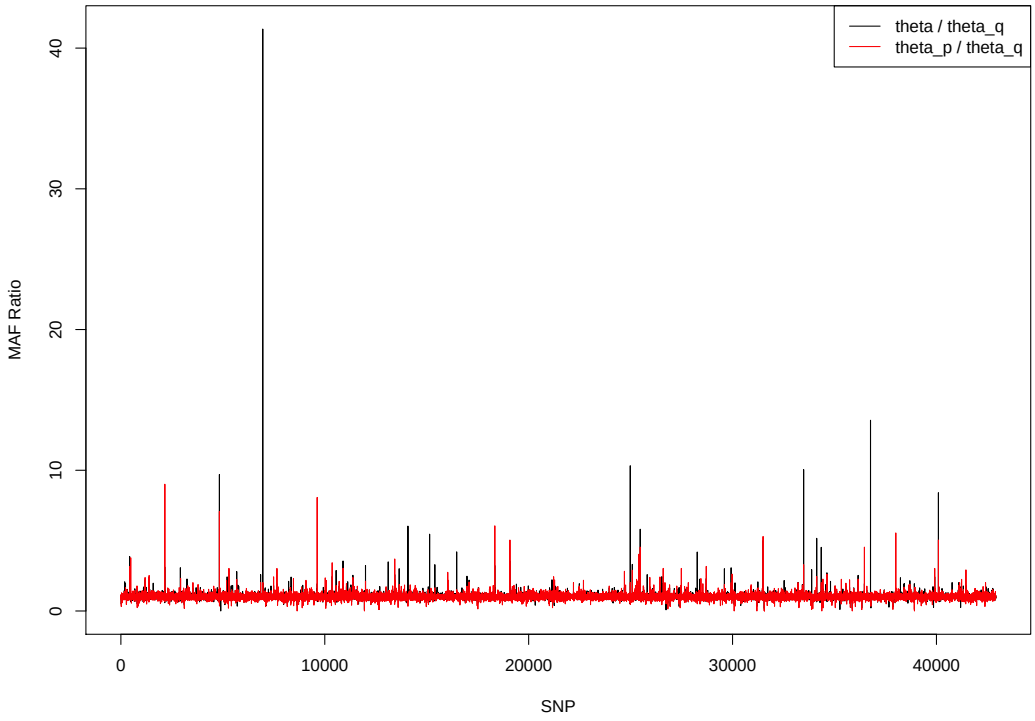


Figure 3.12: The ratios $\frac{\hat{\theta}^j}{\hat{\theta}^j_q}$ (black) and $\frac{\hat{\theta}^j_p}{\hat{\theta}^j_q}$ (red). It is clear that in some markers the BC58 data are significantly different from the CORGI data.

3.3.3 Analysis and Results

Our objective is to calibrate our two-stage LR test in order to increase its power when it comes to identifying true associations but at the same time avoid to increase the probability of identifying false associations as true. Specifically the part of our method that needs to be determined and has an effect in the overall power the testing process is the threshold c we use in the first stage.

As reference in our analysis we use the results from Tomlinson's Phase

II (hereby called simply Phase II). We do not directly compare the power of our method with that of Phase II since they used roughly three times larger case/control sample size than the CORGI data we use. Even with the addition of the BC58 cohort in our study we will certainly not be able to achieve the same power as any appropriate test that is applied on data that contain three times the cases we have in our data. Instead we try to find an optimal threshold c so that the results of our method will be as close as possible to those Phase II results.

We focus our analyses in three directions:

- The p-values for the 15 successfully replicated SNPs. Essentially, if we ignore the first stage of our method, we have two p-values for those SNPs, one from the case-control LRT and another from the case-cohort-control LRT. The threshold in the first stage determines which one we will use. Since the association between CRC and those SNPs has been proven, we want to choose a threshold that results in the most significant, i.e. smaller, p-values for those SNPs, the choice being between the stage 2a and stage 2b p-values.

- The ranking of the SNPs, especially the most significant ones. In GWAS the ranking of SNPs based on their p-values is as important as p-values themselves. This is especially true for studies that involve more than one phases. In this case the top-ranked SNPs are further genotyped in subsequent phases. Therefore in our study we to choose c such that the rank of the most significant SNPs are in accordance to their rank in Phase II.

- False discovery rate (FDR). The risk of identifying false associations concerns particularly the use of the cohort. Potential population heterogeneity may lead to false positives. Thus we want to choose a stage 1 threshold c that keeps FDR at reasonable levels.

The 15 SNPs

In general if the stage 1 p-value (p_{S1}) is less than c we proceed to stage 2a otherwise we proceed to stage 2b. Initially we compare the p-values of the 15 SNPs for $c = 0$ and $c = 1$ which means apply the stage 2b (CCC) test to all SNPs and apply stage 2a (CC) test to all SNPs respectively. For six SNPs (rs22822428, rs16892766, rs2989734, rs3802842, rs4779584 and rs4939827) the CCC test gives more significant (smaller) p-values while for the rest, the CC test gives more significant p-values (Table 3.4). This means that using the CCC design without testing for population homogeneity renders the p-values of nine of the CRC associated SNPs comparatively insignificant. Yet, setting $c = 1$ means we completely ignore the cohorts. The ideal threshold would be one that sends the six SNPs (Group 1 SNPs) in stage 2b and the others (Group 2 SNPs) to stage 2a. So, we begin our analysis with $c = 0$ and then gradually increase the threshold. At $c = 0.05$ two SNPs from Group 2 are directed to stage 2a (Figure 3.13 - Left). So at this point 8 of the 15 SNPs obtain their optimal p-value. Further increasing the thresholld to $c = 0.1$ results in two more SNPs from Group 2 to be directed to stage 2a (Figure 3.13 - Right). No change occures until $c \approx 0.18$ where SNPs from Group 1 begin to being directed to stage 2a. Therefore the best threshold we can achieve with regards to the 15 SNPs is $0.1 < c < 0.18$ (Figure 3.14).

For any value in that range there are 10 SNPs that achieve their optimal p-value. Additionally for the five SNPs that do not (rs11590577, rs4355419, rs10795668, rs2488704, rs12957142), the difference between the optimal CC p-value and their CCC p-value which is the one used for that threshold, is very small, especially compared with the improvement of the other SNPs.

SNP	p_{S1}	$-\log_{10}(p_{S2a})$	$-\log_{10}(p_{S2b})$	$\log_{10}(p_{S2a})-\log_{10}(p_{S2b})$	Phase II rank
rs6983267	0.01	6.60	5.09	-1.51	1
rs16892766	0.81	2.17	2.75	0.58	4
rs4939827	0.44	6.01	9.68	3.67	5
rs4779584	0.79	2.64	3.33	0.69	7
rs4841306	0.01	3.07	1.70	-1.37	9
rs3802842	0.19	0.74	1.86	1.12	11
rs2488704	0.21	2.10	1.75	-0.35	12
rs2164182	0.09	2.59	1.91	-0.68	13
rs11590577	0.29	2.38	2.23	-0.15	16
rs4822442	0.07	2.47	1.70	-0.77	17
rs2989734	0.56	0.89	1.56	0.67	19
rs2282428	0.68	2.12	2.48	0.36	23
rs4355419	0.20	2.14	1.76	-0.38	27
rs12957142	0.49	1.30	1.24	-0.06	31
rs10795668	0.34	2.08	1.95	-0.13	39

Table 3.4: The p-values of the 15 SNPs for all stages of our method.

The ranking of the SNPs

We consider that the SNP with the most significant p-value has rank 1. We consider the following three measures:

1. Top ν Mean Rank Difference (ν-MRD): We consider the SNPs with Phase II ranks $1, \ldots \nu$ and their respective ranks when we apply our method for a Stage 1 threshold c. Then ν-MRD is the average difference of those ranks.

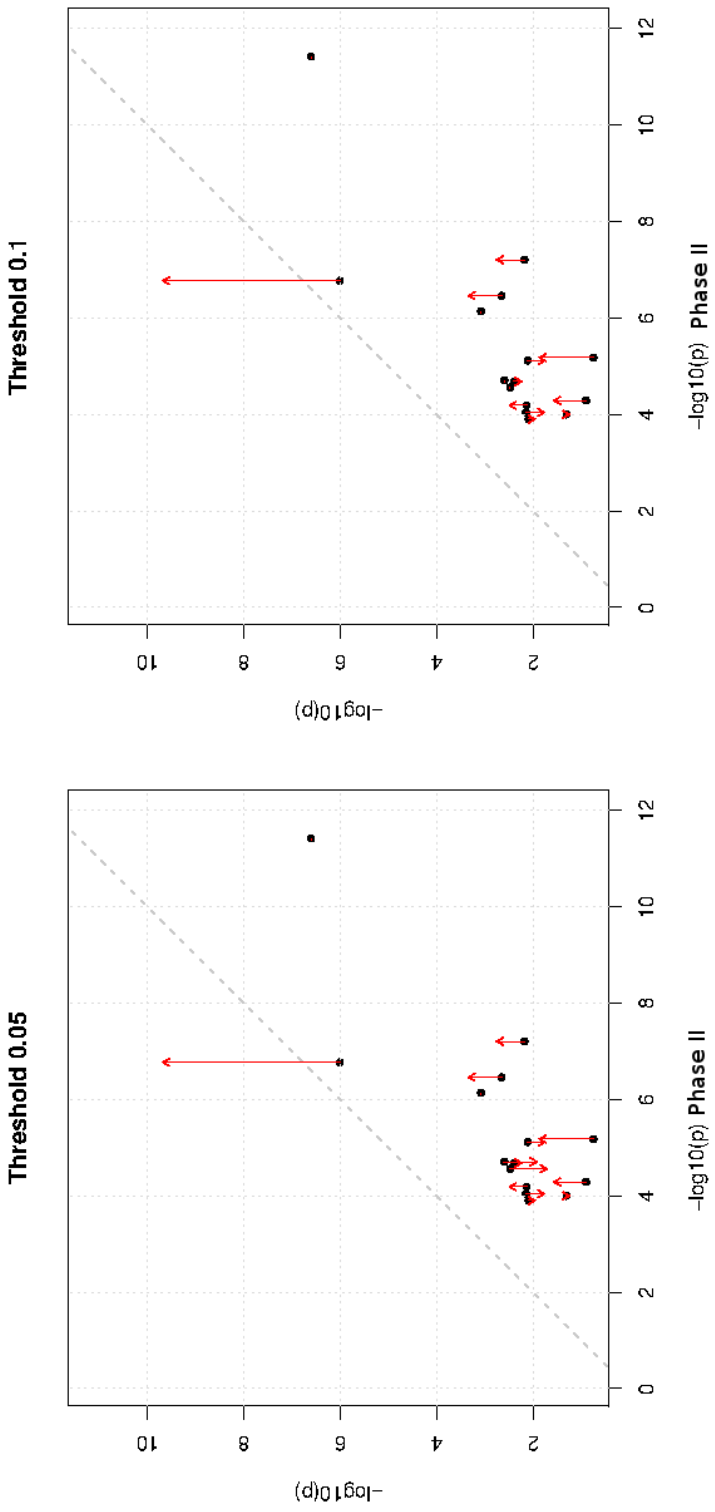


Figure 3.13: Change in the p-value of the 15 SNPs, relative to their CC p-value for $c = 0.05$ (left) and $c = 0.1$ (right). The latter threshold gives more significant p-values for two of those SNPs (seen in the cluster of the three points below the diagonal).

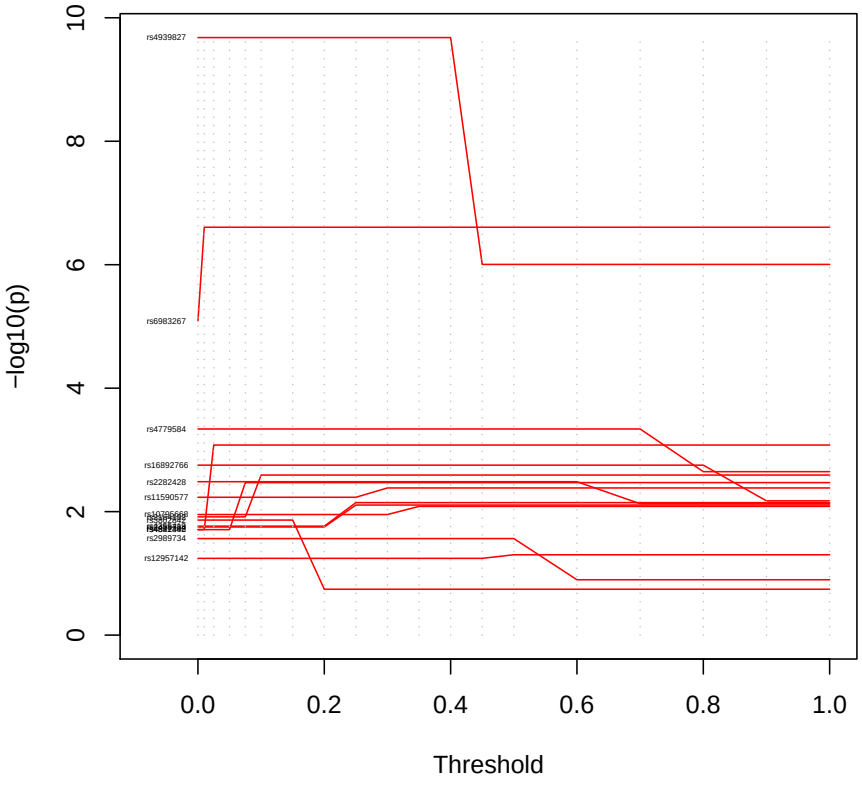


Figure 3.14: The p-value of the 15 SNPs for various values of threshold c.

2. Top ν Mean Absolute Rank Difference (ν-MARD): Similar to ν-MRD, the difference being that we consider the absolute rank differences instead of the raw differences.

3. Weighted ν Mean Absolute Rank Difference (W-ν-MARD): Here we consider the absolute rank difference between Phase II and our method for all SNPs. We calculate the weighted average of those absolute differences using weights equal to 1 for the top ν ranked SNPs based on their Phase II p-values and weights equal to 0.5 for the SNPs with rank greater than ν.

We seek for the Stage 1 threshold that minimises those three measures. The first two take into account only the most significant SNPs, according to their Phase II results. The third measure, in addition, takes into account the rest of the SNPs as well, but gives more weight to the most significant SNPs. If we were equally considering all the ranks, then changes in the ranks of non-significant SNPs will influence greatly the outcome. For example, if a SNP is ranked in the 30,000s and drops to the 40,000s or moves up to 20,000s is not as significant as a SNP that is ranked in the first twenty drops in the hundreds. The least significant SNP among the 15 that were successfully replicated for association has rank 39, so ν does not need to be very large in order to include all the possible SNPs that are associated with CRC. For $\nu = 50$ all three measures are minimised for $c = 0.025$ and for $\nu = 100$ all three are minimized for $c = 0.15$ (Figure 3.15).

False Discovery Rate

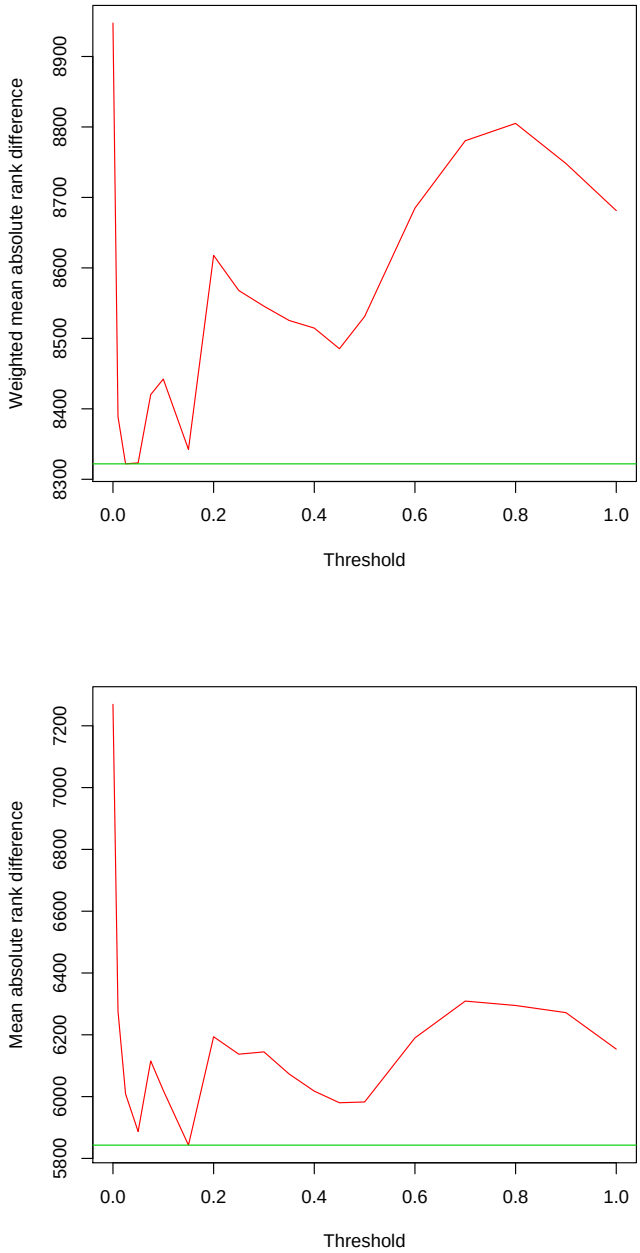


Figure 3.15: Top: Weighted-50-Mean Absolute Rank Difference; Bottom: 100-Mean Absolute Rank Difference

We measure the false discovery rate as the proportion of the SNPs that are ranked in the $k\%$ least significant according to Phase II while they are ranked in the top 10% most significant according to our method with a Stage 1 threshold c. Figure 3.16 shows this for $k = 5, 10, 15$ and 20. There appear to be two local minima in the FDR. One for $c \approx 0.15$ and the other for $c = 1$ i.e. under the CC design.

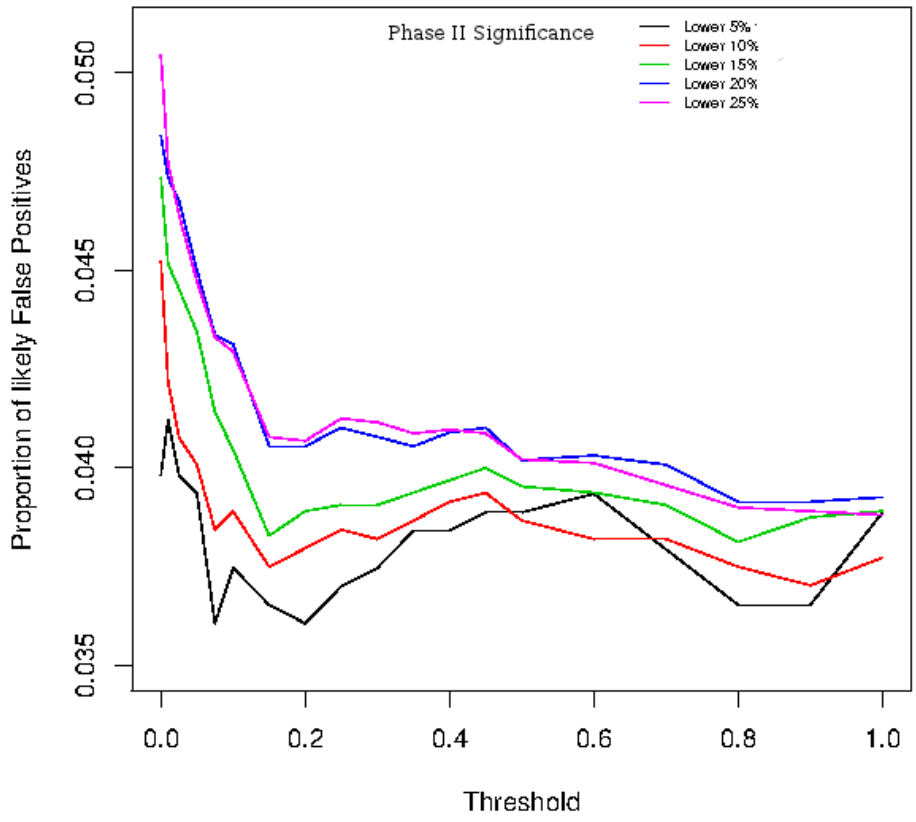


Figure 3.16: False Discovery Rate. Proportion of SNPs that are among the $5\%, 10\%, 15\%, 20\%$ and 25% least significant according to Phase II and move among the 10% most significant when we use our method.

So combining the results concerning the 15 replicated SNPs, the top ranked Phase II SNPs and the FDR we conclude that for this particular dataset the optimum Stage 1 threshold is $c \approx 0.15$. This increases the potential to identify true associations compared to the case-control test but also reduces the risk to obtain false positives which is prevalent when we use the case-cohort-control design due to potential population differences. Of course the choice we make here is based on the results we obtained from Phase II, where much more powerful tests were applied because of the increased sample size. In other studies nothing can guarantee that this threshold will work. However, given the nature of the test we use in Stage 1, which is a population compatibility test, we believe that using a relatively lenient threshoshold, something between 0.05 and 0.15, is reasonable enough, and we believe can be used to pick up all the true population differences in any GWAS, where a cohort sample is available. Due to the lenient threshold, false positives in those differences are likely to occur but for the majority of the markers the optimal between the two designs will be used.

3.4 Discussion

In this chapter we study how the power of genetic association tests can be increased by extending the common Case-Control design to a Case-Cohort-Control design. Initially we compare the power of three commonly used tests under both designs, using simulations. We see that Cochran-Armitage test is more powerful than the other tests under CCC design. Further due to bias-

variance tradeoff neither of the two designs gives more powerful tests, for all parameters' values. The parameter that appears to have the most impact on that is the disease prevalence. We do additional power calculations, adopting a formula for the asymptotic power of Cochran-Armitage test and using phenotype misclassification theory to correctly calculate the power under the CCC design. We introduce the notion of "maximum rare disease prevalence", the value of disease prevalence which depends on the other parameters (MAF and sample size) and essentially determines which of the two designs will give most powerful tests. Finally, we introduce a new method that takes into account potential population heterogeneity. The method comprises of two stages and likelihood ratio tests are used in both stages. We have to choose an optimal threshold for the Stage 1 test. We based the decision on the results we obtained from a study of the same disease that used three times the amount of data we have available. Nevertheless, we believe that a lenient Stage 1 threshold, similar the one we ended up using in our study, can be used in any association study where there are available data for a Case-Cohort-Control design and improve the chances of identifying true associations.

Chapter 4

The Kin-Cohort Design

Family-based designs are not uncommon in GWAS. Hopper et al. (2005) provide a detailed overview and discussion on some of the most common population-based family designs used in genetic epidemiology. Another example is the transmission disequilibrium test (Spielman et al., 1993) where data tranditionally comprise of family trios (two parents and an affected offspring) although studies with more affected children are also possible. Here we consider a hybrid study design known as the kin-cohort design (Gail et al., 1999) although sometimes studies that are based on this design are called case-control family studies. In this design we still have the usual case and control samples, with known genotypes and disease phenotypes, with additional information regarding the disease status (phenotype) of their, usually first-degree, relatives. Under this design the genotyped individuals in the case-control samples are called the probands. The benefit of this design is that, for many diseases, the familial phenotype information can be obtained

by simply interviewing the probands thus having no additional financial cost or in cases where a proper diagnosis is needed, the additional cost is minimal especially compared to the advantages we gain when we properly put this information to use.

4.1 A Bayesian Approach

In this section we describe a Bayesian approach for inference in GWAS, using the kin-cohort design. Here we consider the case where we know the disease status of the probands' first-degree relatives, specifically parents and siblings (Figure 4.1). Bayesian models have been used in family-based GWAS (e.g. Naylor et al., 2010). Most of them, however, focus on identifying and picking up markers that are associated with the disease and not so much on actually quantify the effect size of those markers. In our method we put emphasis on actually estimating the effect size and how to increase the precision of this estimate. We develop a hierarchical model with prospective likelihood where we adopt the logistic disease model. The idea is to probabilistically impute the unknown genotypes and use them in our analysis. We assume Mendelian inheritance and HWE hold. Thus the unconditional distribution of the genotypes in the population is given by (2.2). However the conditional distribution of the children's genotypes given the parental genotypes is independent of the MAF (Table 4.1).

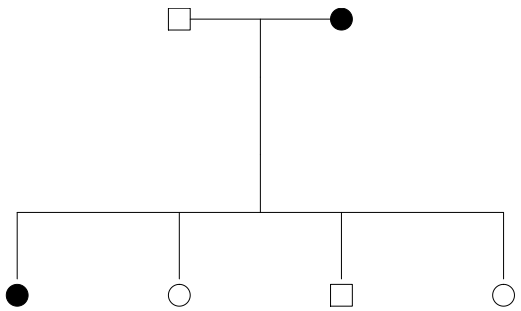


Figure 4.1: An example of a family pedigree used in our kin-cohort design. The shape indicates the sex of an individual and those filled with black have the disease. The proband is one of the children.

4.2 The Statistical Model

Let $(x_i^{(m)}, y_i^{(m)})$ and $(x_i^{(p)}, y_i^{(p)})$ denote the unobserved genotype and observed disease status of the mother and father for the i-th family, and $\{(x_i^{(s_j)}, y_i^{(s_j)})\}_{j=1}^{n_i}$ the genotype and disease status of the n_i children, $i = 1, \ldots, N$. Without loss of generality, we assume that the children with superscript s_1 are the probands and also that those with $i = 1, \ldots, N_0$ are the controls while those with $i = N_0 + 1, \ldots, N$ are the cases. The size of the controls is N_0 and of the cases N_1 ($N = N_0 + N_1$).

4.2.1 Likelihood

We consider the standard logistic disease model (2.5). If we ignore the familial structure and consider all $\sum_{i=1}^{N} n_i + 2N$ individuals independent, then the likelihood of (α, β) is simply the product of all the individual probabilities

Child genotype, $x^{(s)}$	Parental genotypes, $(x^{(p)}, x^{(m)})$								
	(AA, AA)	(AA, Aa)	(AA, aa)	(Aa, AA)	(Aa, Aa)	(Aa, aa)	(aa, AA)	(aa, Aa)	(aa, aa)
AA	1	1/2	0	1/2	1/4	0	0	0	0
Aa	0	1/2	1	1/2	1/2	1/2	1	1/2	0
aa	0	0	0	0	1/4	1/2	0	1/2	1

Table 4.1: Conditional child genotype probabilities given parental genotypes assuming Mendelian inheritence, $\pi(x_i^{(s)}|x^{(p)}, x^{(m)})$. Of course, in our model we use $AA = 0$, $Aa = 1$ and $aa = 2$.

$p_y(x, \alpha, \beta)$, where

$$p_y(x, \alpha, \beta) = P(Y = y|x, \alpha, \beta) = \frac{[\exp(\alpha + \beta x)]^y}{1 + \exp(\alpha + \beta x)}, \quad y \in \{0.1\}. \tag{4.1}$$

However the familial structure in our data lead to ascertainment bias (Thomas, 2004, pp. 141–148) and thus using a likelihood where the observations are assumed independent will lead to heavily biased estimates of the effect size (β). For this reason we introduce weights in our likelihood in order to remove the acertainment bias. Our rationale is similar to what Thornton and McPeek (2010) are using albeit in a different and less complicated setting. We call these weights *relevance weights* because essentially they indicate how relevant an individual is in our sample. The notation and the basis of the methodology we use to construct the weights come from Hu et al. (2000), where the authors discuss the concept of *relevance weighted likelihood for dependent data*. Those weights are the ratios of the unconditional probability to observe the genotype to the conditional probability to observe the same genotype, given the proband's genotype [1]. Of course the families are still assumed independent; the weights are designed to remove the intra-familial

[1] More details can be found in Appendix A.

correlations.

$$\begin{aligned} w_i^{(p)} &= \frac{P(x_i^{(p)})}{P(x_i^{(p)}|x_i^{(m)}, x_i^{(s_1)})}, \\ w_i^{(m)} &= \frac{P(x_i^{(m)})}{P(x_i^{(m)}|x_i^{(p)}, x_i^{(s_1)})}, \\ w_i^{(s_1)} &= 1, \\ w_i^{(s_j)} &= \frac{P(x_i^{(s_j)})}{P(x_i^{(s_j)}|x_i^{(s_1)})}, \quad j = 2 \ldots n_i. \end{aligned} \tag{4.2}$$

Incorporating the weights in (4.2) we get the weighted log-likelihood

$$\begin{aligned} l(\alpha, \beta) = \sum_{i=1}^{N} \Big\{ & w_i^{(p)} \log p_{y_i^{(p)}}(x_i^{(p)}, \alpha, \beta) + w_i^{(m)} \log p_{y_i^{(m)}}(x_i^{(m)}, \alpha, \beta) + \\ & \log p_{y_i^{(s_1)}}(x_i^{(s_1)}, \alpha, \beta) + \sum_{j=2}^{n_i} I(n_i > 1) w_i^{(s_j)} \log p_{y_i^{(s_j)}}(x_i^{(s_j)}, \alpha, \beta) \Big\} \end{aligned} \tag{4.3}$$

4.2.2 Prior Distributions

We consider normal prior distributions for the coefficients of the model,

$$\alpha|\sigma_\alpha^2 \sim N(0, \sigma_\alpha^2) \tag{4.4}$$

$$\beta|\sigma_\beta^2 \sim N(0, \sigma_\beta^2). \tag{4.5}$$

We adopt an Inverse Gamma distribution on the variance parameter of the intercept and Gamma distribution for the variance parameter of the slope

coefficient (effect size).

$$\sigma_\alpha^2 | a_\sigma, b_\sigma \sim IG(a_\sigma, b_\sigma) \tag{4.6}$$

$$\sigma_\beta^2 | \lambda \sim \Gamma(1, \lambda^2/2) \tag{4.7}$$

Although the Inverse-gamma is a conjugate prior for the scale parameter of the Normal distribution and thus simpler in use, we opt for a Gamma prior for the slope, the parameter that the inference will be based upon. Marginally, this particular hierarchical structure induces a double-exponential [2] prior on β which encourages shrinkage,

$$\pi(\beta) = \int \pi(\beta|\sigma_\beta^2)\pi(\sigma_\beta^2|\lambda)d\lambda \overset{\mathcal{D}}{=} DE(\lambda). \tag{4.8}$$

The conditional posterior of σ_β^2, under the Normal-Gamma prior model, is Generalised Inverse Gaussian (Griffin and Brown, 2010).

Finally, for the population MAF we adopt a beta prior.

$$\theta | a_\theta, b_\theta \sim \text{Beta}(a_\theta, b_\theta). \tag{4.9}$$

4.2.3 Posterior Inference

We perform inference for this model using a Markov chain Monte Carlo (MCMC) algorithm. We use Gibbs Sampling from the following conditional distributions except when we update (α, β) where we use a Metropolis-

[2] The double-exponential distribution is also known as Laplace distribution

Hastings step.

4.2.3.1 Updating the MAF (θ)

When we update the minor allele frequency caution must be taken in order to ensure that the expected values of the samples from the conditional posterior distribution are equal to the unconditional population MAF. Because of the ascertainment bias we described previously, we cannot use all the genotypes in a beta-binomial model. This will result to biased posterior samples of θ; the bias will be towards the conditional MAF of the affected subpopulation, provided that we do not deal with a very common disease and the affected individuals are the ones that are over-representent in our sample. Obviously the children's genotypes do not depend on θ given the parental genotypes and thus the conditional posterior of θ depends only on the parental genotypes. However even for the parents of cases and controls, the ascertainment bias is still present. For this reason we consider three subsets of the parental genotypes and use a mixture of posterior conditional distributions. The first subset is the alleles that the controls' parents passed to their proband offspring. Similarly the second subset is the alleles that the cases' parents passed to their proband offspring. These are simply the controls' and cases' genotypes respectively $(x_i^{(s_1)}, i = 1, \ldots, N)$. The third subset, which we denote x^{par}, comprise of the parental alleles that were not passed to the probands $(x_i^{par} = x_i^{(p)} + x_i(m) - x_i^{(s_1)})$. We use a mixture of posterior beta

distributions.

$$\pi(\theta|\{x_i^{(p)}, x_i^{(m)}\}_{i=1}^N) = \pi(\theta|\{x_i^{(s_1)}, x_i^{par}\}_{i=1}^N) = \phi(1-K)\pi(\theta|\{x_i^{(s_1)}\}_{i=1}^{N_0}) + \phi K\pi(\theta|\{x_i^{(s_1)}\}_{i=N_0+1}^N) + (1-\phi)\pi(\theta|\{x_i^{par}\}_{i=1}^N) \quad (4.10)$$

In (4.10), K is the disease prevalence in the population, while ϕ is a constant in $(0, 1)$. We use 0.5, but any value will provide unbiased posterior samples. Since the genotypes of the probands, $x^{(s_1)}$, are known while x^{par} are imputed someone may want to put more weight on the former. To justify the mixture distribution in (4.10) we note that the cases and controls come from the subpopulations of affected and unaffected respectively. If those subpopulations have MAF θ_{af} and θ_{un} respectively then clearly $\theta = K\theta_{af} + (1-K)\theta_{un}$. Additionally we consider that the haplotypes that were not passed to the probands have MAF equal to that of the total population.

4.2.3.2 Updating the imputed parental genotypes $(x^{(p)}, x^{(m)})$

We use a discrete distribution to update the parental genotypes. These are updated in pairs $(x^{(p)}, x^{(m)})$ and hence there are 9 possible pairs. The probabilities of those pairs are obtained from the conditional posterior distribution.

$$\pi(x_i^{(p)}, x_i^{(m)}|y_i^{(p)}, y_i^{(m)}, \{x_i^{(s_j)}\}_{j=1}^{n_i}, \alpha, \beta, \theta) \propto \pi(y_i^{(p)}|x_i^{(p)}, \alpha, \beta)\pi(y_i^{(m)}|x_i^{(m)}, \alpha, \beta)\pi(x_i^{(m)}, x_i^{(p)}|x_i^{(s_1)}, \theta) \times \prod_{j=2}^{n_i} \pi(x_i^{(s_j)}|x_i^{(m)}, x_i^{(p)}) \quad (4.11)$$

4.2.3.3 Updating the imputed siblings' genotypes ($x^{(s_j)}$, $j \geq 2$)

Similar to the parental genotypes, we use a discrete distribution for the genotypes of the probands' siblings. Each child is updated independently so there are only 3 possible values (genotypes). The conditional posterior distribution gives the probabilities of the genotypes.

$$\pi(x_i^{(s_j)}|x_i^{(p)}, x_i^{(m)}, y_i^{(s_j)}, \alpha, \beta) \propto \pi(y_i^{(s_j)}|x_i^{(s_j)}, \alpha, \beta)\pi(x_i^{(s_j)}|x_i^{(p)}, x_i^{(m)}),$$
$$j = 2, \ldots, n_i \quad (4.12)$$

4.2.3.4 Updating the regression coefficients (α, β)

The conditional posterior distribution of the regression coefficients (α, β) is complex and it is not one of the standard distributions. For this reason we use an Independent Metropolis-Hastings step in order to update the coefficients. The proposal distribution is a multivariate normal. The mean and covariance matrix of this distribuition are obtained by fitting the generalised linear model,

$$\text{logit}(y) = \alpha + \beta x \quad (4.13)$$

using the weights we discussed in (4.2). All the genotypes (parents and children) and phenotypes are used. After we fit the model we use the estimates $(\hat{\alpha}, \hat{\beta})$ as the mean and the covariance matrix $\widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta})$ as the covariance matrix of the proposal distribution.

4.2.3.5 Updating the prior variances $(\sigma_\alpha^2, \sigma_\beta^2)$

The conditional posterior distributions of the prior variances σ_α^2 and σ_β^2 are inverse gamma and generalised inverse gaussian respectively.

$$\pi(\sigma_\alpha^2|\alpha, a_\sigma, b_\sigma) \propto \pi(\alpha|\sigma_\alpha^2)\pi(\sigma_\alpha^2|a_\sigma, b_\sigma) = IG(a_\sigma + 0.5, b_\sigma + \frac{\alpha^2}{2}) \quad (4.14)$$

$$\pi(\sigma_\beta^2|\beta, \lambda) \propto \pi(\beta|\sigma_\beta^2)\pi(\sigma_\beta^2|\lambda) = GIG(0.5, \lambda^2, \beta^2) \quad (4.15)$$

4.3 Applications on simulated data

In order to apply our method we simulate familial data. The simulation process comprises of the following steps.

1. Specify values for the population MAF (θ) and disease prevalence (K)
2. Specify the proportion (or δ) of the total disease prevalence that is due to the marker under study. In other words, the genetic effect of the marker "causes" the disease prevalence to increase (or decrease) from $K - \delta K$ ($K + \delta K$) to K. Essentially by choosing δ we indirectly choose the effect size; the larger the proportion we choose the larger the effect size.
3. Given the values of θ, K and δ we calculate the actual values of the effect size (β) and the slope (α) of the logistic model.
4. Simulate a population of parental genotypes using the θ and assuming HWE.

5. For each pair of parents generate the family size using $\min\{3, N(3.7, 1)\}$.

6. Assuming Mendelian inheritance simulate the children genotypes using the parental genotypes. The first child in each family is considered to be the proband.

7. Assign phenotypes to all individuals (parents and children) using the logistic model (4.1).

8. Discard all genotypes other than the probands.

9. Randomly select N_0 individuals from the non-affected probands (controls) and N_1 individuals from the affected probands (cases).

10. Apply the method using the genotypes of the cases and controls as wall as theirs' and their families' phenotypes.

4.3.1 Numerical values

In our simulations we use $\theta = \{0.05, 0.25\}$, so we consider the case of a rare allele and the case of a relatively common one. We choose $K = \{0.01, 0.05\}$ which is close to the prevalence of many diseases studied in GWAS. Finally we choose $\delta = \{0.01, 0.05, 0.1\}$. We apply our method to all $2 \times 2 \times 3 = 12$ combinations of the parameters. Note that the numerical value of β does not depend only on δ and is different for each set of parameters. Also without loss of generality, here we consider only causal alleles.

4.3.2 Results

In order to assess our method, we, additionally, fit the hierarchical Bayesian model for the case-control design. In this model we use the same prior distributions for the probands but we ignore the phenotypes of their first degree relatives and thus omit the "imputation" step in the MCMC algorithm. We compare the two models both in terms of bias and in terms of precision of their effect size estimates.

The MCMC algorithms converge, in general, relatively fast without long burn-in period (Figure 4.2). Based on the partial autocorrelation function (PACF) we thin the posterior samples by taking one every five samples.

Normally the bias should not be an issue for either the case-control or the kin-cohort design (Figures 4.3 and 4.4). Indeed, the weights we describe in (4.2) ensure that the kin-cohort design has the same unbiased properties as the case-control design. Additionally figures show that the precision of the effect size estimate increases when we use the kin-cohort design. The conclusions we discuss in the following paragraph, are valid under the assumption of unbiasedness for the estimates of the effect size. Nevertheless, in the following section (4.3.2.1) we present some additional analyses where we do not make that assumption.

In order to get a better idea on how the three parameters (δ, θ, K) affect the precision of the posterior estimate of β we calculate the ratio of the variance of the posterior sample of βs obtained from the case-control model to the variance of the posterior sample of βs obtained from the kin-cohort

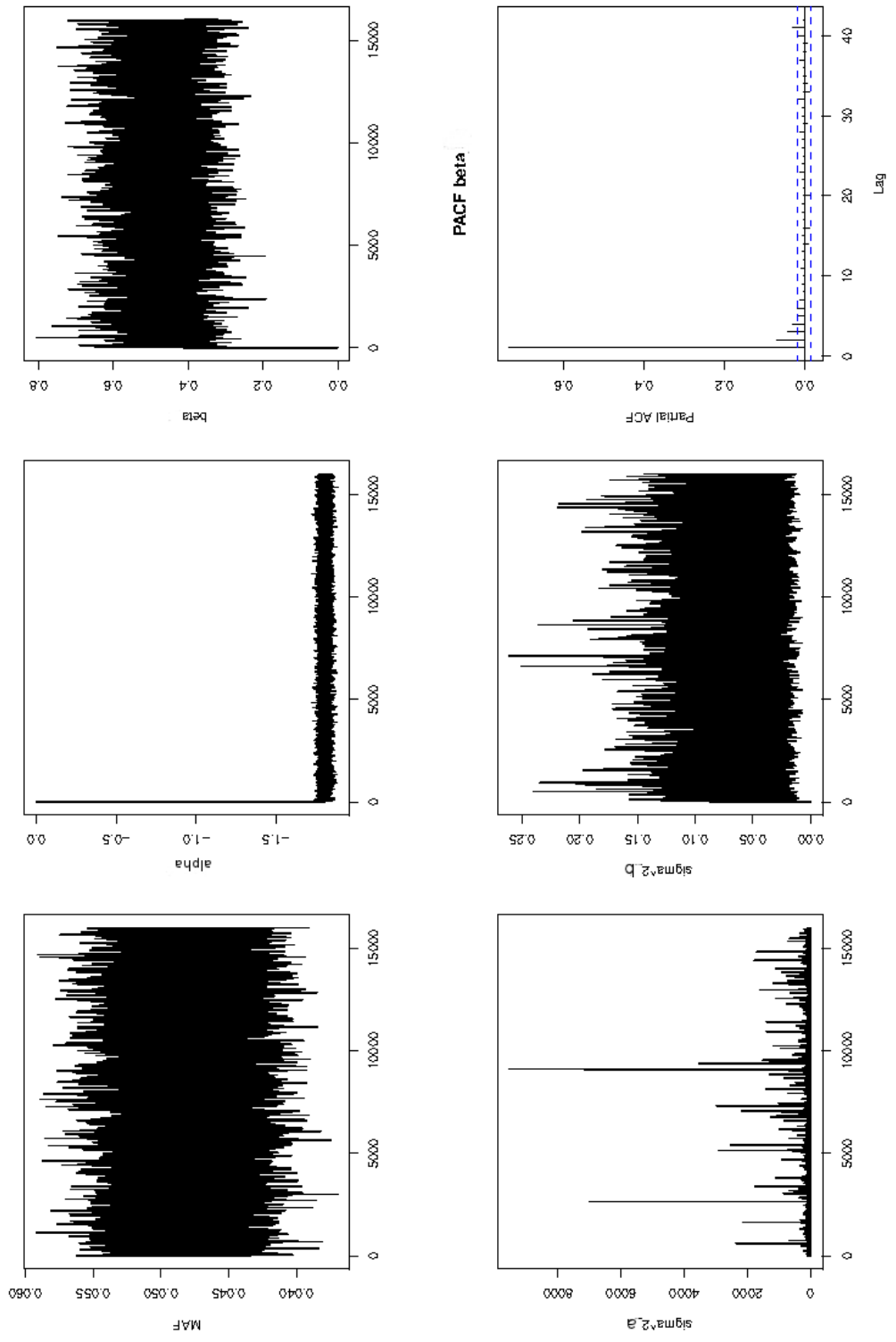


Figure 4.2: Trace plots and PACF plot of the posterior samples of β under the Kin-Cohort design. The values here are $\theta = 0.05$, $\delta = 0.05$ and $K = 0.05$ which results in the true value of $\beta = 0.445$.

model (Table 4.2). Although we do not use a large variety of values for the various parameters we can still derive some preliminary conclusions based on the ones we used. The minor allele frequency does not appear to affect the relative precision of the kin-cohort design compared to the case-control design. The ratios of the variances are roughly equal for the two values of θ. On the other hand the effect of δ seems to depend on the third parameter, K. That is, for a not so rare disease (5% prevalence in our example), δ does not seem to affect the ratio of the precisions. All ratios are between 1.23 and 1.35 in favour of the kin-cohort design, with no clear trend that will indicate dependance on δ. However when we deal with a more rare disease (1% prevalence) then the precision of the kin-cohort deisign estimate increases compared to the precision of the case-control estimate, as δ (i.e. the effect size) increases. For small effect size, when the marker is responsible for 1% of the total disease prevalence the precision of the kin-cohort design estimate is just slightly better that the case-control estimate. For larger effect size, when the marker is responsible for 10% of the disease prevalence the precision obtained from the kin-cohort design is at least 1.5 times better than the precision obtained from the case-control design.

Table 4.2: Ratios $Var\left(\beta^{(C-C)}\right)/Var\left(\beta^{(K-C)}\right)$

	$\theta = 0.05$		$\theta = 0.25$	
	$K = 0.01$	$K = 0.05$	$K = 0.01$	$K = 0.05$
$\delta = 0.01$	1.11	1.26	1.16	1.26
$\delta = 0.05$	1.35	1.23	1.34	1.35
$\delta = 0.10$	1.50	1.28	1.51	1.27

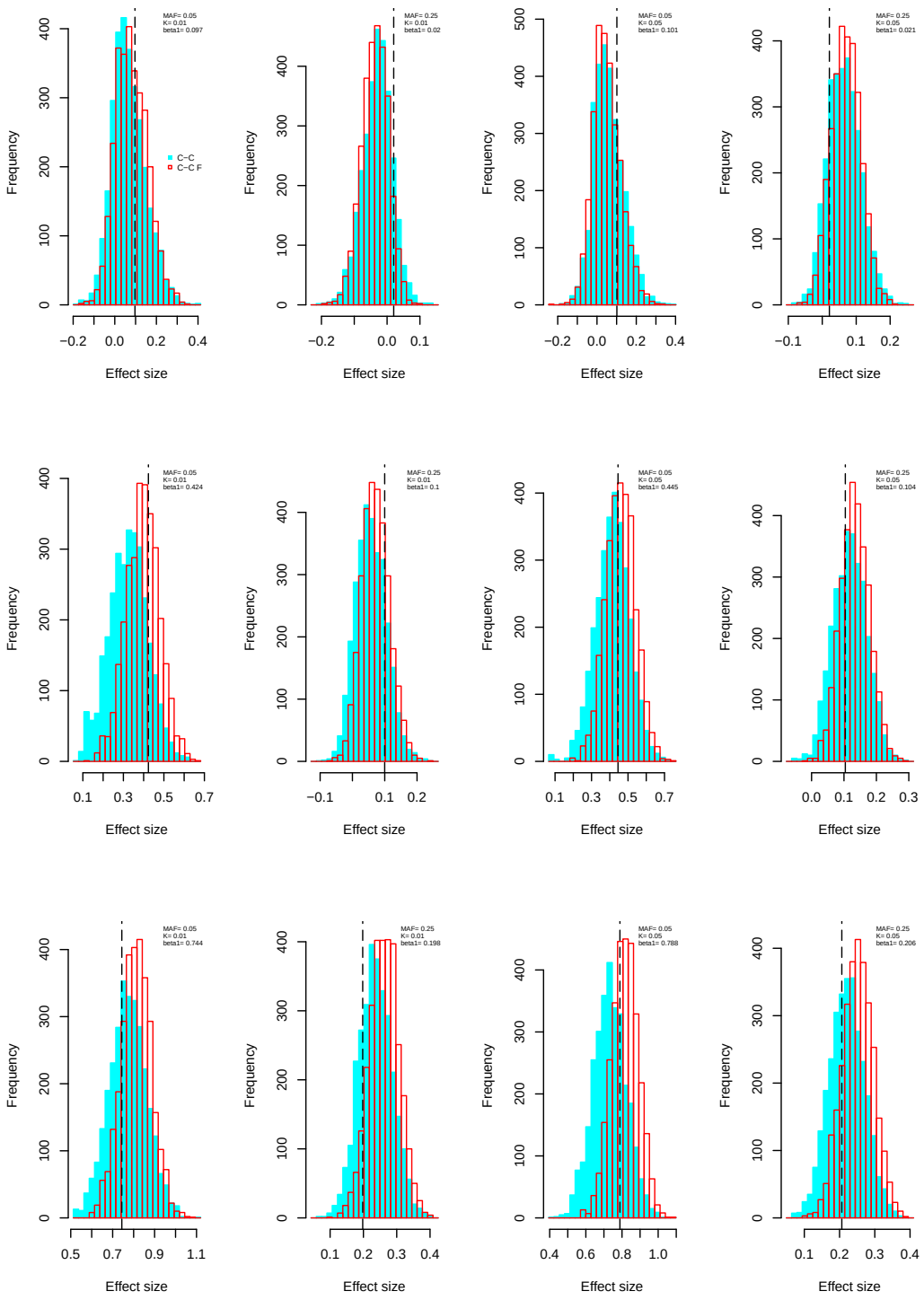


Figure 4.3: Histograms of posterior samples of the effect size, β for both the kin-cohort (red) and case-control (sky blue) designs. The vertical line shows the value used to simulate the data.

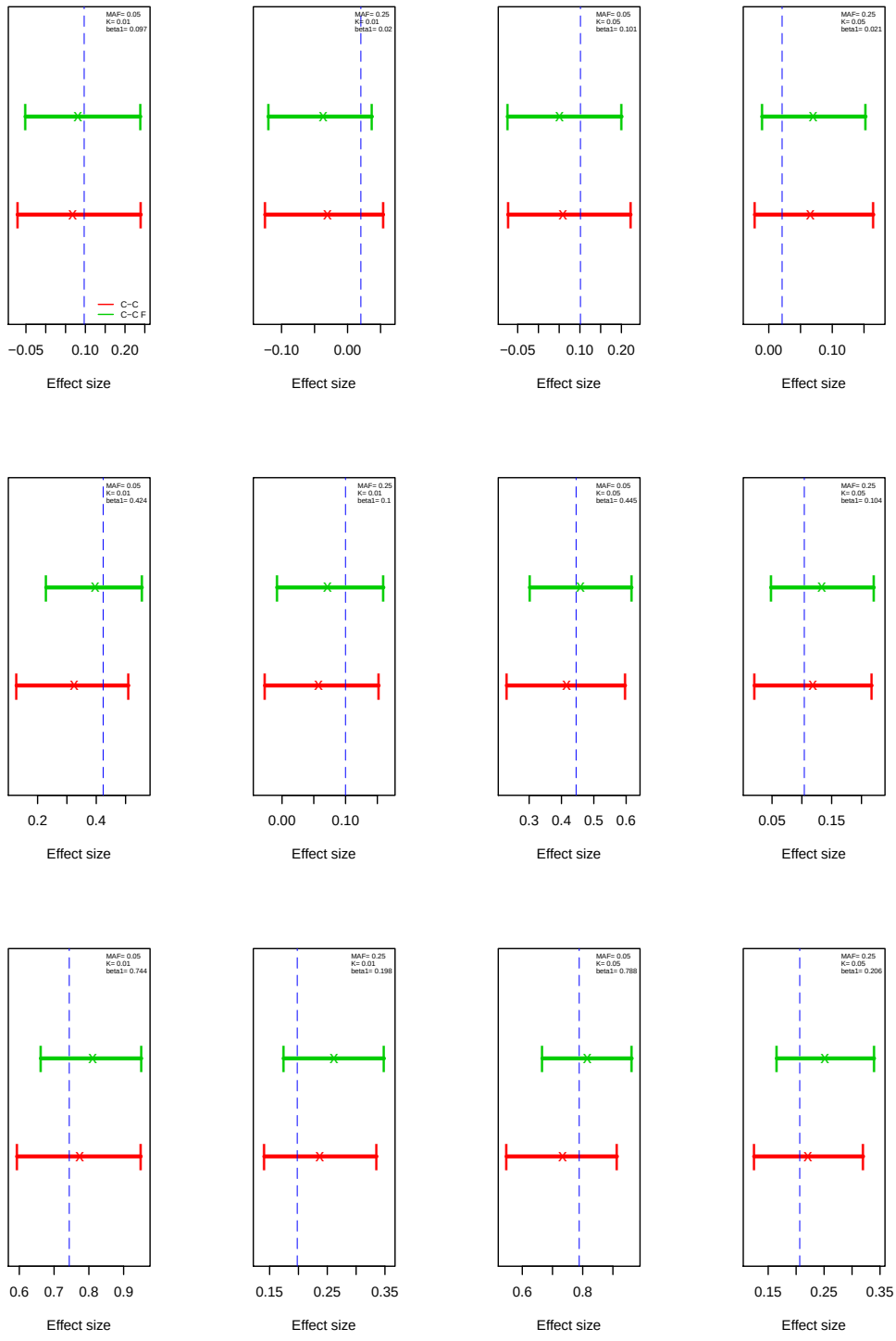


Figure 4.4: 95% credible intervals for the effect size parameter. Kin-cohort (green) and case-control (red). The x indicates the posterior sample mean and the vertical broken line the actual value of β used to simulate the data.

4.3.2.1 Observed Mean Squared Error

The weights we introduce in (4.2) and use in the log-likelihood (4.3) aim to remove the bias of the estimators $\hat{\beta}$ by taking into account the correlations between family members. If all the genotypes are known then the maximum likelihood estimates we obtain for the coefficients of the logistic model through the iteratively reweighted least squares (IRLS) algorithm are indeed unbiased. In our model those ML estimates (using the imputed genotypes) are used as parameters of the proposal distribution in the Independent Metropolis-Hastings step of our MCMC algorithm, to subsequently obtain what is essentially a maximum à posteriori (MAP) estimate.

Therefore we cannot claim beyond any doubt that the MAP estimate has the same unbiased property as the MLE. The fact that all the MAP estimates are always within the boundaries of the credible intervals in figure 4.4 supports their unbiasedness. Withought the use of the relevance weights, the estimates we obtain are far from being inside those intervals. Nevertheless we want to assess the effect size estimators we obtain from the methodology we propose in this chapter without explicitly assuming that they are unbiased.

The mean square error (MSE) is a measure that quantifies the quality of an estimator that takes into account both its bias and precision. If the MAP estimates of our model are indeed unbiased then the previous comparisons based on the variance are equivalent to MSE-based comparisons. The problem is that in our model, the conditional posterior distribution of the coefficients is very complex and this is the reason we use the M-H step in order

to update them in the MCMC algorithm. Therefore we cannot analytically study its properties.

For this reason, in order to further compare the estimates from the case-control and the kin-cohort designs we introduce the *Observed Mean Squared Error* (OMSE) which quantifies the properties of an estimate in the same way MSE quantifies those of an estimator. Specifically,

$$OMSE(\hat{\beta}_{obs}) = (\hat{\beta}_{obs} - \beta)^2 + \hat{\text{Var}}(\hat{\beta}_{obs}), \tag{4.16}$$

where $\hat{\beta}_{obs}$ is the observed value of the MAP estimator $\hat{\beta}$. The variance is calculated, as before, using the posterior samples from the MCMC algorithm.

Table 4.3: OMSE of $\hat{\beta}_{obs}$ under the case-control (C-C) and the kin-cohort (K-C) designs. The last column are the ratios of the two OMSEs for the same parameter values. A ratio greater than 1 favours the kin-cohort design.

δ	θ	K	OMSE($\hat{\beta}_{obs}^{(C-C)}$) ($\times 10^{-3}$)	OMSE($\hat{\beta}_{obs}^{(K-C)}$) ($\times 10^{-3}$)	Ratio $\frac{\text{OMSE}_{(\hat{\beta}_{obs}^{(C-C)})}}{\text{OMSE}_{(\hat{\beta}_{obs}^{(K-C)})}}$
0.01	0.05	0.01	7.1	6.0	1.18
		0.05	4.5	4.9	0.92
	0.25	0.01	7.6	7.3	1.04
		0.05	4.3	4.3	1.00
0.05	0.05	0.01	18.4	7.5	2.45
		0.05	4.0	2.6	1.53
	0.25	0.01	10.1	6.7	1.50
		0.05	2.9	2.8	1.03
0.10	0.05	0.01	11.0	9.8	1.12
		0.05	4.0	6.1	0.66
	0.25	0.01	18.0	6.3	2.86
		0.05	2.6	3.6	0.69

The values of the OMSE from the applications of our method on the simulated data are presented in Table 4.3. The majority of the OMSE sup-

port the use of the kin-cohort over the case-control design. This is especially evident in the datasets that were simulated from a population with small disease prevalence ($K = 0.01$). In fact, in those occasions the support for the kin-cohort design is stronger than the one in the precision-based comparisons. However, using OMSE as a measure to compare the two designs can, sometimes, suggest that the case-control is the preferable design. This happens when we deal with less rare diseases ($K = 0.05$) and it is more profound for large values of the disease penetrance ($\delta = 0.1$). Since the precision of the estimators is always better under the kin-cohort design, this fact raises some concerns regarding the unbiasedness of the MAP estimates under that design. Nonetheless, the values of the observed bias (the error $\hat{\beta}_{obs} - \beta$) of the estimates under the two designs do not show any pattern or systematically favour the one or the other design (Table 4.4).

Overall, based on our current results, we cannot reject our initial assumption that the MAP estimators of β are unbiased, at least for the time being. However the OMSE results raised some concerns which we feel that should be included in this thesis. We address this issue in the subsequent discussion of this chapter (section 4.4) and in the overall discussion, including future work, at the end of this thesis (chapter 6).

4.4 Discussion

We developed a hierarchical bayesian model for genome-wide association analysis under the kin-cohort study design. We make use of a weighted

Table 4.4: Observed bias of $\hat{\beta}_{obs}$ under the case-control (C-C) and the kin-cohort (K-C) designs.

δ	θ	K	Obs.bias($\hat{\beta}_{obs}^{(C-C)}$)	Obs.bias($\hat{\beta}_{obs}^{(K-C)}$)
0.01	0.05	0.01	-0.029	-0.016
		0.05	-0.050	-0.057
	0.25	0.01	-0.043	-0.050
		0.05	0.044	0.049
0.05	0.05	0.01	-0.096	-0.030
		0.05	-0.042	-0.028
	0.25	0.01	-0.030	0.014
		0.05	0.013	0.029
0.10	0.05	0.01	0.026	0.068
		0.05	0.038	0.065
	0.25	0.01	-0.063	0.027
		0.05	0.013	0.043

likelihood in order to correct the bias in the posterior estimates of the effect size. The analysis is carried out using MCMC. We applied our method on simulated data and compare the results with the ones obtain using a hierarchical bayesian model under the ordinary case-control design. The results show that the precision of the posterior effect size estimate increases when we adopt the the kin-cohort design than using the case-control design. This is significant, considering that all we need to do in order to obtain suitable data for a kin-cohort study, is to ask the participants to fill a questionnaire regarding the disease status of their relatives. Thus the transition from the case-control to the kin-cohort design comes at no extra cost when new individuals are going to be genotyped for the study. However when data are already available, and the necessary information was not recorded then it might not be easy or it might be costly to go back and do the interviews. In this case the researcher will have do decide between the benefits of the

kin-cohort design and the added costs needed to implement it.

Concerning the relative increase in precision of the posterior estimate of the effect size when we use the kin-cohort design, compared to the case-control design the only definite conclusions we can draw from our study is that there is one. However there are indications that for rare diseases, this relative precision increases as the effect size increases while for relatively more common diseases the relative precision is roughly constant for different values of the effect size. We provide a possible explanation for this. First we should note that in our simulations we do not increase the effect size parameter (β) directly, but a function of the effect size, δ, which we defined as the proportion of the total disease prevalence that is caused by the marker we study. This way we used a standardised measure of the effect size and thus we can use it to compare results from data that were obtained from data generated using different values for the disease prevalence. Also the use of weights, essentially renders the correlated familial data independent albeit, now, not all individuals carry equal importance in the sample. The estimates become more precise when more of the imputed genotypes are the same as the true unknown ones, especially the imputed genotypes of affected individuals. Now, for a common disease the variance of the distributions in (4.11) and (4.12) is larger than when we deal a rare disease. This can naturally lead to many wrongly imputed genotypes. However since the data are independent there will be many instances where two wrongs will make a right (e.g. Two affected individuals A and B, with true genotypes $x_A = 1$ and $x_B = 2$ and imputed genotypes $x_A = 2$ and $x_B = 1$). Thus we may have more mistaken

genotypes for smaller effect sizes but because they offset the overall accuracy of the imputation remains constant. On the other hand, imputation mistakes in the case of a rare disease will be much more costly, especially when they concern affected individuals. Those mistakes happen more often when he effect size is small while they happen rarely for larger effect size. This is why for a small effect size the relative precision is better when we have a common disease compared to a rare disease while the opposite happens for greater effect sizes.

As a last note, we want to address the issue of potential bias for the estimators of the effect size we use in our model. Even though the relevance weights were specifically introduced to deal with this issue - and indeed they vastly improved the estimates compared to the same models lacking those weights - we nonetheless proceeded and did an additional analysis without the assumption of unbiasedness. In order to compare the effect size estimators under the two designs we introduced the Observed Mean Squared Error which is basically a measure that assesses estimates in the same way that the Mean Squared Error assesses estimators. Despite that the majority of those comparisons still favour the kin-cohort design, our analysis raised some concerns regarding potential bias in the estimates under that design. Conclusive arguments regarding this issue cannot be made at this point since the analysis of bias was based on a relatively small number of estimates - essentially random variables - and any deviation from the true value can be attributed to statistical error. In chapter 6 we discuss the problems that hinder a more comprehensive analysis of the bias with our current framework and suggest

a potential solution as a part of an alternate approach for the analysis of the same type of kin-cohort data we consider in this chapter.

Chapter 5

The Super-control – Control – Case – Super-case Design

In the Super-control – Control – Case – Super-case (SCCS) design we have samples from four different but not mutually exclusive populations (fig 5.1). In reality super-controls, controls and cases share the same definition as the controls, cohort and cases respectively, that we described in Chapter 3. The reason we do not study the case-cohort-control design as a part of the SCCS design is the fact that the methods we present in this chapter, heavily rely on the presence of super-cases. In order to explain this we need to provide the definition of the four groups of our design.

Controls: These are a random sample from the entire population. It is the same sample group we call cohort in Chapter 3. The reason we use the term control here, is mostly to emphasise the symmetry of the SCCS design but also because it is quite common in GWAS that this kind

of data are used and called controls. Normally there is a number of affected individuals in the sample. The expected proportion of affected individuals is equal to the disease prevalence (K). The advantages of using this type of controls in GWAS, instead of actual unaffected individuals, is that we can obtain a larger amount of data since no diagnosis for disease is required. Furthermore the same sample can be used for more than one study concerning different diseases.

Super-controls: These are a random sample from the unaffected sub-population. Sometimes they are required to satisfy additional criteria - e.g. lack of disease in the immediate family - but more often than not we simply require the sample to come from the disease-free population. This is the definition we adopt. Also, we should note that this is the same type of sample we call simply controls in chapter 3.

Cases: These are a random sample from the affected sub-population.

Super-cases: Unlike the other three groups which have more or less a rigid definition, the definition of super-cases is vague and depends on the data available in the study. They are always sampled from a subset of the affected sub-population, which we call the super-affected sub-population. Super-affected individuals are, in most cases, required to have some sort of disease history in their families. However the exact requirements vary from study to study. Some may require a number of diseased first degree relatives, others may extend that to second degree relatives or even take into account the ancestral history. In our study we consider a simple example of super-cases which we discuss in detail.

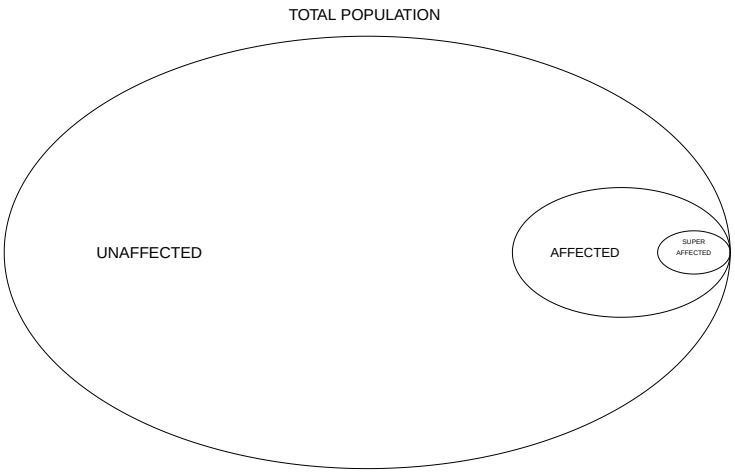


Figure 5.1: Graphical representation of the populations and the sub-populations from which the four sample groups are obtained. Controls from the entire population, super-controls from the unaffected, cases from the affected and super-cases from the super-affected sub-populations respectively.

5.1 The Data in our Study

In our study we use simulated data. Super-controls, controls and cases are straightforward to simulate and we already did that in Chapter 3. In order to simulate the super-cases we need to be more specific with their definition. For this reason we consider a simple example where our data consist of pairs of siblings. For each pair one of the siblings is the proband, that is, it is considered a potential participant in the study if it is selected in the sample, whereas we are only interested in the phenotype of the second sibling when

the proband is affected which determines whether the latter is classified as simply affected or super-affected.

More formally, we consider pairs $\{(Y_i^{(1)}, x_i^{(1)}), (Y_i^{(2)}, x_i^{(2)})\}$ where $Y \in \{0, 1\}$ denotes the phenotype and $x \in \{0, 1, 2\}$ the genotype. Without loss of generality we consider the first sibling to be the proband. Then, the sub-populations of the unaffected (S_0), affected (S_2), super-affected (S_3) as well as the entire population (S_1) are defined as

$$\begin{aligned} S_0 &= \{(Y_i^{(1)}, x_i^{(1)}) : Y_i^{(1)} = 0\} \\ S_1 &= \{(Y_i^{(1)}, x_i^{(1)}) : i \in \{1, \ldots, M\}\} \\ S_2 &= \{(Y_i^{(1)}, x_i^{(1)}) : Y_i^{(1)} = 1\} \\ S_3 &= \{(Y_i^{(1)}, x_i^{(1)}) : Y_i^{(1)} = Y_i^{(2)} = 1\}. \end{aligned} \tag{5.1}$$

In (5.1) M denotes the population size. However this is not the number of individuals in the population but rather the number of individuals that can be potentially included in our sample. In non-family-based association studies we try to avoid including correlated individuals in our sample. We describe the issues that arise from this in Chapter 4. Thus we consider that we use a cluster two-stage sampling scheme to obtain our sample. The clusters in our cases are the pairs while in more general examples might be the families in the population. Thus in the first stage we sample clusters and in the second stage we sample one individual within the selected clusters. So, the population size M in our example is the total number of clusters, i.e. pairs of siblings, in the population.

The sample sizes we use for the simulated data in all the applications of the methods we describe in this chapter are 1000 for the super-cases and super-controls and 2000 for the cases and controls.

5.1.1 Conditional Minor Allele Frequencies

In (2.3) we define the conditional genotypic probabilities p_x and q_x. The same formulas apply here for genotypic probabilities of the affected and unaffected respectively. Also g_x still denotes the unconditional genotypic probabilities in the population. Note that we also assume HWE so g_x can be expressed as a function of the population MAF. Thus $g_x = P(x|S_1) \equiv P(x^{(1)})$ is given by (2.2) while $q_x = P(x|S_0) \equiv P(x^{(1)}|Y^{(1)} = 0)$ and $p_x = P(x|S_2) \equiv P(x^{(1)}|Y^{(1)} = 1)$ are given by (2.3). In addition to those we define the conditional genotypic probabilities of the super-affected, $r_x = P(x|S_3) \equiv P(x^{(1)}|Y^{(1)} = 1 \cap Y^{(2)} = 1)$.

Expressing r_x in a closed form is not as trivial as it is for p_x and q_x. Also r_x may vary within the same study, for example when we define the super affected based on a number of affected first degree relatives and we have differend family sizes, a situation that is almost certain in real applications. In the applications we describe there is no need to have a closed-form expression of r_x and thus they can be used for studies where the definition of super-affected differs from the one we use. However for our siblings example we are able to calculate a closed form of r_x (Appendix A). This is mostly used in simulations and assesment of methods rather than the methods them-

selves.

$$r_x \equiv P(x|S_3) = \frac{f_x P(x^{(1)} = x|y^{(2)} = 1)}{\sum_{x=0,1,2} f_x P(x^{(1)} = x|y^{(2)} = 1)} \tag{5.2}$$

The genotypic probability conditional on the sibling's genotype is given by,

$$P(x^{(1)} = x|y^{(2)} = 1) = \frac{1}{4}g_x + \mathcal{I}_{[x=0]}(1-\theta_1)\left(p_0 + \frac{p_1}{2}\right) + \mathcal{I}_{[x=1]}\left(\theta_1 p_0 + \frac{p_1}{2} + (1-\theta_1)p_2\right) + \mathcal{I}_{[x=2]}\theta_1\left(\frac{p_1}{2} + p_2\right) + \frac{1}{4}p_x \tag{5.3}$$

In (5.3) θ_1 is the population MAF. We use this notation in order to comform with the notation we used for the sub-populations $S_i, i = 1, \ldots, 4$.

$$\theta_i = \frac{E(x|S_i)}{2} = \frac{1}{2}P(x = 1|S_i) + P(x = 2|S_i) \quad , \quad i = 1, 2, 3, 4 \tag{5.4}$$

Equivalent adjustments as the ones we describe in Chapter 3 can be made when we assume a dominant or recessive effect.

We already showed that the value of the population MAF lies between the values of the conditional MAFs of the affected and unaffected subpopulations, $\theta_0 \lesseqgtr \theta_1 \lesseqgtr \theta_2$. The direction is depends on whether the rare allele is causal ($\geq$) or protective ($\leq$). The methods we develop in this chapter are based on an extension of this property where now we include the conditional MAF of the super-affected sub-population. The conditional minor allele frequency of the super-affected is greater than the conditional minor allele frequency of the affected sub-population when the rare allele is causal while is smaller when the rare allele is protective (Appendix A). Intuitively this means that a causal allele is observed more frequently within families with the disease in

their family history rather than families with singular cases. So overall we have,

$$\begin{aligned} &\theta_0 \leq \theta_1 \leq \theta_2 \leq \theta_3 \quad , \quad \text{if the rare allele is causal and} \\ &\theta_0 \geq \theta_1 \geq \theta_2 \geq \theta_3 \quad , \quad \text{if the rare allele is protective.} \end{aligned} \tag{5.5}$$

5.2 Bayesian Models and Importance Sampling

Let $n_i, i = 0, 1, 2, 3$ denote the respective sample sizes of super-controls, controls, cases and super-cases and $N_i = 2n_i$ the number of total alleles in these samples. Also let $x_i^j,\ i = 0, 1, 2, 3,\ j = 1, n_i$ be the genotypes of the individuals in those samples and $X_i = \sum_{j=1}^{n_i} x_i^j$. Assuming that the haplotypes in each sample follow a binomial distribution with probability of success the conditional MAF of the population that the sample comes from, the likelihood of the data is simply the product of the four binomial likelihoods,

$$L(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{N}) = P(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{N}) = \prod_{i=0}^{3} P(X_i|\theta_i, N_i), \tag{5.6}$$

where, $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3)$, $\boldsymbol{X} = (X_0, X_1, X_2, X_3)$, $\boldsymbol{N} = (N_0, N_1, N_2, N_3)$ and

$$P(X_i|\theta_i, N_i) = \binom{N_i}{X_i} \theta_i^{X_i} (1 - \theta_i)^{N_i - X_i} \tag{5.7}$$

We should mention again that even when we assume HWE the genotypes of the sub-populations S_0, S_2 and S_3 do not follow a binomial distribution but a trinomial. For the entire population the trinomial distribution of the genotypes with parameters (g_0, g_1, g_2) is equivalent to the binomial distribution of the haplotypes with parameter θ_1. In (5.7) we make the assumption that the same holds for the sub-populations which is not accurate. However in most cases the binomial distributions we use approximate the respective trinomials fairly accurately.

Using the properties of the minor allele frequencies in (5.5) we develop two Bayesian models for statistical inference in genetic association. The inference is based on Bayes Factors which we calculate using importance sampling. The advantages of this Bayesian approach compared with other, mostly frequentist approaches, is that we can select our test hypotheses so that they cover only the valid parameter space of $\boldsymbol{\theta}$. In a Bayesian framework it is relatively easy to adopt a model that its posterior inference will help us decide between

$$\begin{aligned} H_0 &: \theta_0 = \theta_1 = \theta_2 = \theta_3 \\ H_1 &: \theta_0 < \theta_1 < \theta_2 < \theta_3 \quad \text{or} \quad \theta_0 > \theta_1 > \theta_2 > \theta_3. \end{aligned} \tag{5.8}$$

Any other arrangement of the minor allele frequencies is, at least theoretically impossible. In contrast the competing hypotheses in any frequentist test for our problem cover the entire $[0,1]^4$ space, usually by setting $H_1 : \theta_0 \neq \theta_1 \neq \theta_2 \neq \theta_3$. Thus by using Bayesian methods we are able to restrict the parameter space and consider only the possible values of the parameter $\boldsymbol{\theta}$.

In addition to that we can use any prior knowledge concerning whether the rare allele is causal or protective and adjust the alternative hypothesis accordingly. We can "break" the alternative hypothesis into the causal (H_1^c : $\theta_0 < \theta_1 < \theta_2 < \theta_3$) and protective rare allele hypotheses (H_1^p : $\theta_0 > \theta_1 > \theta_2 > \theta_3$) and assign each with a prior probability, $P(H_1^c|H_1)$ and $P(H_1^p|H_1)$. These probabilities can be used as weights in the prior - and consequently in the posterior - distribution of $\boldsymbol{\theta}$ since those distributions should be able to take into account both potential causal and protective effects.

The two models we develop differ on their prior distributions. Below we describe those models as well as the methods we use to calculate the Bayes Factors.

5.2.1 Truncated-Beta Prior distributions

If $0 \leq l \leq u \leq 1$ we denote with $\text{TrBeta}(a, b, l, u)$ the truncated beta distribution (Nadarajah and Kotz, 2006) that has zero density outside the interval (l, u).

$$X \sim \text{TrBeta}(a, b, l, u) \Leftrightarrow X \sim c(a, b, l, u)\text{Beta}(a, b)\mathcal{I}_{[l \leq X \leq u]} \quad , \tag{5.9}$$

where,

$$c(a, b, l, u) = \left[\int_l^u \frac{X^{a-1}(1-X)^{b-1}}{B(a, b)} dX\right]^{-1} , \tag{5.10}$$

and B is the beta function.

Under the alternative hypothesis we consider two sets of prior distribu-

tions. One under the assumption that the rare allele is causal (H_1^c) and the other under the assumption that it is protecive (H_1^p).

$$\begin{aligned} \pi(\boldsymbol{\theta}|H_1^c) &= \pi(\theta_0|\theta_1; H_1^c)\pi(\theta_1|H_1^c)\pi(\theta_2|\theta_1; H_1^c)\pi(\theta_3|\theta_2; H_1^c) \\ \pi(\boldsymbol{\theta}|H_1^p) &= \pi(\theta_0|\theta_1; H_1^p)\pi(\theta_1|H_1^p)\pi(\theta_2|\theta_1; H_1^p)\pi(\theta_3|\theta_2; H_1^p) \end{aligned} \tag{5.11}$$

The marginal probabilities in (5.11) are given by,

$$\begin{aligned} &\pi(\theta_0|\theta_1; H_1^c) = \text{TrBeta}(a_0, b_0, 0, \theta_1) && \pi(\theta_0|\theta_1; H_1^p) = \text{TrBeta}(a_0, b_0, \theta_1, 1) \\ &\pi(\theta_1|H_1^c) = \text{Beta}(a_1, b_1) && \pi(\theta_2|H_1^p) = \text{Beta}(a_1, b_1) \\ &\pi(\theta_2|\theta_1; H_1^c) = \text{TrBeta}(a_2, b_2, \theta_1, 1) && \pi(\theta_2|\theta_1; H_1^p) = \text{TrBeta}(a_2, b_2, 0, \theta_1) \\ &\pi(\theta_3|\theta_2; H_1^c) = \text{TrBeta}(a_3, b_3, \theta_2, 1) && \pi(\theta_3|\theta_2; H_1^p) = \text{TrBeta}(a_3, b_3, 0, \theta_2). \end{aligned} \tag{5.12}$$

The overall prior distribution under H_1 is taken as a mixture of the two distributions in (5.11). The choice of weights is based on our prior beliefs on whether the rare allele is causal or protective.

$$\pi(\boldsymbol{\theta}|H_1) = \pi(\boldsymbol{\theta}|H_1^c)P(H_1^c|H_1) + \pi(\boldsymbol{\theta}|H_1^p)P(H_1^p|H_1) \tag{5.13}$$

When we choose to adopt non-informative prior distributions we set $P(H_1^c|H_1) = P(H_1^p|H_1) = 1/2$ and $a_i = b_i = 1$, $i = 0, 1, 2, 3$, which is equivalent to choosing uniform distributions. In this case the priors have a much simpler closed

form [1].

$$\begin{aligned} \pi(\boldsymbol{\theta}|H_1^c) &= \frac{1}{\theta_1}\frac{1}{1-\theta_1}\frac{1}{1-\theta_2}\mathcal{I}_{[\theta_0<\theta_1<\theta_2<\theta_3]} \\ \pi(\boldsymbol{\theta}|H_1^p) &= \frac{1}{1-\theta_1}\frac{1}{\theta_1}\frac{1}{\theta_2}\mathcal{I}_{[\theta_0>\theta_1>\theta_2>\theta_3]} \end{aligned} \tag{5.14}$$

The assumption under the null hypothesis of no association is that all the genotypes were generated from the same distribution and thus there is only one "global" MAF.

$$H_0 : \theta_0 = \theta_1 = \theta_2 = \theta_3 \equiv \theta^* \tag{5.15}$$

We adopt a beta prior distribution for θ^*.

$$\pi(\theta^*|H_0) = \text{Beta}(a^*, b^*) \tag{5.16}$$

Ideally, the parameters in (5.16) should be chosen such that the distribution has the same expectation as the prior distribution of θ_1 in (5.12) although we can choose the prior of θ^* to be more vague. A non-informative prior in this case is simply a $U[0, 1]$ distribution.

[1] Further justification in Appendix A

5.2.1.1 Bayes Factors and Importance Sampling

In order to calculate the the Bayes Factor of our model we need the marginal likelihood of the data $\boldsymbol{X}$ under the two hypotheses (Kass and Raftery, 1995).

$$\begin{aligned} m(\boldsymbol{X}|\boldsymbol{N}; H_1) &= \int_{\boldsymbol{\theta}} P(\boldsymbol{X}|\boldsymbol{N}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|H_1) \\ m(\boldsymbol{X}|\boldsymbol{N}; H_0) &= \int_{\theta^*} P(\boldsymbol{X}|\boldsymbol{N}, \theta^*\mathbf{1}_4)\pi(\theta^*|H_0) \end{aligned} \tag{5.17}$$

The likelihood $P(\boldsymbol{X}|\boldsymbol{N}, \boldsymbol{\theta})$ is given in (5.6) and (5.7) while $\mathbf{1}_4$ is the vector $(1, 1, 1, 1)$. The marginal likelihood under H_0 can easily be calculated analytically.

$$m(\boldsymbol{X}|\boldsymbol{N}; H_0) = \frac{B\left(\sum_{i=0}^{3} X_i + a^*, \sum_{i=0}^{3}(N_i - X_i) + b^*\right)}{B(a^*, b^*)} \prod_{i=0}^{3} \binom{N_i}{X_i} \tag{5.18}$$

On the other hand the integral that gives the marginal likelihood under the alternative is extremely difficult, if not impossible, to calculate analytically. Instead we aalculate numerically the equivalent integral

$$m(\boldsymbol{X}|\boldsymbol{N}; H_1) = \int_{\boldsymbol{\theta}} \frac{P(\boldsymbol{X}|\boldsymbol{N}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|H_1)}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{5.19}$$

using importance sampling (Robert and Marin, 2010). That is we draw samples $\boldsymbol{\theta}^{(i)}, i = 1, \ldots, m$ from an appropriate distribution $g(\boldsymbol{\theta})$ and estimate the marginal likelihood using Monte-Carlo integration.

$$\hat{m}(\boldsymbol{X}|\boldsymbol{N}; H_1) = \frac{1}{m} \sum_{i=1}^{m} \frac{P(\boldsymbol{X}|\boldsymbol{N}, \boldsymbol{\theta}^{(i)})\pi(\boldsymbol{\theta}^{(i)}|H_1)}{g(\boldsymbol{\theta}^{(i)})} \equiv \frac{1}{m} \sum_{i=1}^{m} \lambda(\boldsymbol{\theta}^{(i)}) \tag{5.20}$$

We choose $g(\boldsymbol{\theta})$ to be the unconstrained posterior distribution of $\boldsymbol{\theta}$. The unconstrained prior distribution is

$$\begin{aligned} \pi_u(\boldsymbol{\theta}) &= \prod_{i=0}^{3} \pi_{u;i}(\theta_i), \quad \text{where,} \\ \pi_{u;i}(\theta_i) &= \text{Beta}(a_i, b_i) \end{aligned} \tag{5.21}$$

Then the joint unconstrained posterior is just the product of the marginal posterior distributions.

$$g(\boldsymbol{\theta}) \equiv \pi_u(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{N}) \stackrel{\mathcal{D}}{=} \prod_{i=0}^{3} \text{Beta}(X_i + a_i, N_i - X_i + b_i) \tag{5.22}$$

Thus we can draw samples from g using the marginal Beta distributions; that is draw $\theta_j^{(i)}$ from the Beta$(X_j + a_j, N_j - X_j + b_j)$ distribution and then set $\boldsymbol{\theta}^{(i)} = (\theta_0^{(i)}, \theta_1^{(i)}, \theta_2^{(i)}, \theta_3^{(i)})$. To make our subsequent calculations easier we can also write $g(\boldsymbol{\theta})$ in the form,

$$g(\boldsymbol{\theta}) = \frac{B_\pi}{B_P Q} P(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{N}) \pi_u(\boldsymbol{\theta}) \tag{5.23}$$

where,

$$\begin{aligned} B_\pi &= \prod_{i=0}^{3} B(a_i, b_i) \\ B_P &= \prod_{i=0}^{3} B(X_i + a_i, N_i - X_i + b_i) \\ Q &= \prod_{i=0}^{3} \binom{N_i}{X_i} \end{aligned} \tag{5.24}$$

Thus $\lambda(\boldsymbol{\theta}^{(i)})$ becomes,

$$\lambda(\boldsymbol{\theta}^{(i)}) = \frac{B_P Q}{B_\pi} \frac{\pi(\boldsymbol{\theta}^{(i)}|H_1)}{\pi_u(\boldsymbol{\theta}^{(i)})}. \tag{5.25}$$

Much simpler, when we choose non-informative prior distributions we have,

$$\lambda(\boldsymbol{\theta}^{(i)}) = \begin{cases} \frac{B_P Q}{2} \frac{1}{\theta_1^{(i)}} \frac{1}{1-\theta_1^{(i)}} \frac{1}{1-\theta_2^{(i)}} & , \quad \theta_0^{(i)} < \theta_1^{(i)} < \theta_2^{(i)} < \theta_3^{(i)} \\ \frac{B_P Q}{2} \frac{1}{1-\theta_1^{(i)}} \frac{1}{\theta_1^{(i)}} \frac{1}{\theta_2^{(i)}} & , \quad \theta_0^{(i)} > \theta_1^{(i)} > \theta_2^{(i)} > \theta_3^{(i)} \\ 0 & , \quad \text{otherwise} \end{cases} \tag{5.26}$$

Finally we assume that the prior probabilities of the two hypotheses are equal, $P(H_0) = P(H_1) = 1/2$ (prior odds are equal to 1), an assumption that is common in Bayesian decision theory if there is lack of strong evidence that supports either hypothesis. In this case the posterior odds are equal to the Bayes Factor which is calculated as the ratio of the marginal likelihoods.

$$\mathrm{BF}(\boldsymbol{X}|\boldsymbol{N}) = \frac{\hat{m}(\boldsymbol{X}|\boldsymbol{N}; H_1)}{m(\boldsymbol{X}|\boldsymbol{N}; H_0)} \tag{5.27}$$

Values significantly larger than one (1) favour the alternative hypothesis while values close to zero (0) provide strong evidence towards the null hypothesis.

5.2.2 Joint Constrained-Uniform Prior Distribution

Let $Z_i \in [0, 1]$, $i = 1, \ldots, M$ be random variables and σ any given permutation of $\{1, \ldots, M\}$. Without loss of generality we set $\sigma(i) = i$. We define the

joint constrained-unifrom distribution of $\boldsymbol{Z} = (Z_1, \ldots Z_M)$ to be,

$$f(\boldsymbol{z}) = c \times \mathcal{I}_{[z_1<z_2<\ldots<z_M]} = \begin{cases} c & , \quad \text{if} \quad z_1 < z_2 < \ldots < z_M \\ 0 & , \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{5.28}$$

where $\boldsymbol{z} = (z_1, \ldots z_M)$ is a realisation of $\boldsymbol{Z}$. In other words the variables Z_i have a joint uniform distribution inside the sub-space of $[0,1]^M$ that is defined by the constraint $Z_1 < Z_2 < \ldots < Z_M$ while outside this sub-space the density is zero. The constant c like any normalizing constant of any distribution, specifically in this case we want the integral of f in $[0,1]^M$ to be equal to 1.

$$c = \left[\int_0^1 \int_0^{Z_M} \ldots \int_0^{Z_2} dZ_1 \ldots dZ_{M-1} dZ_M\right]^{-1} = \left[\int_0^1 \int_{Z_1}^1 \ldots \int_{Z_{M-1}}^1 dZ_M dZ_{M-1} \ldots dZ_1\right]^{-1} \tag{5.29}$$

Using mathematical induction we can show (Appendix A) that $c = M!$.

Now, for our model we adopt this constraint-uniform distribution as prior under the alternative hypothesis. Similar to the previous model we consider the two sub-hypotheses H_1^c and H_1^p.

$$\begin{aligned} \pi(\boldsymbol{\theta}|H_1^c) &= 24 \times \mathcal{I}_{[\theta_0<\theta_1<\theta_2<\theta_3]} \\ \pi(\boldsymbol{\theta}|H_1^p) &= 24 \times \mathcal{I}_{[\theta_0>\theta_1>\theta_2>\theta_3]} \\ \pi(\boldsymbol{\theta}|H_1) &= P(H_1^c|H_1)\pi(\boldsymbol{\theta}|H_1^c) + P(H_1^p|H_1)\pi(\boldsymbol{\theta}|H_1^p) \end{aligned} \tag{5.30}$$

Under the null hypothesis we choose a uniform prior distribution on θ^*.

$$\pi(\theta^*|H_0) = 1 \tag{5.31}$$

Thus the posterior marginal likelihood under H_0 is

$$m(\boldsymbol{X}|\boldsymbol{N}; H_0) = B\left(\sum_{i=0}^{3} X_i + 1, \sum_{i=0}^{3}(N_i - X_i) + 1\right) \prod_{i=0}^{3} \binom{N_i}{X_i}. \tag{5.32}$$

For the calculation of the marginal likelihood under H_1 we, once more, resolve to importance sampling. The unconstrained prior distribution is uniform in the entire $[0,1]^4$ space. It can also be written as the product of the marginal unconstrained posterior distributions, $\pi_{u;i}(\theta_i)$, $i = 0, 1, 2, 3$, which are all $U(0,1)$.

$$\pi_u(\boldsymbol{\theta}) = \prod_{i=0}^{3} \pi_{u;i}(\theta_i) = 1 \tag{5.33}$$

Thus, in the importance sampling algorithm each $\theta_j^{(i)}$ is drawn from the $U(0,1)$ distribution. Therefore we can obtain an estimate of the posterior marginal likelihood under H_1,

$$\hat{m}(\boldsymbol{X}|\boldsymbol{N}; H_1) = \frac{1}{m} \sum_{i=0}^{3} \lambda(\boldsymbol{\theta}^{(i)}) \tag{5.34}$$

where,

$$\lambda(\theta^{(i)}) = \begin{cases} 12\frac{B_P Q}{B_\pi}, & \text{if} \quad \theta_0^{(i)} < \theta_1^{(i)} < \theta_2^{(i)} < \theta_3^{(i)} \\ & \text{or} \quad \theta_0^{(i)} > \theta_1^{(i)} > \theta_2^{(i)} > \theta_3^{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{5.35}$$

The quantities B_π, B_P and Q in (5.35) are the same as the ones we defined

in (5.24), only this time we have $a_i = b_i = 1$ for $i = 0, 1, 2, 3$. Hence we use $m(\boldsymbol{X}|\boldsymbol{N}; H_0)$ and $\hat{m}(\boldsymbol{X}|\boldsymbol{N}; H_1)$ according to (5.27) and calculate the Bayes Factor for our decision problem.

5.3 An Allelic Trend Test

In this section we develop an allelic trend test. The rationale is the same as in the Cochran-Armitage test and the derivation of our test follows similar steps. However unlike CA test where the inference is based on the genotypic risks $P(Y = 1|x = i) \equiv E(Y|x = i)$, $i = 0, 1, 2$ the inference in our test is based on the conditional allele frequencies $E(x|S_i)$ or equivalently on θ_i, $i = 0, 1, 2, 3$. We should also note that this test does not take into account the order of the conditional MAFs. The test hypotheses are

$$\begin{aligned} H_0 &: \theta_0 = \theta_1 = \theta_2 = \theta_3 \\ H_1 &: \theta_i \neq \theta_j \quad \text{for at least one pair} \quad (i, j), \quad i, j = 0, 1, 2, 3 \end{aligned} \tag{5.36}$$

We consider the distribution of the alleles in the four sub-populations (Table 5.3). Similar to the CA test which basically modifies the Pearson's chi-square test, we want to construct a test statistic that is also approximately chi-square distributed under the null hypothesis. Using the same tools as in the derivation of CA test (Slager and Schaid, 2001), and given scores z_i,

Sample Group Population	Super-control S_0	Control S_1	Case S_2	Super-case S_3	Total
Allele					
0 (Common)	$\tilde{X}_0$	$\tilde{X}_1$	$\tilde{X}_2$	$\tilde{X}_3$	$\tilde{X}_T$
1 (Rare)	X_0	X_1	X_2	X_3	X_T
Total	N_0	N_1	N_2	N_3	N

Table 5.1: Allelic distribution for the four sub-populations

$i = 0, 1, 2, 3$, we set

$$U = \sum_{i=0}^{3} z_i \left(\frac{X_T}{N} \tilde{X}_i - \frac{\tilde{X}_T}{N} X_i \right) \tag{5.37}$$

Under the null hypopothesis of no association the expected value of U is zero and an estimate of its variance is

$$\hat{Var}(U) = \frac{X_T \tilde{X}_T}{N^2} \left(\sum_{i=0}^{3} z_i^2 N_i - \frac{1}{N} \sum_{i=0}^{3} (z_i N_i)^2 \right) \tag{5.38}$$

Based on the central limit theorem the asymptotic distribution of $U/\sqrt{U}$ is standard normal. Thus we define our allelic trend test statistic

$$X^2_{all} = \frac{U^2}{\hat{Var}(U)} \overset{H_0}{\sim} \chi^2_1. \tag{5.39}$$

However although the derivation and the resulting statistic of the allelic trend test can be seen as a generalisation of the CA test for a 2×4 table there is a major difference that may potentially cause complications in the asymptotic distribution (5.39). In the Cochran-Armitage test the marginal row totals of the data (R and S in the notation we used in Table 3.1) are fixed and the marginal column totals are random. This is also the case when we

apply the traditional Pearson's chi-square test to test for diffrences between populations. The data are in the form of a contingency table where the marginal row totals are the sizes of the different population samples. The opposite holds for the table we use in the allelic trend test. The marginal row totals $(X_T, \tilde{X}_T)$ are random while the marginal column totals are fixed $(N_i, i = 0, 1, 2, 3)$.

In order to make sure that this does not cause any problems with the asymptotic distribution of X^2_{all} under the null hypothesis, we simulate data under H_0 and apply the test. Some of the resuts are shown in figure 5.2. Other than the justifiable Monte-Carlo error in the extreme values of the distribution, there is no evidence to suggest against the use of the asymptotic χ^2_1 distribution.

5.4 Applications and Comparison of the Bayesian Models and Trend Tests

Tranditionally, comparison and assessement of different frequentist tests is based on their power. On the other hand Bayesian decision theory is usually based on Bayes factors and it is more probabilistic in nature. Although, in theory, the two are not directly comparable, there have been many studies (e.g. Wakefield, 2009) where Bayes factors are compared with p-values, the main tool in frequentist decision theory. There have also been attempts to establish a direct connection between p-values and Bayes factors. For example under general assumptions the following result holds (Stephens and

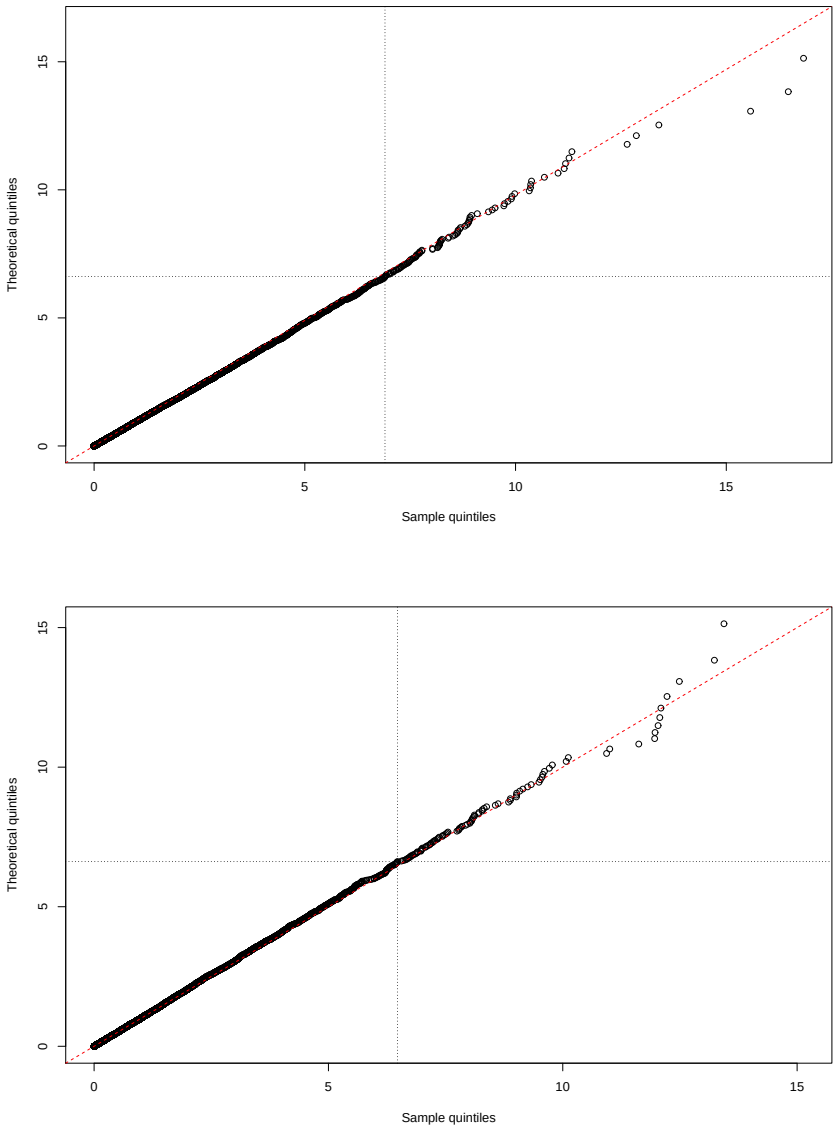


Figure 5.2: QQ-plots that compare the distribution of the allelic trend test with the χ^2_1 distribution. The data were generated using $\theta^* = 0, 2$ (top) and $\theta^* = 0.4$ (bottom). The dotted lines indicate the 0.99 quantiles.

Balding, 2009):

$$\text{If} \quad p < \frac{1}{e} \quad \text{then} \quad BF < -\frac{1}{ep\log(p)} \tag{5.40}$$

In our study take two different approaches for comparing Bayesian and frequentist methods. First, rather simplistically, we treat the Bayes Factor as a test statistic and use its empirical distribution under H_0 in order to calculate the power. The second approach is based on the ranking of markers. We use small-scale genome simulations and assess the performance of the test in identifying the significant markers. We use the following tests:

1. The first bayesian model (BM1) we descibe in section 5.2.1.
2. the second bayesian model (BM2) we describe in section 5.2.2.
3. The allelic trend test (X^2_{all}) we describe in section 5.3.
4. A modified version of the allelic trend test (X^2_{all2}) where we use the joint sample of controls and super-controls. Thus in this test the data are tabulated in a 2×3 contingency table and the test statistic is adjusted accordingly.
5. The Cochran-Armitage test using only the cases and controls (CA_{cc}).
6. The Cochran-Armitage test using only the super-cases and super-controls (CA_{ss}).
7. The Cochran-Armitage test using the joint sample of cases and super-cases and the joint sample of controls and super-controls (CA_{scsc}).
8. The Cochran-Armitage test using the super-cases and the joint sample

of controls and super-controls (CA_{scs1}), i.e. discard the cases.

9. The Cochran-Armitage test using the super-controls and the joint sample of cases and super-cases(CA_{scs2}), i.e. discard the controls.

5.4.1 Single-SNP Power Comparisons

We simulate single-SNP data both prospectively and retrospectively. In prospective simulations we first generate a very large number of parental genotypes (10^7 couples) and then use those to obtain the children's genotypes (pair of siblings). We adopt the logistic disease model and use it to assign the children's disease status ie. the phenotype. One child in each family is chosen to be the proband. Thus based on the phenotypes we can determine which probands are unaffected, affected or super-affected. The cases, super-cases, controls and super-controls are then selected from the affected, super-affected, all and unaffected probands respectively. In retrospective simulations we directly generate the genotypes of the four samples using the probabilities q_x for the super-controls, g_x for the controls, p_x for the cases and r_x for the super-cases. The results we obtain are not different regardless the method we use.

We simulate the data using various different population parameters. The penetrance parameter (β) is chosen so that it takes into account both causal and protective effects. All the frequentist test statistics follow an asymptotic χ^2_1 distribution under the null hypothesis. The scores we use is $\boldsymbol{z} = (0, 1, 2, 3)$ for X^2_{all} and $\boldsymbol{z} = (0, 1, 2)$ for all the other trend tests. In order to be able

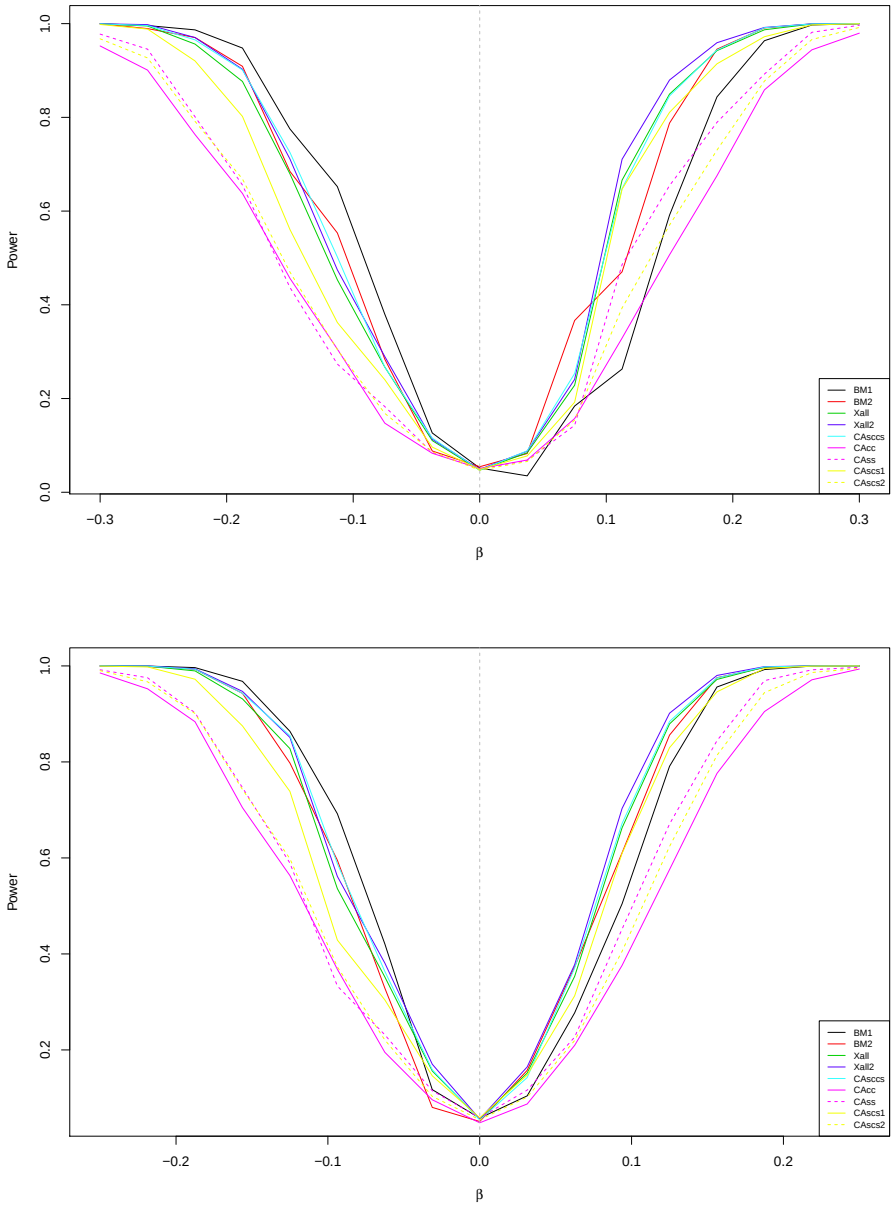


Figure 5.3: Power curves of the two Bayesian and frequentist tests. The data were generated using $\theta_1 = 0.1$ (top) and $\theta_1 = 0.2$ (bottom). Here, we conventionally treat the Bayes factors as if they were test statistics.

to make comparisons we treat the Bayes factors of the two bayesian models as if they were a test statistics. In our power simulations we apply all the

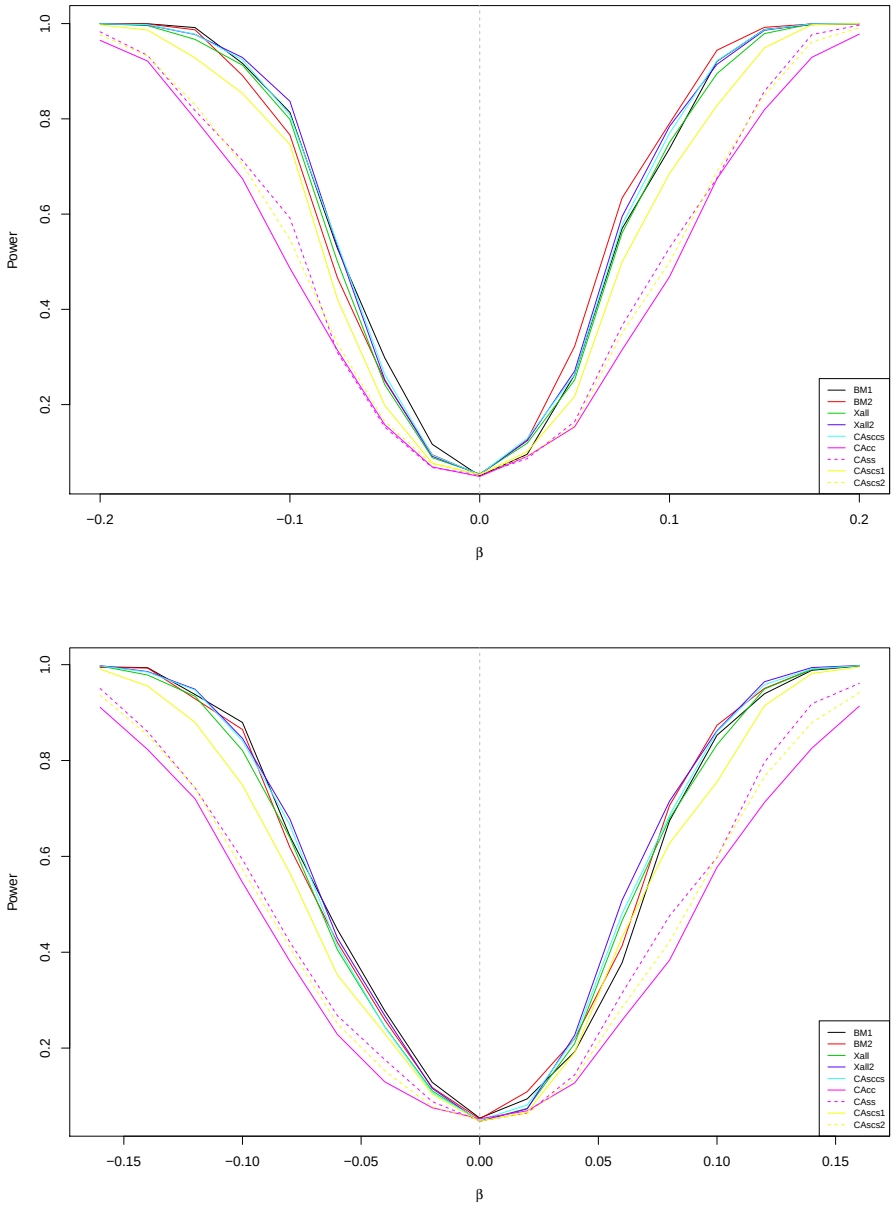


Figure 5.4: Power curves of the Bayesian and frequentist tests. The data were generated using $\theta_1 = 0.3$ (top) and $\theta_1 = 0.4$ (bottom).

tests repeateadly on a large number of datasets that are generated using the same simulation parameters. This means that we have samples of both Bayes

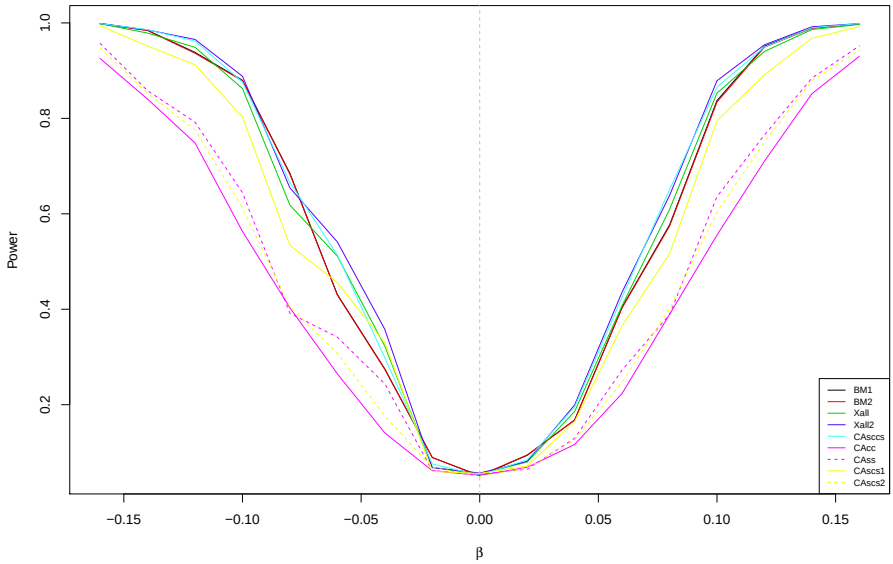


Figure 5.5: Power curves of the Bayesian and frequentist tests. The data were generated using $\theta_1 = 0.5$.

factors that were calculated using data that correspond to the null hypothesis of no association. We use those samples to get an empirical distribution of the BF "test-statistics" under H_0. More precisely we calculate the sample quantile that is equal to the significance level we use to calculate the power of the frequentist tests.

The results (figures 5.3-5.5) show that the two allelic trend tests we use, as well as the Cochran-Armitage test where we use all four samples (CA_{sccs}), are the most powerful among the frequentist tests, with a slight edge in favour of the X^2_{all2} test. The frequentist treatment of the two Bayes Factors also gives relatively powerful tests. Both Bayesian models are often more powerful than any of the frequentist tests. However, their power-curves show some warning signs that this treatment may not be fully reliable. The power-

curves themselves are slightly more rough and not as symmetric with respect to causal and protective effects as the power-curves of the frequentist tests. The latter point can be especially seen when we compare the power-curves of the two Bayesian models. The first model (truncated-beta priors) is more powerful than the second one (joint constrained-uniform prior) when the rare allele has a protective effect. This changes when the rare allele is causal, with the second test being, now, more powerful. Nevertheless, in real life applications it is not very common to use the Bayes Factors as test statistics, if nothing else because their distribution under H_0 is not known. In the next section we adopt a more realistic measure to assess both Bayesian and frequentist methods.

5.4.2 Small-scale genome simulations – Comparisons based on the ranking of the markers

As we mentioned earlier in this thesis, GWAS usually take the form of a series of single-marker tests. Since there is no golden rule of when a p-value or a Bayes factor is significant, the inference is usually based on the ranking of those p-values or Bayes factors. Although when using a powerful test is more likely to rank the significant markers close to the top there is no guarantee that this will always be the case nor that the ranking of the significant markers will correspond to their true significance. This is especially crucial in multi-stage GWAS, which nowadays are not uncommon, where we choose to genotype the top ranked markers on additional samples for subsequent stages of the analysis. Therefore it is no exaggeration to say

that it is equally important for a test to rank the markers according to their true significance as it is to be powerful.

For this reason we use small-scale genome simulations, where we have more than one marker associated with the disease. Since we know the actual significance ranking of those markers, we can assess the tests based on how they rank the markers. We consider the same siblings example we used thus far, to define the sub-populations.

We simulate genomes $\{c_{i.j}\}_{i=1,...C,j=1,...,n_C}$ where C is the number of chromosome pairs and $c_{i.j}$ is the j^{th} marker in the i^{th} chromosome pair. For each marker we randomly choose a minor allele frequency. Using those MAFs we simulate the parental haplotypes - a total of $2C$ haplotypes for each parent. Then using Mendelian inheritance we generate the genotypes of their two children. For every chromosome pair, each child inherits one haplotype from each parent. We do not consider any linkage or recombination effects.

5.4.2.1 Disease Model

We choose a small subset of the markers to be associated with the disease. So we want to define a disease model that is a function of those markers. In single-marker disease models we can view $f_0 = P(Y = 1|x = 0)$ as a baseline probability of disease. This probability takes into account genetic effects from other markers as well as environmental or other behavioural effects. The idea is that if the entire genome is known, we can adjust the disease model so that the input will be the entire genome $\mathcal{G}$ and the baseline

probability $P(Y = 1|\mathcal{G} = \mathbf{0})$ will account for the non-genetic effects. Also we develop the disease model so that the marginal disease models are the single-marker models that are assumed for each marker.

In general, let $\mathcal{G} = (x_1, x_2, \ldots, x_N)$ be the entire genome, and $\{\beta_i\}_{i=1}^N$ be the penetrance parameter of the disease model for marker i. Depending on the single marker disease model, there is a value β_i^0 such that, if $\beta_i = \beta_i^0$, then the i^{th} marker is not associated with the disease. So essentially if

$$\mathcal{G}^* = (x_1^*, \ldots x_m^*) \equiv \{x_i : \beta_i \neq \beta_i^0\}, \tag{5.41}$$

then the disease model depends only on $\mathcal{G}^*$ and not on the entire genome.

$$P(Y = 1|\mathcal{G}) = P(Y = 1|\mathcal{G}^*) \equiv f(x_1^*, \ldots x_M^*) \quad M \leq N \tag{5.42}$$

We consider four of the marginal (single-marker) disease models we describe in section 2.1.2: the logistic (f^L), additive (f^A), recessive (f^R) and dominant (f^D) models. Each marker in $\mathcal{G}^*$ has one of the aforementioned marginal disease models. Without loss of generality suppose that we re-order the markers in $\mathcal{G}^*$ so that the markers $1, \ldots, m_1$ have the logistic disease model, the markers $m_1+1, \ldots, m_2$ have the additive disease model, the markers $m_2+1, \ldots, m_3$ have the recessive disease model and the markers $m_3 + 1, \ldots, M$ have the dominant disease model. Then the joint disease model, or alternatively the

genome-based disease model is defined as,

$$P(Y=1|\mathcal{G}^*) \equiv f(\mathcal{G}^*) = \alpha_0 \left\{ \frac{(1+\exp(\alpha))\exp(\beta_1 x_1^* + \ldots \beta_{m_1} x_{m_1}^*)}{1+\exp(\alpha+\beta_1 x_1^* + \ldots \beta_{m_1} x_{m_1}^*)} \times \right.$$
$$\left. f^A(x_{m_1+1}^*) \cdots f^A(x_{m_2}^*) \cdot f^R(x_{m_2+1}^*) \cdots f^R(x_{m_3}^*) \cdot f^D(x_{m_3+1}^*) \cdots f^D(x_M^*) \right\} \tag{5.43}$$

The genome-based disease model is roughly a product of the marginal disease models. The only exception are the markers with logistic disease models. The parameter α_0 in (5.43) determines the baseline disease probability that is not due to genetic effects, while $\alpha = \log(\alpha_0) - \log(1-\alpha_0)$. This model maintains the marginal disease models for the individual markers.

We use this model to simulate the disease status of the siblings. Similar to our previous procedure, one sibling is considered the proband while the other helps determine whether the proband is super-affected or affected when the latter is diseased.

5.4.2.2 Relative Risk

In order to be able to assess our results we need a uniform measure for the significance of the associated markers. Obviously the penetrance parameters β_i are not an option since they have different parameter space for different models but also the significance of a marker also depends on its MAF. Even if two markers have exactly the same disease model - including same β - but different MAF, the effect of their respective rare alleles will be different. For example if they are causal, the rare allele of the marker with lower MAF will

cause the probability of disease to increase more than the rare allele of the marker with higher MAF. For this reason we use the marginal relative risk of the markers.

The genotypic relative risk,

$$RR_{gen;i} = \frac{P(Y=1|x=i)}{P(Y=1|x=0)} \quad , \quad i=1,2 \tag{5.44}$$

indicates how much more (or less) likely is for an individual to have the disease when it has one or two copies of the rare allele compared to having none. The genotypic relative risk is not really helpful for ranking purposes because for each marker we have two values. So we introduce the allelic relative risk.

$$RR_a = \frac{P(Y=1|\text{rare allele})}{P(Y=1|\text{common allele})} =: \frac{P(Y=1|\theta)}{P(Y=1|1-\theta)} \tag{5.45}$$

The notation in the right hand side of (5.45) is conventionally used. This is because the basis of our calculations of relative risk is the expression:

$$\theta|Y := P(\text{rare allele}|Y) = \frac{1}{2}P(x=1|Y) + P(x=2|Y) \tag{5.46}$$

In reality $\theta|Y$ is either θ_0, the conditional MAF of the unaffected subpopulation (when $Y=0$) or θ_2 the conditional MAF of the affected subpopopulation (when Y=1). Using the Bayes formula we can easily show that

$$RR_a = \frac{\theta_2(1-\theta_1)}{(1-\theta_2)\theta_1}, \tag{5.47}$$

where θ_1 is the unconditional population MAF. The allelic relative risk is good to compare markers that have the logistic or additive disease models. However it will underestimate the significance of markers with dominant or recessive effects. For such markers we use:

$$RR_{rec} = \frac{P(Y=1|x=2)}{P(Y=1|x<2)} = \frac{p_2(g_0+g_1)}{(p0+p1)g_2} \quad \text{and} \tag{5.48}$$

$$RR_{dom} = \frac{P(Y=1|x>0)}{P(Y=1|x=0)} = \frac{(p_1+p_2)g_0}{p_0(g_1+g_2)}. \tag{5.49}$$

Another issue that arises is that we have to compare markers with protective rare alleles and markers with causal rare alleles. So for instance we will always have $RR_a > 1$ when the rare allele is causal and $RR_a < 1$ when the rare allele is protective. Fortunately due to symmetry when the rare allele is causal then the common allele is protective and vice versa. Furthermore if RR_a^m is the relative risk of the rare allele and RR_a^M the relative risk of the common allele, then $RR_a^m = (RR_a^M)^{-1}$. This is also true for the relative risk associated with recessive and dominant effects. In fact a recessive model with causal rare allele is equivalent to a dominant model with protective common allele. Thus for the sake of comparison we use the allele that has the causal effect. This does not affect the significance of a given marker since the rare allele and consequently the minor allele frequency are used in most studies by convention if nothing else. The fact is that, when testing for association, we will get the exact same results whether we choose to code the rare allele with 1 and the common allele with 0 or choose the opposite coding. Therefore we use the causal allelic relative risk $RR_a^c = \max(RR_a^M, RR_a^m)$. Overall, adjust-

ing both for dominant and recessive effects as well as causal and protective alleles we have the adjusted, for comparisons, relative risk:

$$RR^c_{adj} = \begin{cases} \max\left\{\frac{\theta_2(1-\theta_1)}{(1-\theta_2)\theta_1}, \frac{(1-\theta_2)\theta_1}{\theta_2(1-\theta_1)}\right\} & \text{If the marker has the logistic} \\ & \text{or additive disease model,} \\ \max\left\{\frac{p_2(g_0+g_1)}{(p0+p1)g_2}, \frac{(p_1+p_2)g_0}{p_0(g_1+g_2)}\right\} & \text{If the marker has the recessive} \\ & \text{or dominant disease model.} \end{cases} \tag{5.50}$$

5.4.2.3 Examples

The objective of our simulation study is to assess and compare the various methods - trend tests and Bayesian models - in the presence of more than one marker that is associated with the disease and in addition a number of markers that are not associated with that disease. The scale we use is small compared to actual GWAS but it is sufficient to draw some conclusions about the methods we consider. We present three examples. In the first two examples we only consider the logistic disease model for all associated markers. In the third example we consider different disease models for different associated markers. Furthermore, in the first example we consider relatively large effect sizes while in the last two examples we use smaller effect sizes that provide greater challenge to our methods and make it harder to identify the associated markers. The sample sizes for all three examples are 1000 super-cases and super-controls and 2000 cases and controls.

Additional figures and tables for the examples we present below, can be

found in Appendix C.

Example 1

In the first example we simulate a hypothetical population with 5 chromosome pairs and 50 biallelic markers in total. We consider those markers to be SNPs. There are 14 SNPs on the first chromosome, 11 on the second, 10 on the third, 8 on the fourth and 7 on the fifth. There are six markers that are associated with the disease, c1.3, c1.13, c2.6, c3.9, c4.8 and c5.2. Table 5.2 provides details regarding those markers as well as figure 5.6.

Table 5.2: Information on the 6 markers that are associated with the disease in Example 1.

	c1.3	c1.13	c2.6	c3.9	c4.8	c5.2
MAF	0.08	0.29	0.22	0.16	0.32	0.49
β	0.643	-0.261	0.285	-0.236	0.411	0.166
RR_a	1.805	0.776	1.300	0.807	1.454	1.167
RR_a^c	1.805	1.289	1.300	1.240	1.454	1.167

The results (Figure 5.7 and Table 5.3) show that neither method has any real problems in identifying the markers that are truly associated with the disease and in addition most of them are able to pick them in the correct order. There is an issue with marker c2.6 (rank 3) and marker c1.13 (rank 4) where some tests rank 4th and 3rd respectively. However the effect sizes of those those markers are very similar; RR_a^c of 1.300 and 1.289 respectively, so this is not a major issue concerning the accuracy of those methods. The only other major problem concerns the CA test where we only use the cases and controls (CA_{cc}). While all the other methods identify the 6 markers as the top 6 most significant ones; based on the p-values or Bayes factors, CA_{cc} ranks marker c3.9, 8th (actual rank 5) while it includes marker c5.4, which is

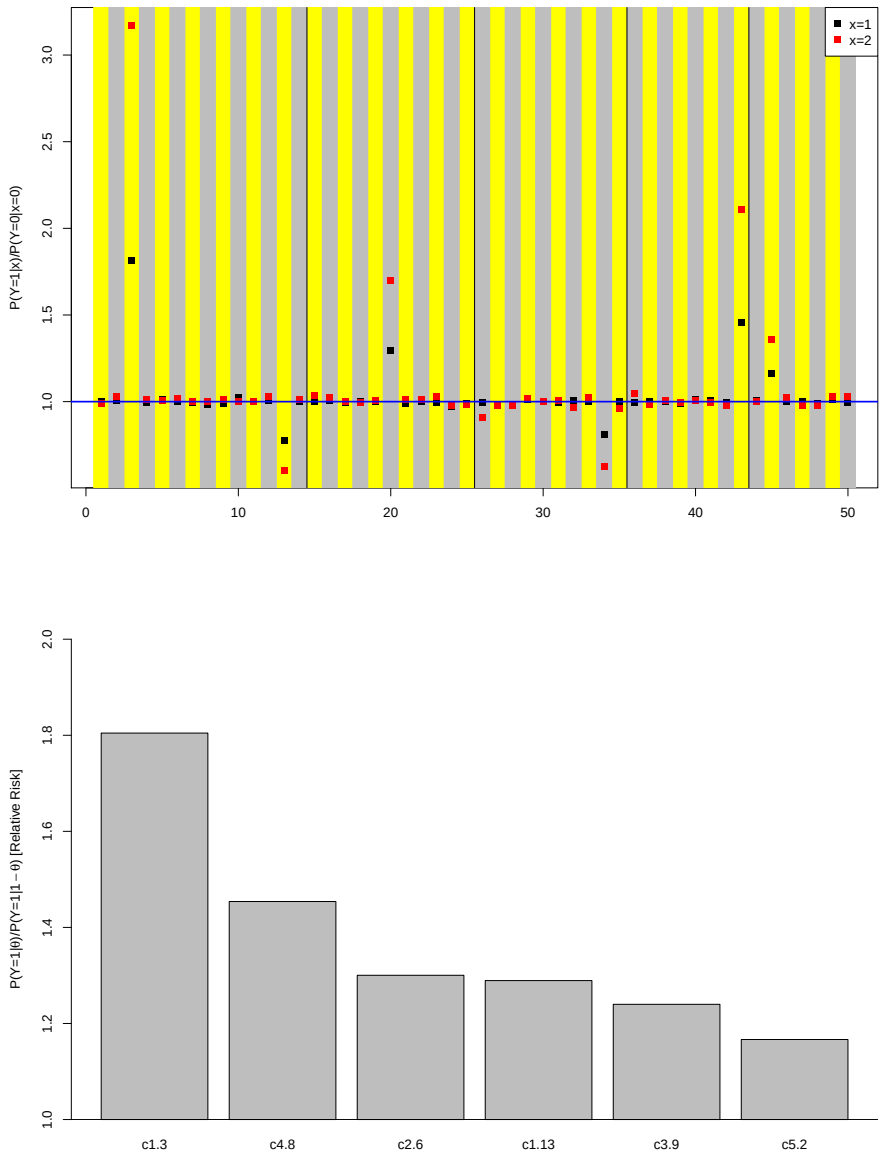


Figure 5.6: Genotypic Relative Risk for the markers in Example 1 (top) and Causal Allelic Relative Risk (RR_a^c) for the 6 markers that are associated with the disease (bottom).

extremely well.

Table 5.3: The ranking of the 6 markers according to the 9 methods we consider (Example 1).

Marker	c1.3	c4.8	c2.6	c1.13	c3.9	c5.2
Actual Order	1	2	3	4	5	6
	Observed Order					
Bay. Mod. 1	1	2	4	3	5	6
Bay. Mod. 2	1	2	3	4	5	6
X^2_{all}	1	2	3	4	5	6
X^2_{all2}	1	2	3	4	5	6
CA_{cc}	1	2	4	3	8	5
CA_{ss}	1	2	3	4	5	6
CA_{scsc}	1	2	4	3	6	5
CA_{scs1}	1	2	3	4	5	6
CA_{scs2}	1	2	3	4	5	6

Example 2

In the second example we simulate a hypothetical population with 3 chromosome pairs and 26 markers in total. There are 10 markers on the first chromosome and 8 markers on the second and third chromosomes. Four markers (c1.5, c1.9, c.26 and c3.1) are associated with the disease. In this example we assume that the rare allele in all four markers is causal however we choose relatively small effect sizes. Details on the population and the four markers are provided in table 5.4 and figure 5.8.

Table 5.4: Information on the 4 markers that are associated with the disease in Example 2

	c1.5	c1.9	c2.6	c3.1
MAF	0.46	0.20	0.38	0.13
β	0.056	0.114	0.045	0.087
RR_a $(= RR^c_a)$	1.052	1.128	1.055	1.093

Due to the small effect sizes we simulate and apply our methods in three

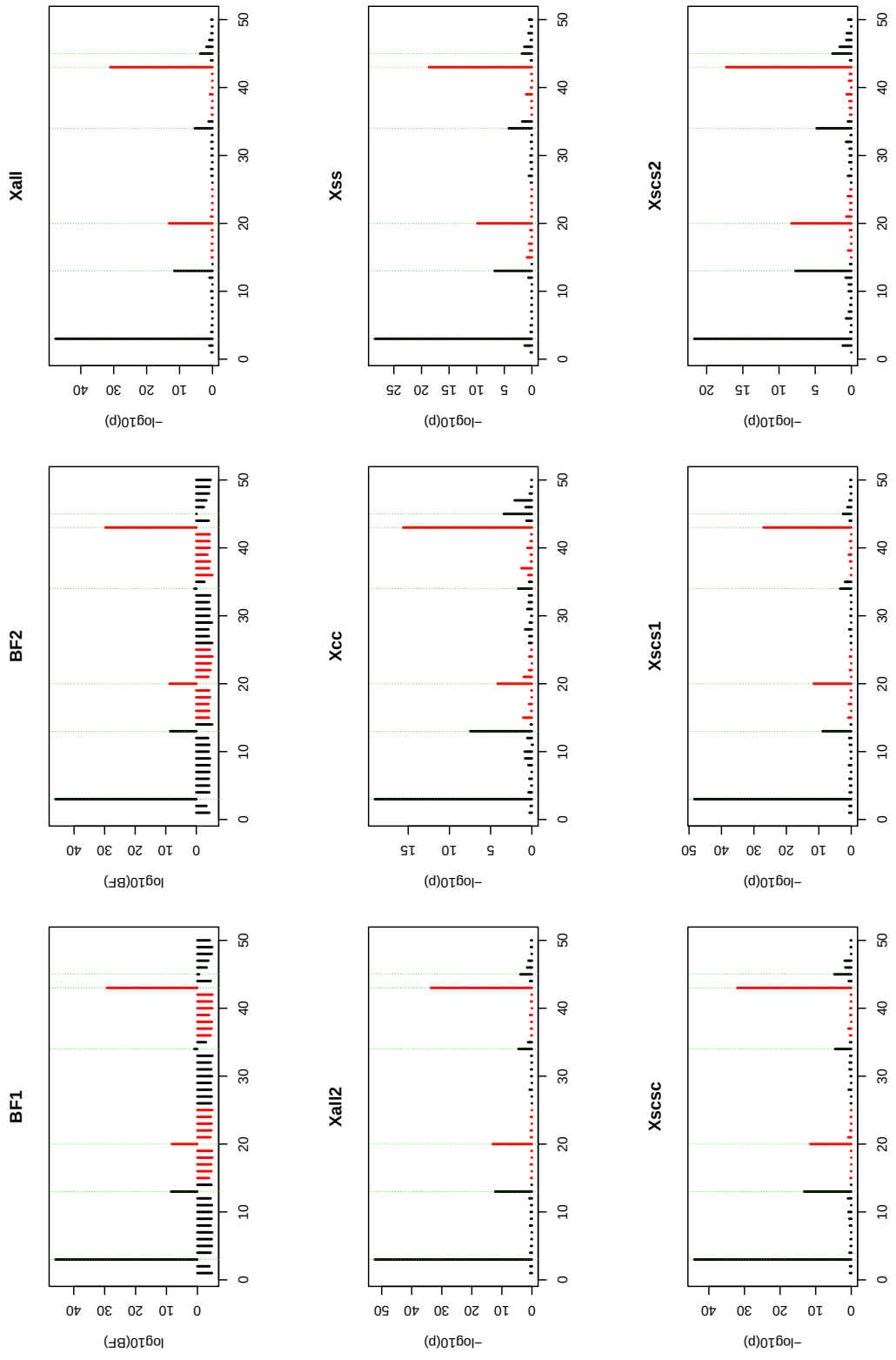


Figure 5.7: The results of the 9 methods applied in the data generated for Example 1. The first 2 figures show the Bayes factors in $\log_{10}$–scale while the rest show p-values in $-\log_{10}$–scale

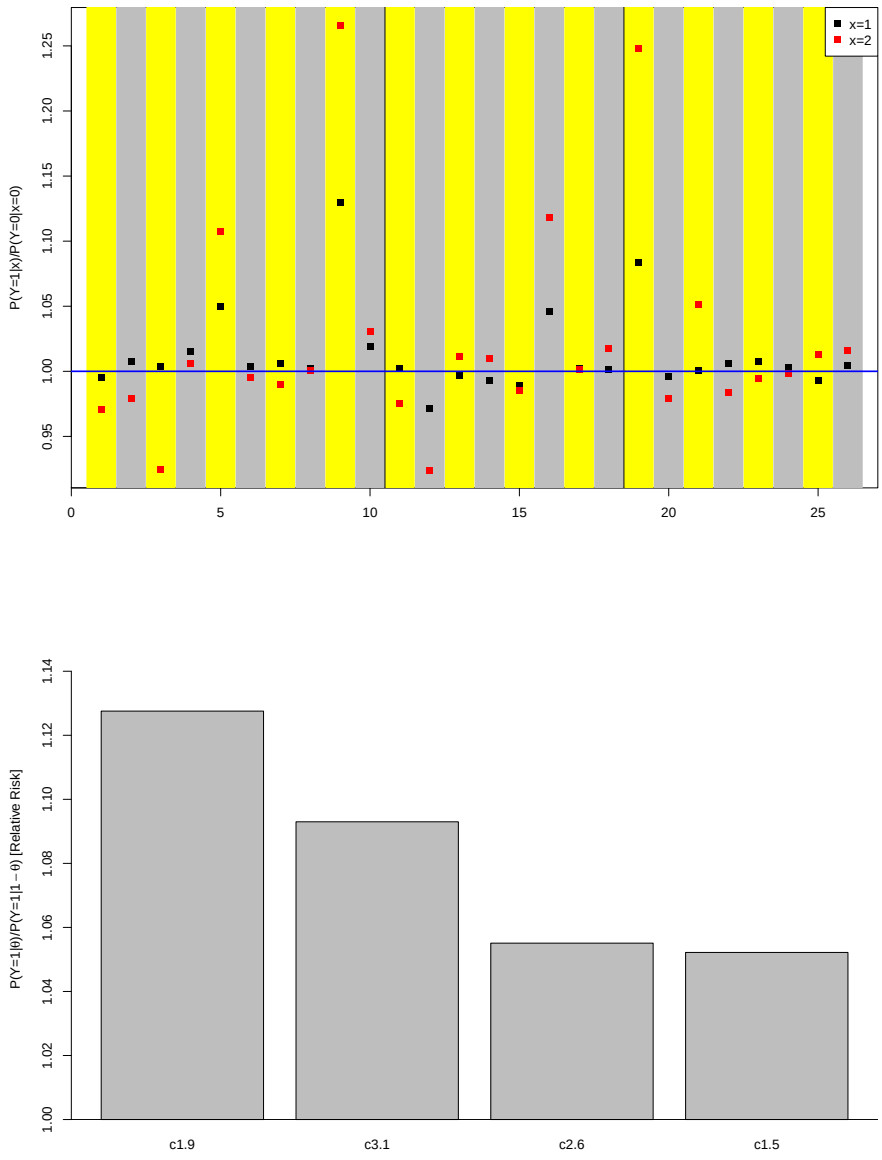


Figure 5.8: Genotypic Relative Risk for the markers in Example 2 (top) and Causal Allelic Relative Risk (RR_a^c) for the 6 markers that are associated with the disease (bottom).

different datasets. Figure 5.9 shows the results for one of those datasets. The results for the other two can be found in Appendix C. Table 5.5 shows the average ranking of the four markers based on the application of the methods on the three datasets; the rankings for each dataset are presented in table C.1 in Appendix C. No method is able to correctly identify all four markers. Especially marker c3.1 is ranked very low for 2 of the 3 datasets. This probably has to do mostly with "bad" samples of either super-cases or super-controls since CA_{cc} which only uses cases and controls is the only test that, on average, ranks that marker among the top 10 markers. As for the other three markers, they are, on average, picked among the top four by the second Bayesian model (BM2) , the allelic trend test X^2_{all} and the CA test that uses all four samples (CA_{scsc}). The first Bayesian model (BM1) and the CA test that discards the controls (CA_{scs2}) also perform fairly well. In general if we had to choose one method for this setting, i.e. small effect size, that woud be the Cochran-Armitage test CA_{scsc}. This is exactly because of the small effect size, any of the samples is susceptible to be a "bad" sample. Because in CA_{scsc} we use the joint sample of super-controls and controls and the joint-sample of super-cases and cases, if only one of the samples is not good, it does not affect the outcome as it does for the other methods.

Example 3

In our third and final example we consider a hypothetical population with six chromosome pairs and sixty markers in total. The chromosomes have 15, 12, 10, 10, 8 and 5 markers respectively. In this example in addition to relatively low risk alleles (small effect sizes) we consider different disease

Table 5.5: The average ranking (based on 3 datasets) of the 4 markers according to the 9 methods we consider (Example 2).

Marker	c1.9	c3.1	c2.6	c1.5
Actual Order	1	2	3	4
	Av. Observed Order			
Bay. Mod. 1	3.66	17.66	4.66	4.66
Bay. Mod. 2	3.33	15.66	3.66	3.33
X^2_{all}	3.33	12.66	4.00	3.66
X^2_{all2}	2.66	13.33	3.66	5.00
CA_{cc}	3.66	8.66	4.00	6.33
CA_{ss}	4.33	15.00	8.66	6.33
CA_{scsc}	2.66	10.00	3.66	3.00
CA_{scs1}	2.66	14.66	6.33	8.00
CA_{scs2}	4.33	13.66	4.00	4.00

models. We assume the logistic disease model for the rare allele in markers c1.12 and c3.5, the additive model for the rare allele in markers c1.4 and c6.1, the recessive model for the rare allele in marker c2.5 and the dominant model for the rare allele in marker c4.7. Details are presented in table 5.6 and figure 5.10 as well as in Appendix C.

Table 5.6: Information on the 6 markers that are associated with the disease in Example 3

	c1.4	c1.12	c2.5	c3.5	c4.7	c6.1
MAF	0.12	0.40	0.33	0.06	0.26	0.19
Disease Model	A	L	R	L	D	A
β	1.187	-0.086	1.316	0.177	1.115	1.081
RR_a	1.183	0.922	1.098	1.185	1.074	1.084
RR^c_a	1.183	1.085	1.098	1.185	1.074	1.084
RR^c_{adj}	1.183	1.085	1.306	1.185	1.111	1.084

Similar to our approach in example 2 we apply our methods in three different datasets. The results of the application in one of those datasets are presented in figure 5.11 and the average ranking of the 6 markers that are

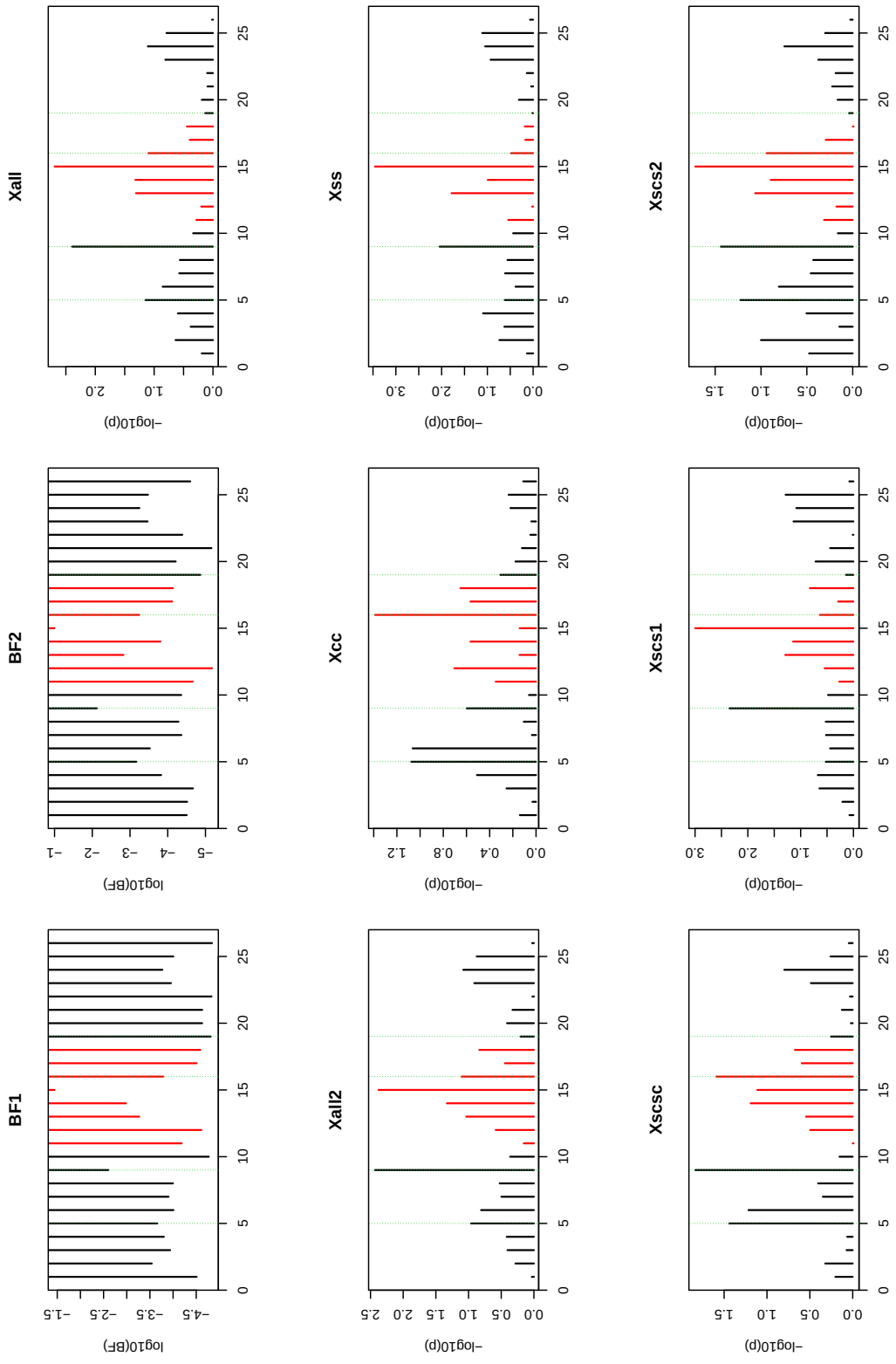


Figure 5.9: The results of the 9 methods applied to one of the datasets generated for Example 2.

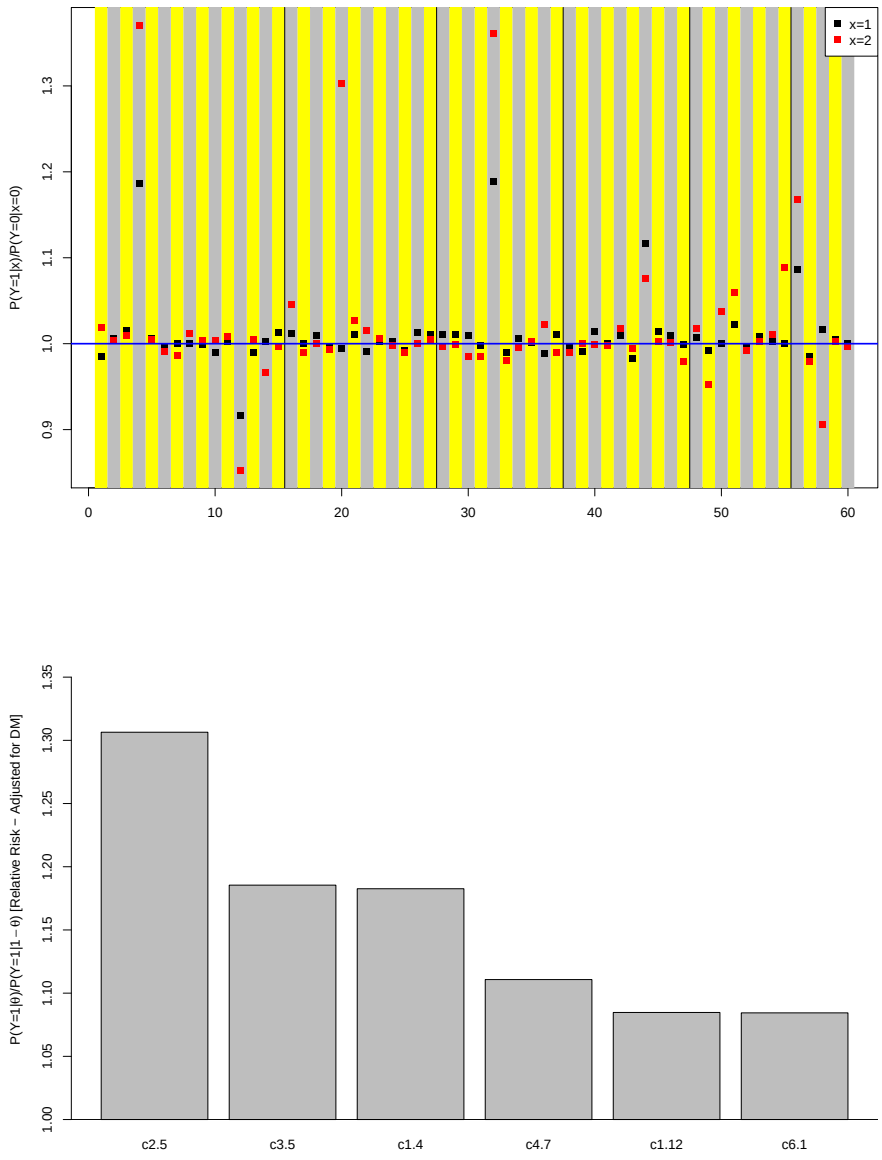


Figure 5.10: Genotypic Relative Risk for the markers in Example 3 (top) and Adjusted Causal Allelic Relative Risk (RR^c_{adj}) for the 6 markers that are associated with the disease (bottom).

associated with the disease are presented in table 5.7. More detailed results for each dataset can be found in Appendix C. Except for the case-control CA test (CA_{cc}), no other test has any problems to identify the most significant marker c2.5. Of course its effect size is not as small as the effect size of the other five. Problems arise with the fourth and fifth ranked markers (c4.7 and c1.12). In the case of c4.7, the marker is ranked very low in only one dataset and thus its low average rank can be attributed to that one "bad" sampling. Otherwise, the methods that use all four samples (BM1, BM2, X^2_{all}, X^2_{all2} and CA_{scsc}) have no issues, although BM1 might not be considered as good as the others. The other marker, c1.12, consistently fails to be identified by any method we use. This is probably not just because of its small effect size but also the presence of another, most significant marker (c1.4), in the first chromosome. Overall the methods that perform better in this example are the second Bayesian model (BM2) and the two allelic trend tests (X^2_{all} and X^2_{all2}) and to lesser extend the first Bayesian Model (BM1) and the CA test that uses the four samples (CA_{scsc}).

5.5 Testing for Association Using Simple Linear Regression

Ordinary linear regression is rarely used in GWAS. The reason is that the response - most commonly the disease status - is usually a dichotomous variable or sometimes a discrete variable with finite domain or even a categorical variable. This, by definition violates the assumption of normality of the lin-

Table 5.7: The average ranking (based on 3 datasets) of the 6 markers according to the 9 methods we consider (Example 3).

Marker	c2.5	c3.5	c1.4	c4.7	c1.12	c6.1
Actual Order	1	2	3	4	5	6
	Observed Order					
Bay. Mod. 1	1.33	5.00	2.66	24.33	27.33	5.66
Bay. Mod. 2	1.33	6.66	2.66	19.00	22.00	4.00
X^2_{all}	2.00	3.33	2.33	15.33	25.66	4.00
X^2_{all2}	1.33	4.33	2.66	13.33	28.00	3.33
CA_{cc}	5.66	9.66	14.00	16.33	17.66	6.66
CA_{ss}	3.00	7.00	1.66	24.66	37.00	7.33
CA_{scsc}	3.00	8.00	3.66	14.00	22.00	4.00
CA_{scs1}	1.66	4.66	2.00	16.33	35.66	3.66
CA_{scs2}	3.00	5.33	2.33	27.33	22.33	15.33

ear regression model. However studies involving analysis of binary response data using linear regression exist in literature (David Wilson and Meinert, 1984; Cantor et al., 2010).

Ordinary linear regression analysis is certainly inadequate in GWAS when we want to estimate the effect size of the marker. However if our aim is just to identify whether a marker is associated with a disease or not, then linear regression can be used with fairly accurate results. We show that the linear model we propose for the SCCS design is as powerful as the allelic trend test. Thus, linear regression can be used, for example, in early stages of multi-stage GWAS, when our main concern is to identify potential significant markers and not as much to quantify their effect size.

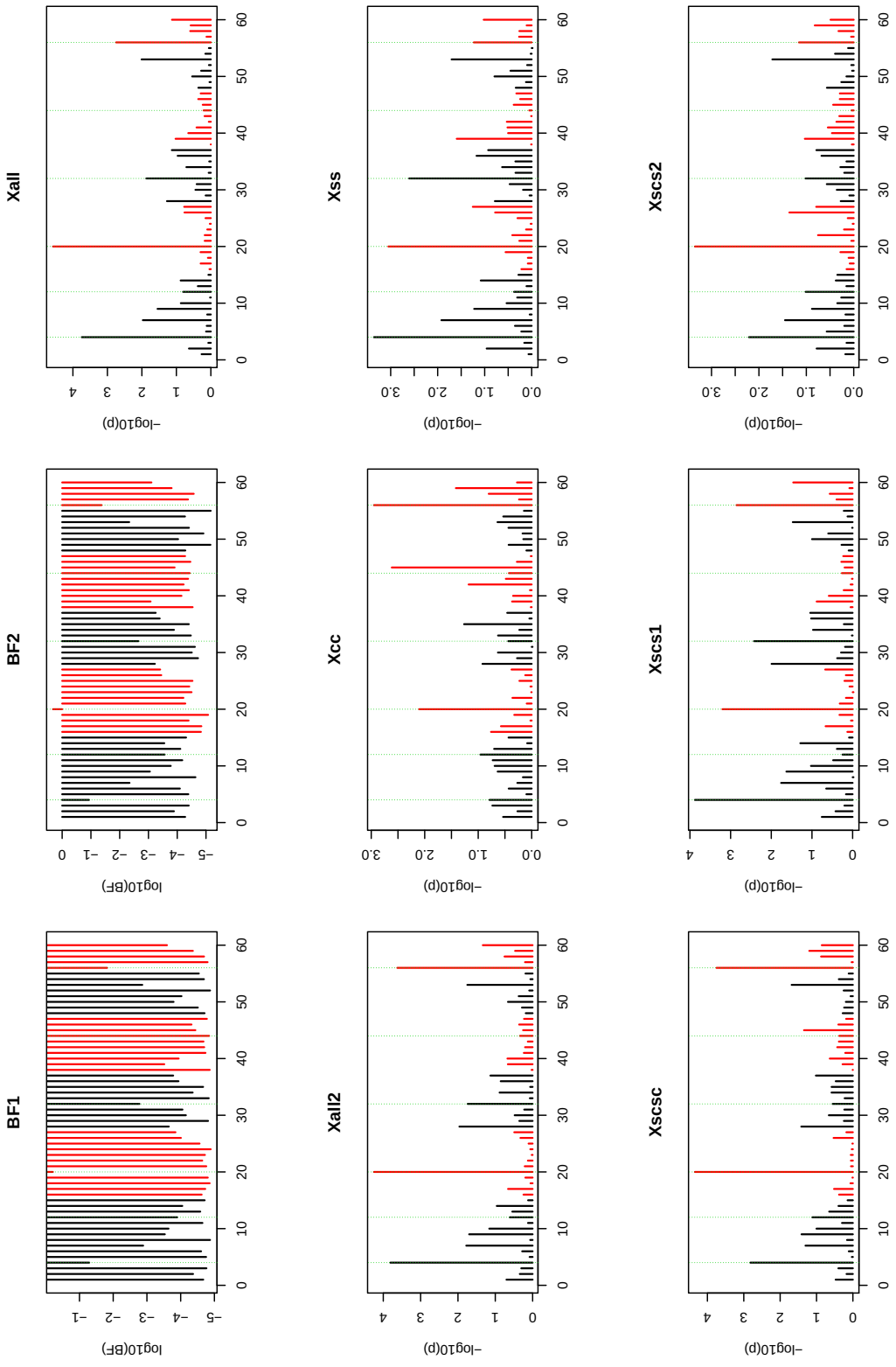


Figure 5.11: The results of the 9 methods applied to one of the datasets generated for Example 3.

5.5.1 The Linear Regression Model

We consider the model,

$$\psi = a + bx + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \tag{5.51}$$

where $\psi := \psi(S_j)$, $j = 0, 1, 2, 3$ is a discrete variable that takes four values depending on the sample group that the individual belongs. We want to maintain a connection between the actual disease status Y and the response in our linear model. If we choose $\psi(S_j) = P(Y = 1|S_j)$ this will lead to cases and super-cases having the same value of the response. Hence we keep that choice of ψ for the super-controls, controls and cases, while for the super-cases we set $\psi(S_3) = P(Y = 1|S_3) + \gamma$ where gamma is a constant. We discuss later in this chapter about the choice of γ. Overall we have:

$$\psi_j := \psi(S_j) = \begin{cases} 0, & \text{if the individual is super-control } (j = 0) \\ K, & \text{if the individual is control } (j = 1) \\ 1, & \text{if the individual is case } (j = 2) \\ 1 + \gamma, & \text{if the individual is super-case } (j = 3) \end{cases} \tag{5.52}$$

It is useful to remind that K is the disease prevalence in the population. Additionally we understand that the assumption of the errors' normality in (5.51) is a bold statement, however it is an assumption we have to make in order to fit the model using the ordinary least squares (OLS) method.

When we use the linear model in (5.51), testing for genetic association

is equivalent to testing whether the slope coefficient b is equal to zero or not. A zero slope means that the genotype x is not ssociated with ψ and consequently, due to the choice of ψ (5.52), it is not associated with the disease status.

5.5.2 Connection between the Linear Regression Model and the Allelic Trend Test

In general, when we have two variables, say U and V, then the fitted line from the regression of U on V has slope that is the reciprocal of the slope of the fitted line obtained from the regression of V on U. If the two variables are not associated, the fitted line in both cases will be a horizontal (zero slope) line with intercept equal to the sample mean of the response. Thus when we are interest in identifying potential association between U and V we can use either regression.

This example roughly reflects the connection between the linear regression model we discuss in this section and the allelic trend test we discussed in section 5.3. The outcome of the allelic trend test is based on the slope of line that fits the points $E(x|z_i)$, $i = 0, 1, 2, 3$. On the other hand, the linear regression model fits a least-square line for the model $E(\psi|x)$ $x = 0, 1, 2$. Since all the the variables - the genotype x, the scores of the trend test z and the response of the linear model ψ - are discrete and finite there are are essentially four points that determine the fitted slope in the allelic trend test and three points that determine the fitted slope in the linear model. Those

are the conditional sample means $\left\{\overline{x|z_i}\right\}_{i=0}^{3}$ and $\left\{\overline{\psi|x}\right\}_{x=0}^{2}$ respectively. Thus if we use $z_i = \psi_i$ we expect that the two methods will give the same evidence for association.

We use power simulations for different values of ψ (equivalently z). The results of the simulations show that when we use $z_i = \psi_i$ the power of the test for slope in the regression model is exactly the same as the power of the allelic trend test (Figure 5.12). The significance of this property will become more evident in the following section.

5.5.3 An Asymptotic Method for Calculating the Power of the Linear Regression Model

First we should clarify that when we say the power of the linear regression model we mean the power of the test for slope of the linear model, which in the case of the simple linear regression is equivalent to the overall test for regression. Thus we consider the linear regression model as a method that provides evidence for or against the association of the genotype with the disease status.

Since the tests for the significance of the individual coefficients in a linear model are t-tests, calculating the power of the linear regression model is analogous to the calculation of the power of the t-test albeit more complex due to the presence of the independent variable (x). Consider the general

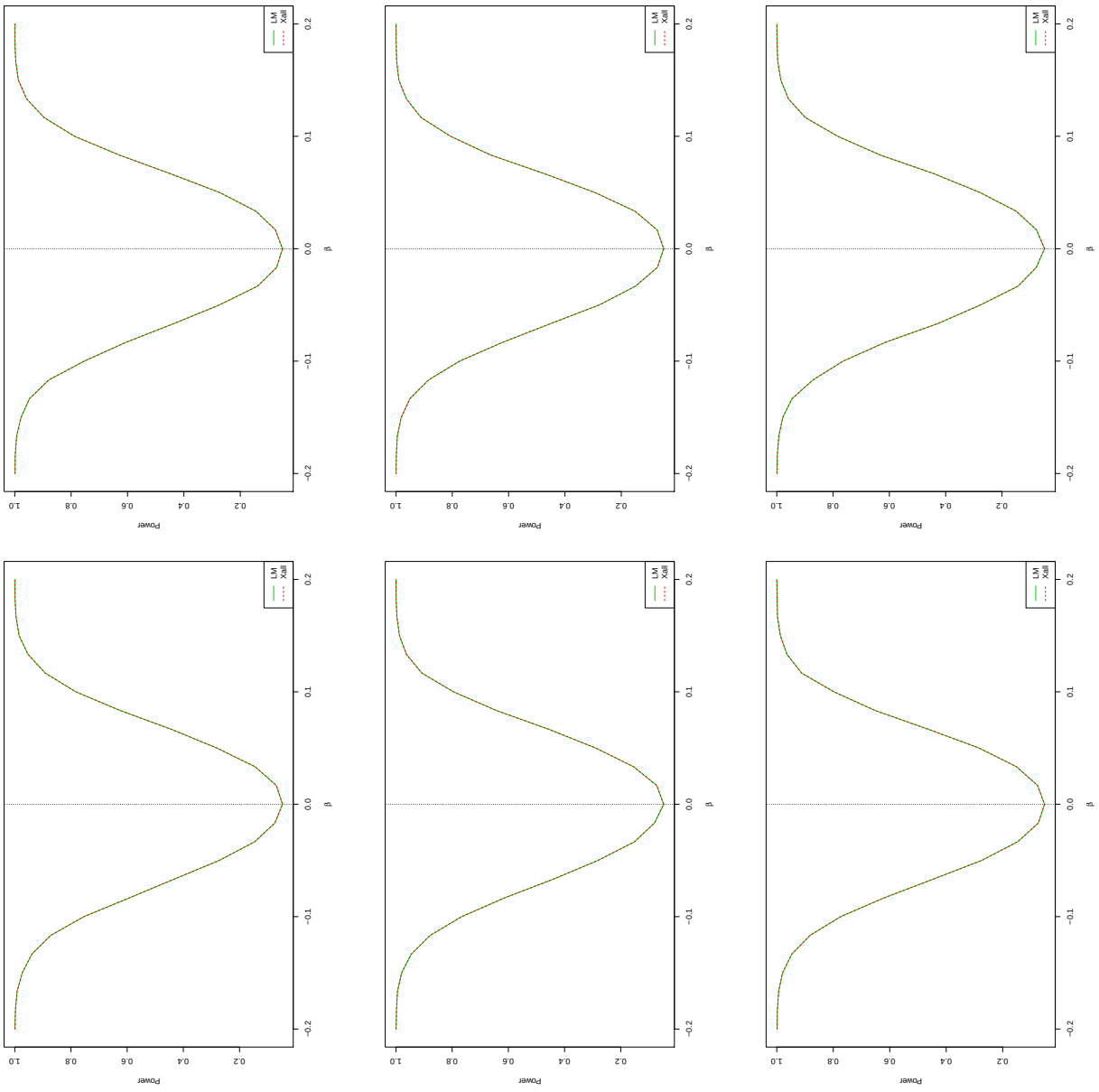


Figure 5.12: Power simulations comparing the power of the allelic trend test with the power of the test for slope in the linear regression model. In this example we used $\psi_0 = z_0 = 0$, $\psi_1 = z_1 = K$, $\psi_2 = z_2 = 1$ and $\psi_3 = z_3 = 1 + \gamma$. From left to right and top to bottom we use $\gamma = 0.05, 0.1, 0.2, 0.5, 0.75, 1$.

case of a linear regression model,

$$y = a + bx + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{5.53}$$

The Fisher information matrix of $\phi = (a, b, \sigma^2)$ is:

$$I(\phi) = \frac{1}{\sigma^2} \begin{pmatrix} n & \sum_{i=1}^n x_i & 0 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \\ 0 & 0 & \frac{n}{2\sigma^2} \end{pmatrix} \tag{5.54}$$

Under Gaussianity the ordinary least squares (OLS) estimates $\hat{a}$, $\hat{b}$ of a and b are also maximum likelihood estimates (MLEs). Therefore, based on the propertis of MLEs we know that those estimates have an approximate normal distribution. Specifically for $\hat{b}$ we have:

$$\hat{b} \approx N\left(b, \{I(\phi)\}_{2,2}^{-1}\right) = N\left(b, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \tag{5.55}$$

Equivalently,

$$\frac{\hat{b}^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \approx \chi_1^2(\delta), \quad \text{where } \delta = \frac{b^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2}. \tag{5.56}$$

The non-central chi-square distribution $\chi_1^2(\delta)$ is the one that is used in the power analysis of the linear model. One would choose a significance level α, choose the values of β and σ^2 and calculate the power $1 - F(q_{1-\alpha})$, where $q_{1-\alpha}$ is the $1 - \alpha$ quantile of the central χ_1^2 distribution and F is the cumulative density function (CDF) of the non-central $\chi_1^2(\delta)$ distribution. Thus power

analysis comes down to the calculation of δ. However, even in this general case there is a problem due to the presence of $\sum_{i=1}^{n}(x_i - \bar{x})^2$ in the formula of δ. This means we need to observe the data before doing power analysis while on the other hand power analysis is normally carried out without the need to observe any data.

In our problem, where we use linear regression to identify genetic association, things are even more complicated. It is very unlikely that the data were generated from a linear model; a more realistic assumption would be one of the models we describe in section 2.1.2. Therefore both the slope and the variance component of the linear model have no direct significance in our power analysis. We are interested to study the power as a function of the parameters that are involved in the genetic model, namely the MAF, the disease prevalence, the effect size and the disease model itself. Only the sample sizes have the same interpretation in the context of a linear model.

For this reason we develop an asymptotic method for calculating the power of the test for association, when the data are generated from any given disease model and are analysed using linear regression. An additional benefit is that this method can also be used for the power analysis of the allelic trend test. Even though we present an asymptotic formula for the power of the trend test in chapter 3, this cannot be used for the allelic test. That formula assumes that the row marginal totals are fixed and known, something that does not hold for the allelic table of counts. Since we already showed that both the linear model and the allelic trend test have the same power, we can use this method for sample size determination or effect size

power analysis for either test.

5.5.3.1 Description of the Asymptotic Method

Our objective is to find a consistent estimator of δ, which will be a function of the parameters we are interested in $(\theta_1, K, \beta, \boldsymbol{N})$. If we have an estimate $\hat{\delta}$ we can then use the CDF of $\chi^2_1(\hat{\delta})$ distribution to asymptotically calculate the power of the linear model. However instead of finding an estimate of δ directly we instead do that indirectly, using plug-in estimators. Since δ can be expressed as a product of three factors; b^2, $\sum_{i=1}^{n}(x_i - \bar{x})^2$ and σ^{-2}, we opt to find consistent estimators for those quantities and then plug them in the formula of δ (5.56) to obtain $\hat{\delta}$. Furthermore instead of an estimate of b^2, we find an estimate an b and then use its square $(\hat{b}^2)$. Similarly instead of σ^{-2} we estimate σ^2 and use the reciprocal $\hat{\sigma}^{-2}$. For the sake of conciseness we use the notation $Q = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

The first thing we need to make is an assumption about the disease model f_x, $x = 0, 1, 2$, that generated the data. This can be any model as long as we can calculate the conditional genotype frequencies. It is those frequencies that we need in our estimators. We remind that those are:

Unaffected Genotype Frequencies (Super-Controls)

$$q_x = P(x|S_0) = \frac{g_x(1 - f_x)}{1 - K} \tag{5.57}$$

Population Genotype Frequencies (Controls)

$$g_x = P(x|S_1) = \begin{cases} (1-\theta_1)^2, & x = 0 \\ 2\theta_1(1-\theta_1), & x = 1 \\ \theta_1^2, & x = 2 \end{cases} \tag{5.58}$$

Affected Genotype Frequencies (Cases)

$$p_x = P(x|S_2) = \frac{g_x f_x}{K} \tag{5.59}$$

Super-affected Genotype Frequencies (Super-cases)

$$r_x = P(x|S_3) = \frac{f_x P(x|\text{Sibl. Aff.})}{\sum_{x=0,1,2} f_x P(x|\text{Sibl. Aff.})} \tag{5.60}$$

where,

$$P(x|\text{Sibl. Aff.}) = \frac{1}{4}g_x + \mathcal{I}_{[x=0]}(1-\theta_1)\left(p_0 + \frac{p_1}{2}\right) + \mathcal{I}_{[x=1]}\left(\theta_1 p_0 + \frac{p_1}{2} + (1-\theta_1)p_2\right) + \mathcal{I}_{[x=2]}\theta_1\left(\frac{p_1}{2} + p_2\right) + \frac{1}{4}p_x$$

The effect size is taken into account in the power analysis through those conditional genotype frequencies. The effect size can be expressed as a function of the penetrance parameter, β, and since β is an integral part of any disease model, when we use the disease model for the above calculations we involve the effect size as well.

5.5.3.2 The plug-in estimator of b

The basis for the derivation of the plug-in estimator $\hat{b}$ of b is its MLE/OLS estimator $\hat{b}_{MLE}$, which in the case of the general linear model (5.53) is given by:

$$\hat{b}_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5.61}$$

If we divide both the nominator and the denominator in (5.11) by n, then $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ converges to $E(x - E(x))(y - E(y))$ as $n \to \infty$. Similarly $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ converges to $E(x - E(x))^2$. Taking these into account, for our linear model - (5.51) and (5.52) - we consider the estimator:

$$\hat{b} = \frac{E_{x,\mathcal{S}}\left\{[x - E_x(x|\boldsymbol{N})]\,[\psi(\mathcal{S}) - E_{\mathcal{S}}(\psi(\mathcal{S})|\boldsymbol{N})]\,|\boldsymbol{N}\right\}}{E_x\left\{[x - E_x(x|\boldsymbol{N})]^2\,|\boldsymbol{N}\right\}} \tag{5.62}$$

Here $\boldsymbol{N} = (N_0, N_1, N_2, N_3)$ are the sample sizes and $\mathcal{S}$ is the categorical random variable with levels S_0, S_1, S_2 and S_3. The notation $E_v(\cdot|\boldsymbol{N})$ means that the expectation is conditional on the sample sizes and it is taken with respect to the variable v - which is x, $\mathcal{S}$ or both (joint expectation) in the estimator above. We define:

$$\hat{\bar{x}} := E_x(x|\boldsymbol{N}) = \frac{\sum_{i=0}^{3} N_i\big(P(x=1|S_i) + 2P(x=2|S_i)\big)}{N} \tag{5.63}$$

$$\hat{\bar{\psi}} := E_{\mathcal{S}}(\psi(\mathcal{S})|\boldsymbol{N}) = \frac{\sum_{i=0}^{3} N_i\psi(S_i)}{N} \tag{5.64}$$

where N is the total sample size. Also the probability for an individual to belong in a certain sub-population conditional on the sample sizes, is equal

to the proportion of the sub-population's sample in the total sample.

$$P(S_i|\boldsymbol{N}) := P(\mathcal{S} = S_i|\boldsymbol{N}) = \frac{N_i}{N} \tag{5.65}$$

The genotype probability, conditional on the sample sizes is given by:

$$P(x|\boldsymbol{N}) = \sum_{j=0}^{3} P(x, S_j|\boldsymbol{N}) = \sum_{j=0}^{3} P(x|S_j)P(S_j|\boldsymbol{N}) \tag{5.66}$$

Using the estimates and probabilities, (5.63)-(5.66), plus the conditional genotype probabilities, (5.57)-(5.60), we can calculate the plug-in estimator $\hat{b}$.

$$\hat{b} = \frac{\sum_{x=0}^{2}\sum_{j=0}^{3} P(x|S_j)P(S_j|\boldsymbol{N})(x - \hat{\bar{x}})(\psi(S_j) - \hat{\bar{\psi}})}{\sum_{x=0}^{2}\sum_{j=0}^{3} P(x|S_j)P(S_j|\boldsymbol{N})(x - \hat{\bar{x}})^2} \tag{5.67}$$

5.5.3.3 The plug-in estimator of σ^2

We propose two methods to obtain a plug-in estimator of σ^2. The first is based on estimating the variance of the response directly, while the second has its basis in the residual sum of squares estimator that is used in linear regression analysis.

Method 1

First we consider the distribution of the response varible in a linear model. In the general linear model, $y \sim N(a + bx, \sigma^2)$. We use the definition of the variance. If Y is a random variable, then $\text{Var}(Y) = E(Y - E(Y))^2$. Thus for the power analysis of our linear model the plug-in estimator of σ^2 is given

by:

$$\hat{\sigma}^2 = E_{\mathcal{S}}\Big(\psi(\mathcal{S}) - E_{\mathcal{S}}\big(\psi(\mathcal{S})|\boldsymbol{N}\big)|\boldsymbol{N}\Big)^2 = \sum_{j=0}^{3} P(S_j|\boldsymbol{N})\big(\psi(S_j) - \hat{\bar{\psi}}\big)^2 \quad (5.68)$$

Method 2

In the second method we consider the error variance and its OLS estimate. In general $\hat{\sigma}^2_{OLS} = \frac{RSS}{n-2}$, where the residual sum of squares (RSS) is $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Thus, we calculate the expected value of RSS in our model. However, first we need an estimate of the fitted values of the response and in order to do this we need an estimate of the intercept coefficient. Based on its OLS/MLE estimator, we define:

$$\hat{a} = \hat{\bar{\psi}} - \hat{b} \cdot \hat{\bar{x}} \quad (5.69)$$

where $\hat{b}$ is the estimate given by (5.67). Thus, the expected value of RSS is:

$$\widehat{E[RSS]} = E_{x,\mathcal{S}}\big(\psi(\mathcal{S}) - \hat{a} - \hat{b}x|\boldsymbol{N}\big)^2. \quad (5.70)$$

For the simple linear regression, it can be shown that $E(RSS) = (n-2)\sigma^2$ (Draper and Smith, 1998). In order to make the bias correction, we consider the estimate of RSS,

$$\widehat{RSS} = N \cdot \widehat{E[RSS]}. \quad (5.71)$$

Thus, overall the unbiased plug-in estimator of σ^2 is:

$$\hat{\sigma}^2 = \frac{\widehat{RSS}}{N-2} \equiv \frac{N \cdot \widehat{E[RSS]}}{N-2} = \frac{N}{N-2} \sum_{x=0}^{2} \sum_{j=0}^{3} P(x|s_j) P(S_j|\boldsymbol{N}) \big(\psi(S_j) - \hat{a} - \hat{b}x\big)^2 \quad (5.72)$$

5.5.3.4 Plug-in estimator of Q

Again, we provide two alternate methods for the calculation of the plug-in estimate of Q. The first has, in fact, already been used in the derivation of $\hat{b}$, since the denominator of $\hat{b}_{MLE}$ is Q. The second method is more complex and calculation intensive however it has the benefit that among the intermediate results are estimates of conditional Fisher information matrix for an individual. The matrix is conditional on the sub-population, so overall we have four individual Fisher information matrices. These matrices are very useful when we do power analysis, mostly for sample size determination. They can be used to quantify the change in the precision of the score function of the parameters that are associated with the effect size when we add an additional control, super-control, case or super-case.

Method 1

Here, we recycle the same idea we used to obtain $\hat{b}$ in section 5.5.3.2. The details and the rationale are fully explained there and it follows that the

plug-in estimator of Q is given by:

$$\hat{Q} = N \cdot E_x \left\{ [x - E_x(x|\boldsymbol{N})]^2 \, |\boldsymbol{N} \right\} = N \sum_{x=0}^{2} \sum_{j=0}^{3} P(x|S_j) P(S_j|\boldsymbol{N}) (x - \hat{\bar{x}})^2 \quad (5.73)$$

Method 2

In 5.54 we present the Fisher information matrix of $\phi = (a, b, \sigma^2)$ for all the available data. Here we consider the individual Fisher information matrix, $I_1(\phi)$; the Fisher information of ϕ for a single observation (y, x).

$$I_1(\phi) = \frac{1}{\sigma^2} \begin{pmatrix} 1 & x & 0 \\ x & x^2 & 0 \\ 0 & 0 & \frac{n}{2\sigma^2} \end{pmatrix} \quad (5.74)$$

We use the individual Fisher information to obtain an estimator for the expected value of the Fisher information $I(\phi)$ in our data. First we calculate the conditional, on the sub-population, expected value of individual Fisher information.

$$I_1^{S_j} := E_x(I_1(\phi)|S_j) = \frac{1}{\sigma^2} \begin{pmatrix} 1 & P(x=1|S_j) + 2P(x=2|S_j) & 0 \\ P(x=1|S_j) + 2P(x=2|S_j) & P(x=1|S_j) + 4P(x=2|S_j) & 0 \\ 0 & 0 & \frac{1}{2\sigma^2} \end{pmatrix} \quad (5.75)$$

Then we estimate the Fisher information of the data,

$$\hat{I}(\phi) = \sum_{j=0}^{3} N_j I_1^{S_j}. \tag{5.76}$$

We recall that $\text{Var}(\hat{b}) = [I^{-1}(\phi)]_{2,2} = \sigma^2/Q$. Therefore we can use $\widehat{\text{Var}(\hat{b})} = [\hat{I}^{-1}(\phi)]_{2,2} = \sigma^2/\hat{Q}$ in order to obtain the estimator $\hat{Q}$ of Q. We have that:

$$\frac{\sigma^2}{\hat{Q}} = \hat{I}^{-1}(\phi)]_{2,2} = \frac{\sigma^2 \sum_{j=0}^{3} N_j}{D} = \frac{\sigma^2 N}{D}, \tag{5.77}$$

where D is the determinant of $\hat{I}(\phi)$:

$$D = \left(\sum_{i=0}^{3} N_i\right)\left(\sum_{i=0}^{3} N_i \left[P(x=1|S_i) + 4P(x=2|S_i)\right]\right) - \left(\sum_{i=0}^{3} N_i \left[P(x=1|S_i) + 2P(x=2|S_i)\right]\right)^2 \tag{5.78}$$

Thus, overall, combining (5.77) and (5.78), the plug-in estimator of Q is given by:

$$\hat{Q} = \sum_{i=0}^{3} N_i \left[P(x=1|S_i) + 4P(x=2|S_i)\right] - \frac{\left(\sum_{i=0}^{3} N_i \left[P(x=1|S_i) + 2P(x=2|S_i)\right]\right)^2}{N} \tag{5.79}$$

5.5.3.5 Accuracy of the Asymptotic Method

In order to assess the asymptotic method we develop for calculating the power, we compare it with power calculated using simulations. We consider

both methods of calculating the plug-in estimators of σ^2 and Q. Figure 5.13 shows those comparisons. It is clear that the differences between the asymptotic and simulated power are imperceptible. Additionally both methods for either $\hat{\sigma}^2$ or $\hat{Q}$ appear to work equally well.

5.5.3.6 Choice of the Super-Cases Response

We use the asymptotic method to in order to choose the optimal value of γ for the super-cases response ($\psi(S_4)$). Equivalently to chose the score z_4 that gives the greater power when we apply the allelic trend test. We try different values of γ starting wiith $\gamma = 0$ and going up to $\gamma = 10$ (Figure 5.14). We can see that the best results are achieved when γ is slightly larger than zero. The exact optimal value appears to depend on the other parameters, specifically the MAF and disease prevalence K. If we choose a large γ, for example greater than one there is a large reduction in power compared to $\gamma = 0$. As a rule of thumb we propose choosing γ between 0.1 and 0.5.

5.5.4 Environmental and other Non-Genetic Factors

We showed in section 5.5.2 that whether we choose to use the allelic trend test or the linear model we describe in this chapter, the power of both methods in identifying markers that are associated with the disease is the same. However, an advantage of the linear model is that we can include environmental or other non-genetic effects in the model. When we use the allelic trend test, we can only to test for those effects using separate possibly independent

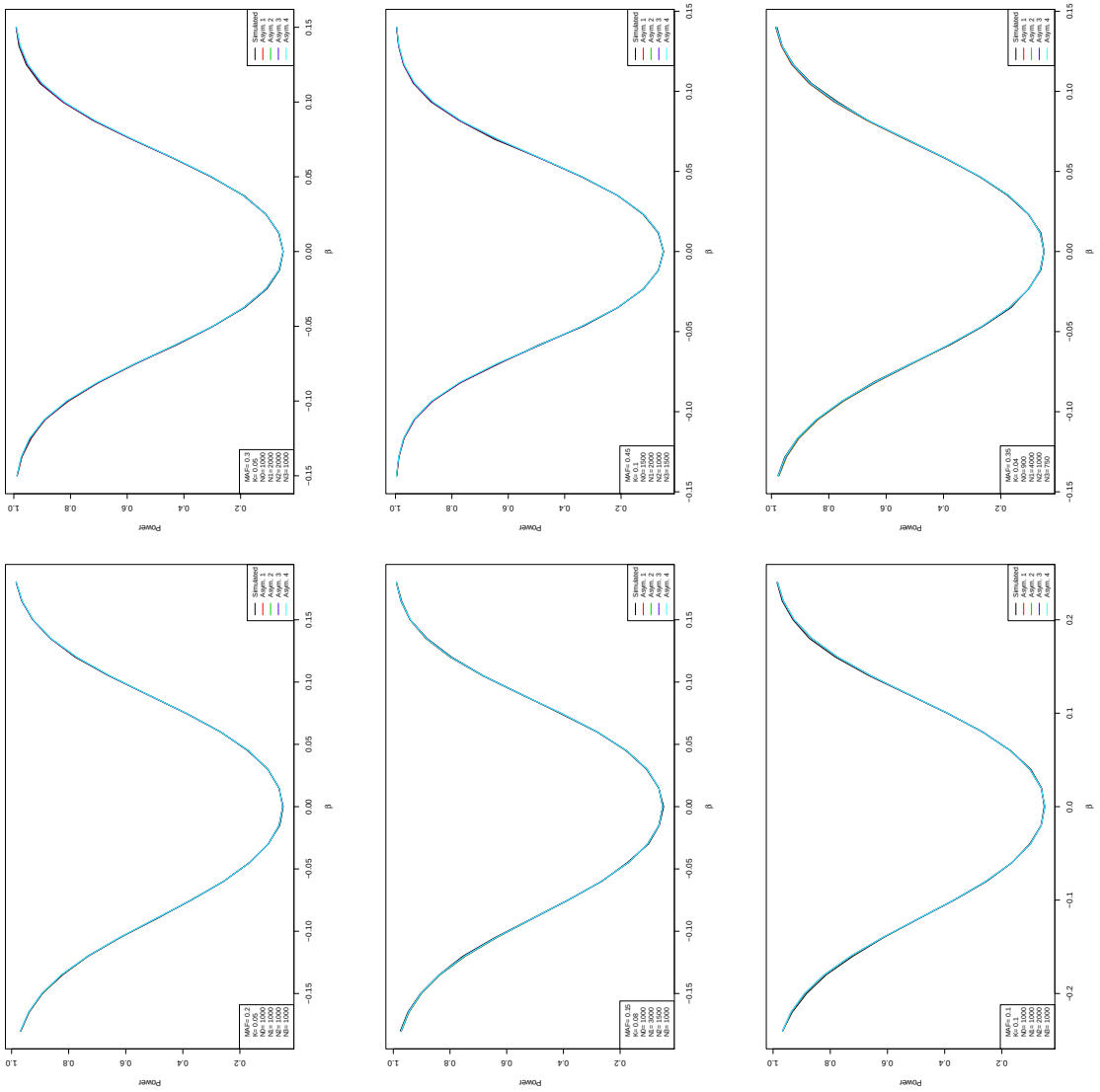


Figure 5.13: Comparison of asymptotic and simulated power. In asymptotic 1 we use the use the first method for the plug-in estimator of σ^2 and the first method for the plug-in estimator of Q. In asym. 2 we use the first method for $\hat{\sigma}^2$ and the second method for $\hat{Q}$; in asym.3 we use the second method for $\hat{\sigma}^2$ and the first method for $\hat{Q}$ and in asym. 4 we use the second methods for both $\hat{\sigma}^2$ and $\hat{Q}$.

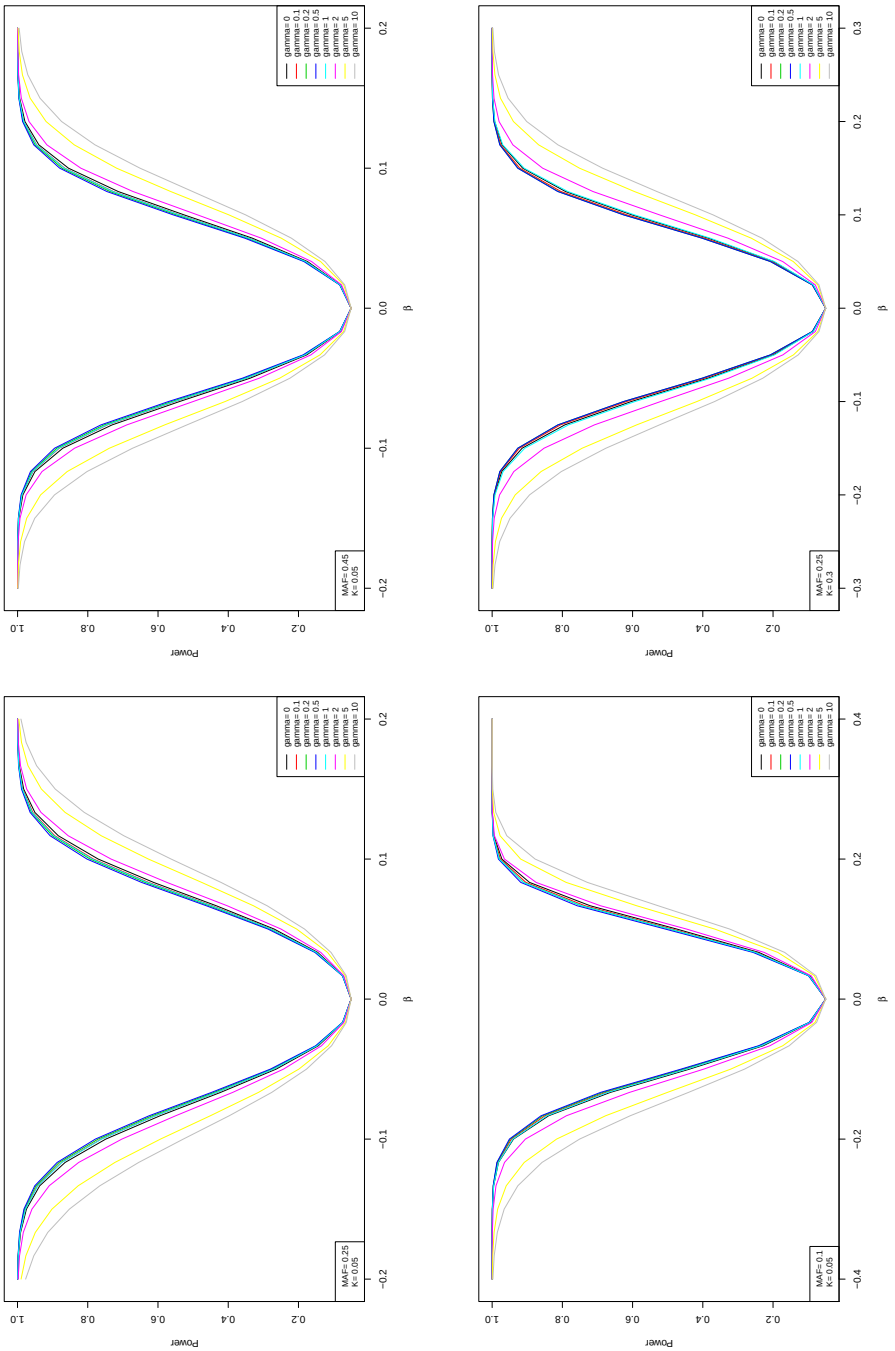


Figure 5.14: The power for different values of γ ($\psi(S_4) = 1 + \gamma$)

methods. A common example of those non-genetic factors, is to take into account the smoker status when we study lung cancer.

To demonstrate this, we consider a simple disease model, where the risk of disease is the sum of the genetic and external risk (Sepulveda et al., 2007) plus a possible risk due to interaction between genetic and external effects.

$$P(Y = 1|x, \xi) = f_x + h(\xi) + k(x, \xi) \tag{5.80}$$

In (5.80), f_x is the marginal probability of disease, conditional only on the genotype, $P(Y = 1|x)$; ξ is the non-genetic factor and h and k are appropriate risk functions. In order to simplify things we consider a binary ξ and set fixed coefficients β_h and β_k instead of the more general functions h, k. Thus we have,

$$P(Y = 1|x, \xi) = f_x + \beta_h I_{[\xi=1]} + \beta_k I_{[\xi=1]} xJ, \tag{5.81}$$

where J is a variable that takes the value -1 if the rare allele is protective and $+1$ if the rare allele is causal. Hence the model we fit in this case is,

$$\psi(S) = b_1 x + b_2 \xi + b_3 x\xi + \epsilon \tag{5.82}$$

We consider three cases of external effects. In the first case we consider only main effects for both the genetic and the non-genetic factors, i.e. $\beta_k = 0$. In the second case we consider that the non-genetic factor, affects the risk for disease only through interactions with the genetic factors, i.e. $\beta_h = 0$. In the third case we consider the full model in (5.81). That is, there are main

effects for both genetic and non-genetic factor plus an interaction term.

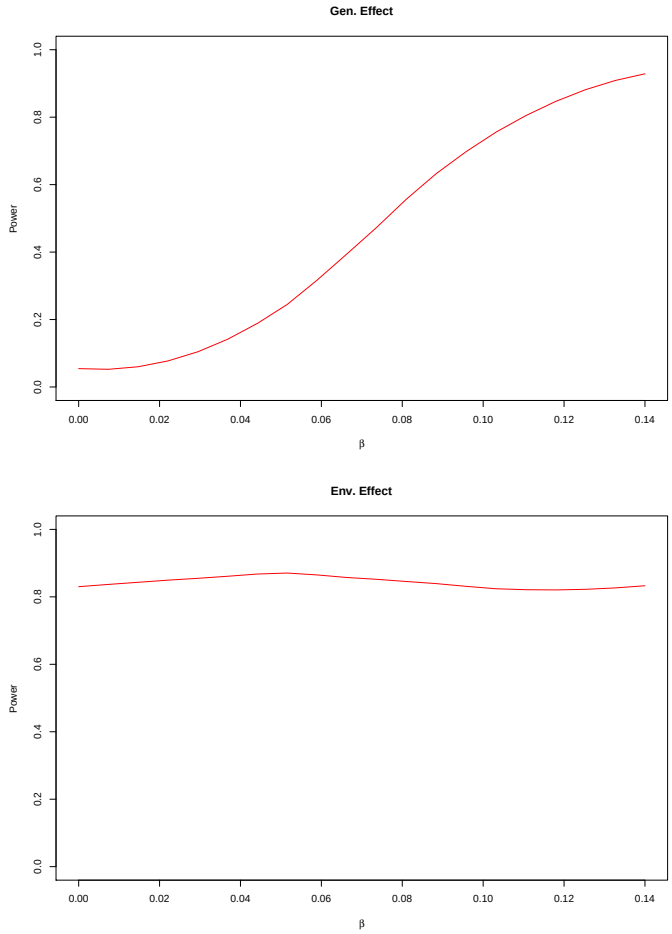


Figure 5.15: Power of the allelic trend test for the genetic effect (top) and the Pearson chi-square test for the non-genetic effect(bottom). The data generated with main effects only.

For comparison purposes, we also apply, in all cases, the allelic trend test to identify the genetic effect and also an independent Pearson chi-square test to identify the non-genetic effect. We assume that the genetic disease model (f_x) is the logistic and we calculate the power as a function of the penetrance parameter β of this model. In our applications we use $\beta_h = 0.01$ and $\beta_k = 0.009$ (in the cases which they are not zero). Also the frequency of

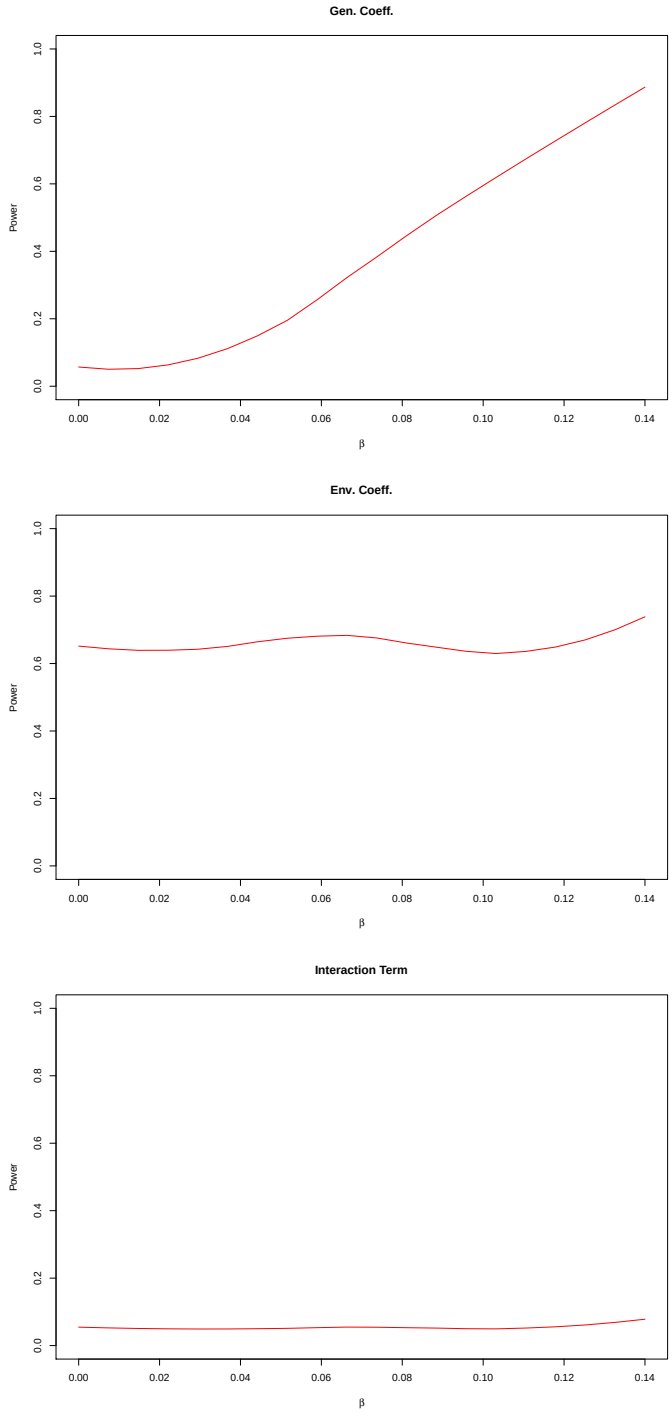


Figure 5.16: Power of the tests based on the coefficients of the linear model. Genetic effect (top); Non-genetic effect (middle); Interaction (bottom). The data were generated with main effects only.

the non-genetic effect in the population is independent of the genotype and we use $P(\xi = 1) = 0.3$.

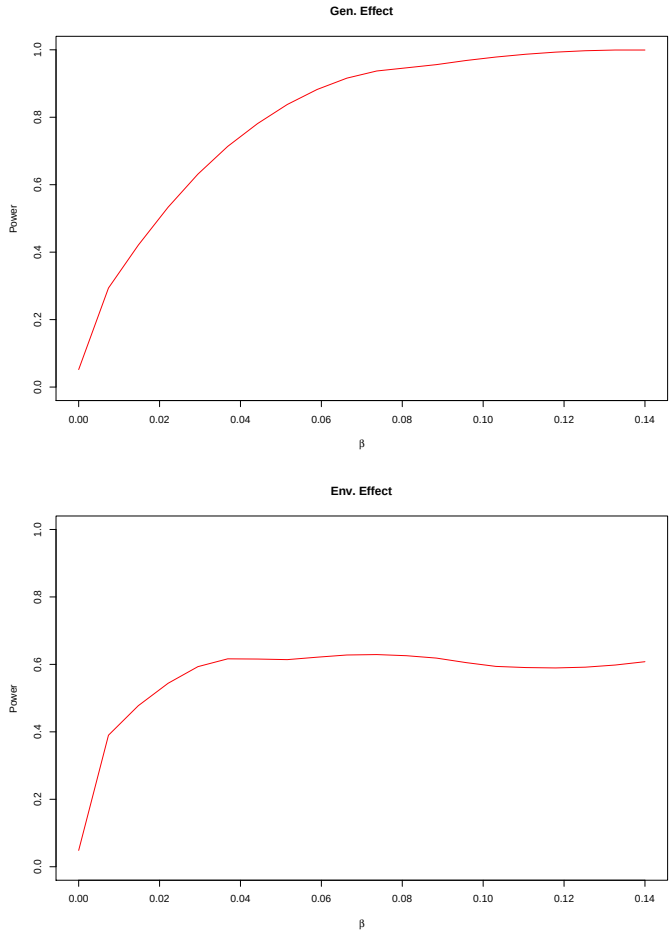


Figure 5.17: Power of the allelic trend test for the genetic effect (top) and the Pearson chi-square test for the non-genetic effect (Bottom). The data were generated with a main effect for the genetic term plus an interaction term between genetic and environmental effects.

When there are no interactions (Figures 5.15 and 5.16), using the linear model does not offer much, on the contrary, the power of the separate chi-square tests is greater than the power of the tests based on the fitted coefficients of the model. However when we include an interaction term in the

generative model of the data, either without main effect for the non-genetic factor (Figures 5.17 and 5.18) or when that factor has a separate main effect (Figures 5.19 and 5.20), then the linear model can help us to identify the type of those effects. We can see that in figure 5.18 the linear model offers clear evidence that the non-genetic factor affects the risk of disease only through interaction with the genotype. Similarly, figure 5.20 shows that is very likely to identify both the main effect and the interaction term when those exist in the generative model.

In general, the power of the chi-square tests is always greater than the power of the tests based on the cofficients of the linear model. Therefore, if we simply want to test whether an external, non-genetic factor, affects the risk of disease but we are not interested in the type of this effect, then it is preferable to use a marginal test, e.g. the Pearson's chi-square, focusing only on that factor. However if we are interested on identifying the type of this effect, e.g. main effect, interaction or both, then the linear model is the method that should be used.

5.6 Discussion

In this chapter we study the Super-Case–Case–Control–Super-Control design. Here, the additional data are not as easily obtainable as in the other designs we discuss in this thesis. We assume that the data which are originally available are the super-cases and super-controls. The additional control sample is relatively easy to obtain and it is essentially the same sample as the

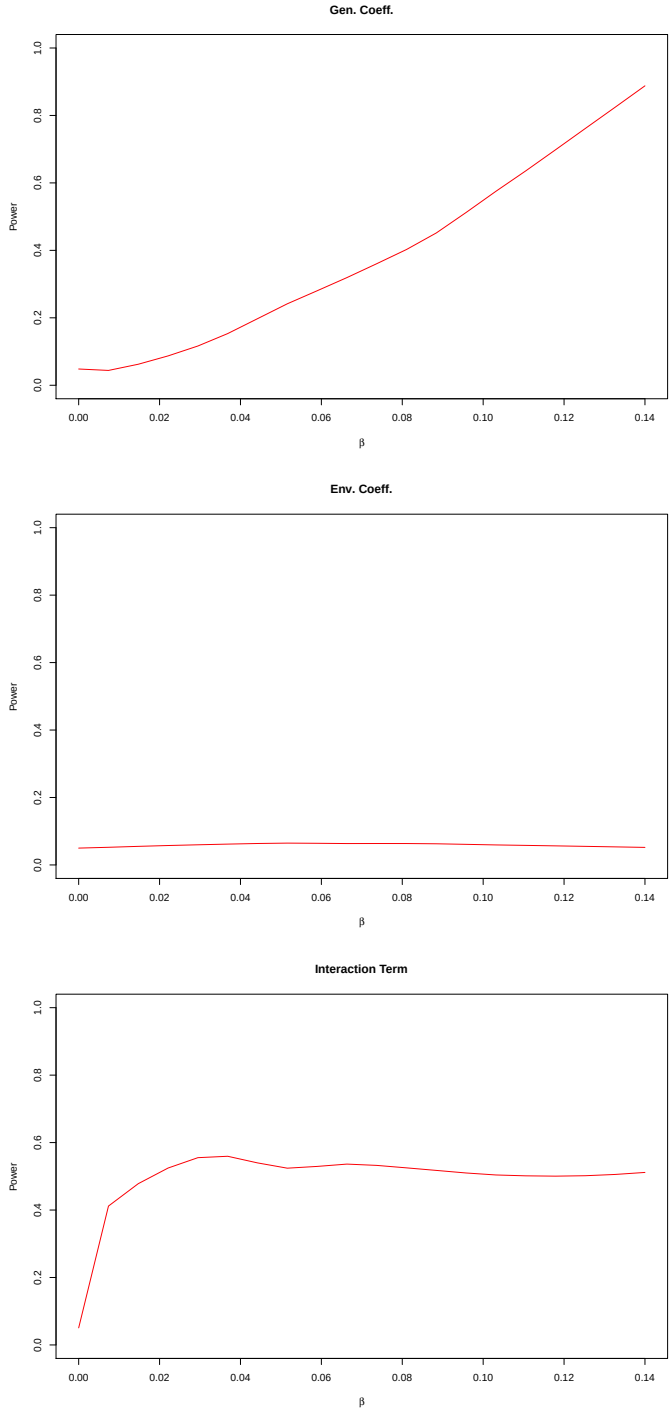


Figure 5.18: Power of the tests based on the coefficients of the linear model. Genetic effect (top); Non-genetic effect (middle); Interaction (bottom). The data were generated with a main effect for the genetic term plus an interaction term between genetic and environmental effects.

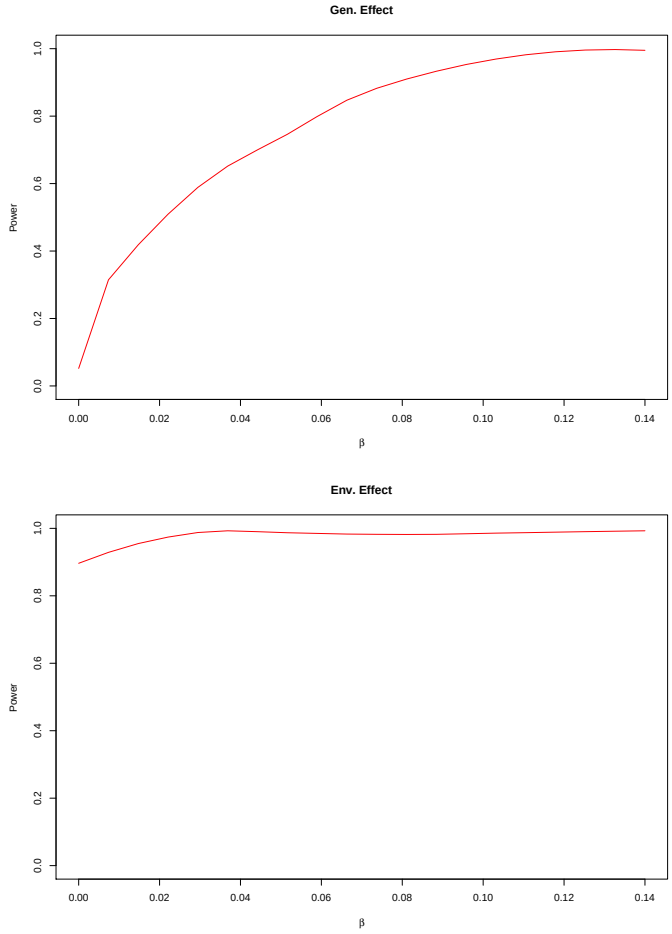


Figure 5.19: Power of the allelic trend test for the genetic effect (top) and the Pearson chi-square test for the non-genetic effect (bottom). The data generated with both genetic and environmental main effects plus their interaction term.

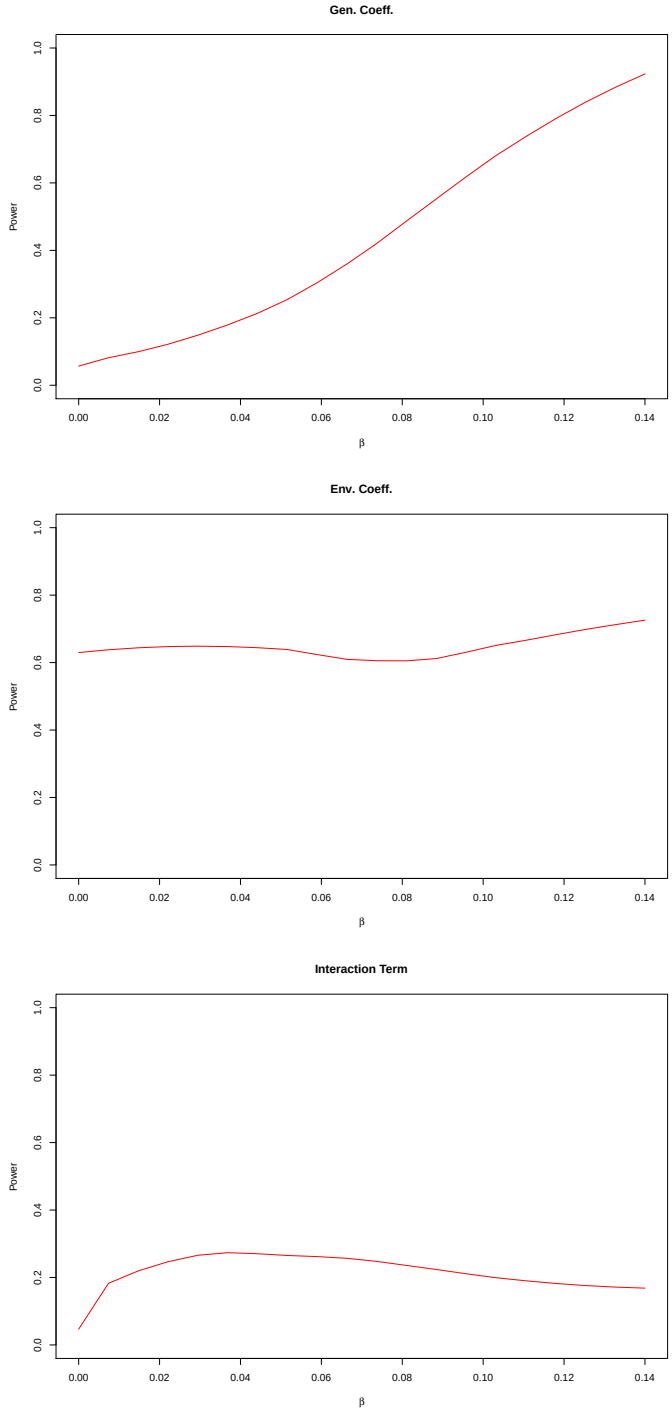


Figure 5.20: Power of the tests based on the coefficients of the linear model. Genetic effect (top); Non-genetic effect (middle); Interaction (bottom). The data generated with both genetic and environmental main effects plus their interaction term.

cohort we describe in chapter 3. The difficult part is to obtain the additional case samples. These may possibly be data from older studies of the disease where not much attention was paid in securing strong presence of the disease within the families of the cases.

We study three new methods in addition to the standard Cochran-Armitage test, which we adjust by merging the case/super-cases and controls/super-controls samples. All methods are based on a property of the conditional minor allele frequencies under the hypothesis of genetic association. Specifically, the order $\theta_0 < \theta_1 < \theta_2 < \theta_3$ holds when the allele is causal and the opposite direction of inequalities when it is protective.

The first method we consider is a set of two Bayesian models, that differ in their prior specifications. In our examples and applications we used non-informative prior distributions. However, given the subjective nature of Bayesian methods, those models can become much more useful and efficient in the precence of some prior information which can allow us to use - not arbitrarily - tighter priors.

The other two methods we consider is an allelic trend test and the simple linear model. The former is derived from the standard Pearson chi-square test. Essentially, under certain conditions, the two methods have equal power in identifying genetic associations. We also provide an asymptotic formula for the calculation of said power. The advantages of using the linear model is discussed in the last section of this chapter and concerns the abililty to take into account non-genetic factors when we perform a genetic association study.

Chapter 6

Discussion and Future Work

The obvious answer to the question "how can I increase the power of a test used in a genome-wide association study?", is to get more samples by genotyping additional individuals. However, in practice this is still a relative expensive procedure. In this thesis we provide alternatives, where by using relatively easy to obtain and non-costly additional data, we can extend the traditional case-control design that is used in GWAS. We discuss in detail three such extensions, the case-cohort-control design, the kin-cohort design and the super-case–case–control–super-control design as well as provide appropriate methods that can be used under each of those designs. Those methods are usually modified standard methods that are used in the case-control design. A brief summary about each of those designs is provided at the end of the relevant chapters.

In this thesis we focus primarily on method development and assessment. However, in our study we also applied our methods to real genetic data,

where we unfortunately came across issues that prevented us from including those applications in this thesis. The most notable issue was that of population heterogeneity, or in other words incompatibility between the different samples, where instead of disease associations the methods identify population differences. This problem was the subject and what gave rise to the idea behind section 3.3. We would like to re-visit this problem and search for methods and approaches to resolve it, without the need to use a reference results set as we did in our study.

The same problem was present in our applications under the SCCS design. Here, the CORGI cases and controls were taken to be the super-cases and super-controls in the design while the BC58 data were taken to be the controls. For cases in our design, we used the VICTOR-QUASAR[1] (VQ) data (Tomlinson et al., 2008). Unfortunately VQ data were far worse in terms of compatibility with CORGI than BC58 were. We tried to address the problem with principal components analysis (PCA). Specifically when we applied the linear model (section 5.5) in those data, we tried to include all the combinations of the first three principal components (PCs) of the data, as additional explanatory variables in the model. The idea was that population differences will be explained by the PC terms, thus letting the genetic term explain only potential associations with the disease. Even though graphical PCA showed that there were obvious population differences between the samples, the coefficients of the PCs in the models were always not significant (equal to zero) and thus had no effect in the model. Again, this is a challenging subject for future work since in most practical applications, such issues are very likely

[1]http://www.octo-oxford.org.uk/

to arise.

Another problem that prevented us from applying our methods to real data was the very lack of appropriate data. We developed the methodology for the kin-cohort design in chapter 4 under the assumption that we will be able to apply those methods on the CORGI data. To our disapointment, only about half of the CORGI cases had familial information (pedigrees) and even not all of those were the type of pedigrees we describe in our method. Finding that type of data and being able to apply our methods is one of our objectives for the future.

Another thing we would like to include in our methods is the age of onset. All the methods we present in this thesis assume that the individuals in the control and super-control samples are at no risk to develop the disease in the future. It would be more realistic to take age of onset into account by including a time variable in our methods. Antoniou and Easton (2003) adopt a design that can be consider a hybrid of our kin-cohort and SCCS designs. The authors study genetic associations of the breast cancer and use mother-daughter data. They take age of onset into account by adopting a proportional hazards model for the breast cancer incidence rate.

In chapter 4 we developed a Bayesian hierarchical model for inference in GWAS with familial data where the analysis is carried out by implementing an MCMC algorithm. Something that needs to be addressed, which is not something specific to our model but it is inherent in the MCMC methods themselves, is that this approach is extremely computationally expensive in

terms of time. Using the R software[2], we needed roughly two days for a single run of the algorithm (20000 iterations). Potentially, a more optimised code written in a different computer language may improve those computer times. An alternative approach is to adjust our model so that the analysis can be carried out by using an expectation-maximisation (EM) algorithm. Considering that the EM algorithm is well suited for problems with missing data (family members' genotypes) while usually being considerably faster than the MCMC algorithms make its implementation an intersting and promising prospect worth exploring. The developement of faster methods will also enable us to study another issue we encountered in our analysis. Specifically the issue of potential bias for the MAP estimates of the effect size in the kin-cohort design. Even though we believe that with the introduction of the relevance weights we remove the bias from the estimators, we were not able to prove that. We could not do it analytically because we do not have a convenient form of the conditional distribution of β, hence the independent Metropolis-Hastings step in our algorithm. Extensive simulations can be used to resolve this issue however the computational time required makes this impractical. Therefore a faster algorithm or method to obtain the MAP estimates of β will have the added benefit that it will allow us to study the bias properties of the estimators more thoroughly.

Also worth trying, is the use of polytomous logistic regression under the SCCS design in the same vein we use the simple linear model (Ananth and Kleinbaum, 1997; Engel, 1988). We did some preliminary analyses in our research, considering both nominal and ordinal responses but we chose to

[2]http://cran.r-project.org/

focus on the more simple ordinary linear regression. Given that the logistic disease model is the most commonly assumed model in GWAS, polytomous logistic regression appears to be more suitable and natural model for the data under the SCCS design. The main problem we faced when we were considering this approach was the interpretation of the fitted model. Both ordinal, but especially the nominal, logistic regression models fit a relatively a large number of coefficients and it is not clear which of those are indicative of a genetic effect when they are deemed significant in the model. Nevertheless, polytomous logistic regression is something that we keep in mind for the future.

Appendix A

Proofs

Odds ratio and Relative risk for the Logistic disease model

$$OR = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\left(\frac{1}{1+\exp(\alpha+\beta x)}\right)^{-1} / \frac{\exp(\alpha)}{1+\exp(\alpha)}\left(\frac{1}{1+\exp(\alpha)}\right)^{-1} = \frac{\exp(\alpha+\beta)}{\exp(\alpha)} = \exp(\beta)$$

$$RR = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\left(\frac{\exp(\alpha)}{1+\exp(\alpha)}\right)^{-1} = \frac{\exp(\alpha+\beta)+\exp(2\alpha+\beta)}{\exp(\alpha)+\exp(2\alpha+\beta)} = \frac{\exp(\beta)\exp(\alpha+\beta)}{1+\exp(\alpha+\beta)}$$

Minor Allele Frequency: Property 1

We show that the population MAF, θ, can be written as a linear combination of the conditional MAFs of the uffected θ_p and unaffected sub-populations.

$$\begin{aligned}
\theta &= \frac{1}{2}E(x) = \frac{1}{2}P(x=1) + P(x=2) = \\
&\sum_{i=0,1}\left[\frac{1}{2}P(x=1, Y=i) + P(x=2, Y=i)\right] = \\
&\sum_{i=0,1}\left[\frac{1}{2}P(x=1|Y=i)P(Y=i) + P(x=2|Y=i)P(Y=i)\right] = \\
&\frac{1}{2}P(x=1|Y=0)P(Y=0) + P(x=2|Y=0)P(Y=0) + \\
&\frac{1}{2}P(x=1|Y=1)P(Y=1) + P(x=2|Y=1)P(Y=1) = \\
&(1-K)\left[\frac{1}{2}P(x=1|Y=0) + P(x=2|Y=0)\right] + \\
&K\left[\frac{1}{2}P(x=1|Y=0) + P(x=2|Y=0)\right] = \\
&\frac{1-K}{2}E(x|Y=0) + \frac{K}{2}E(x|Y=1) = (1-K)\theta_q + K\theta_p
\end{aligned}$$

Minor Allele Frequancy: Property 2

We show that the value of the population MAF, θ, lies always between the values of the conditional MAFs θ_p and θ_q.

Based on the result in Property 1:

$$\theta = (1-K)\theta_p + K\theta_q \tag{A.1}$$

But also:

$$\theta = (1 - K)\theta + K\theta \tag{A.2}$$

Combining A.1 and A.2, it is obvious that if $\theta_q < \theta$ then necessarily $\theta < \theta_p$ and vice versa.

Derivation of the relevance weighted likelihood

The derivation of the relevance weights and their use in the likelihood is based on the works published by Hu et al. (2000) and Hu (1997). In the latter paper the relevance weighted likelihood is defined as a method that weights individual contributions to the likelihood differently, according to their relevance in some sence. Therefore – and here we use the notation introduced in chapter 4 – in the unweighted likelihood

$$L(\alpha, \beta) = \prod_{i=1}^{N} p_{y_i^{(m)}}(x_i^{(m)}, \alpha, \beta) p_{y_i^{(p)}}(x_i^{(p)}, \alpha, \beta) \prod_{j=1}^{n_i} p_{y_i^{(s_j)}}(x_i^{(s_j)}, \alpha, \beta) \tag{A.3}$$

all individuals contribute equally and thus have the same relevance. What causes the need, in our case, to consider different contribution for each individual in the likelihood is the sampling scheme. In a simple random sampling scheme all individuals have equal contributions. Our sampling scheme is simple random only when it comes to the probands' genotypes. The genotypes of other family members are sampled conditional on the probands' genotype. The contribution of each individual in the likelihood should be chosen so that the relevance weighted likelihood will emulate an unweighted likelihood that

is made up of observations collected from a simple random sampling scheme.

If $\boldsymbol{S}$ denotes the simple random sampling scheme and $\boldsymbol{S}^*$ the sampling scheme, then in order the emulate the former in a likelihood comprised of observations collected using the former the contribution of an observation (x, y) should be:

$$w = \frac{P(x|\boldsymbol{S})}{P(x|\boldsymbol{S}^*)}. \tag{A.4}$$

Note that given the genotype, the phenotype and the sampling scheme are conditionally independent. The weights in equation (4.2) result directly from (A.4).

Now, the contribution of (x, y) in the likelihood is incorporated in the likelihood by substituting the term $p_y(x, \alpha, \beta)$ with the term $p_y(x, \alpha, \beta)^w$. Combining all the above, we obtain the log-likelihood presented in equation (4.3).

Derivation of r_x

$$\begin{aligned} r_x := P(x^{(1)} = x|S_3) &= P(x^{(1)} = x|y^{(1)} = 1, y^{(2)} = 1) \\ &= \frac{f_x P(x^{(1)} = x \mid y^{(2)} = 1)}{\sum_{x=0}^{2} f_x P(x^{(1)} = x \mid y^{(2)} = 1)} \end{aligned}$$

Now we just need $P(x^{(1)} = x|y^{(2)} = 1)$. We use the notion of identity by

descent (IBD).

$$\begin{aligned} P(x^{(1)} = x \mid y^{(2)} = 1) &= \sum_{i=0}^{2} P(x^{(1)} = x \mid y^{(2)} = 1 \wedge \text{share } i \text{ alleles IBD}) \times \\ &\quad P(\text{share } i \text{ alleles IBD}) \\ &= \frac{1}{4} g_x + \frac{1}{2} P(x^{(1)} = x \mid y^{(2)} = 1 \wedge \text{share 1 allele IBD}) + \frac{1}{4} p_x \end{aligned}$$

Now,

$$\begin{aligned} & P(x^{(1)} = 0 \mid y^{(2)} = 1 \wedge \text{share 1 allele IBD}) \\ = \; & P(\text{non-IBD allele is not MA} \wedge \text{IBD allele is not MA} \mid y^{(2)} = 1) \\ = \; & P(\text{non-IBD allele is not MA}) P(\text{IBD allele is not MA} \mid y^{(2)} = 1) \\ = \; & (1 - \theta_1) P(\text{IBD allele is not MA} \mid y^{(2)} = 1) \\ = \; & (1 - \theta_1) \sum_{j=0}^{2} P(\text{IBD allele is not MA} \mid x^{(1)} = j) P(x^{(1)} = j | y^{(2)} = 1) \\ = \; & (1 - \theta_1) \left(P(x^s = 0 | z^s = 1) + \frac{1}{2} P(x^s = 1 | z^s = 1) + 0 \times P(x^{(1)} = 2 | y^{(2)} = 1) \right) \\ = \; & (1 - \theta_1) \left(p_0 + \frac{p_1}{2} \right) \end{aligned}$$

where MA is the minor (rare) allele.

Similarly,

$$\begin{aligned} P(x^{(1)} = 1 \mid y^{(2)} = 1 \wedge \text{share 1 allele IBD}) &= \theta p_0 + \frac{p_1}{2} + (1 - \theta_1) p_2 \\ P(x^{(1)} = 2 \mid y^{(2)} = 1 \wedge \text{share 1 allele IBD}) &= \theta_1 \left(\frac{p_1}{2} + p_2 \right) \end{aligned}$$

Thus overall,

$$P(x^{(1)} = x \mid y^{(2)} = 1) = \frac{1}{4}g_x + I_{[x=0]}(1-\theta_1)\left(p_0 + \frac{p_1}{2}\right) + I_{[x=1]}\left(\theta_1 p_0 + \frac{p_1}{2} + (1-\theta_1)p_2\right) + I_{[x=2]}\theta_1\left(\frac{p_1}{2} + p_2\right) + \frac{1}{4}p_x$$

Justification of equation 5.14

We show this result for the case of a causal minor allele ($\pi(\theta|H_1^c)$). The derivation of $\pi(\theta|H_1^p)$ is equivalent. We consider the prior distributions in (5.12). When we adopt non-informative priors we use $a_i = b_i = 1$, $i = 0, 1, 2, 3$.

Now, we know that a beta distribution with both parameters equal to 1 is identical to the continuous uniform distribution in the interval $(0, 1)$.

$$\text{Beta}(1, 1) \overset{\mathcal{D}}{\equiv} U(0, 1)$$

Furthermore we can easily see that a truncated beta distribution with $a = b = 1$ and zero density outside (l, u) is identical to the continuous uniform distribution in the interval (l, u). To prove this, we use the definition of the truncated-beta distribution (Equations 5.9 and 5.10).

$$X \sim \text{TrBeta}(1, 1, l, u) \overset{\mathcal{D}}{\equiv} c(1, 1, l, u)\text{Beta}(1, 1)\mathcal{I}_{[l \leq X \leq u]} \overset{\mathcal{D}}{\equiv}$$

$$c(1, 1, l, u)U(0, 1)\mathcal{I}_{[l \leq X \leq u]}$$

Now, the p.d.f of $U(0,1)$ is 1. Equation 5.10 gives: $c(1,1,l,u) = (u-l)^{-1}$. Therefore the probability density function of the truncated-beta distribution $f_{TrB}(X)$ can be written as,

$$f_{TrB}(X) = (u-l)^{-1} \cdot 1 \cdot \mathcal{I}_{[l \leq X \leq u]} = \frac{1}{u-l}\mathcal{I}_{[l \leq X \leq u]} \equiv f_{U(l,u)}(X),$$

where $f_{U(l,u)}$ denotes the c.d.f of the U(l,u) distribution.

Using those properties and substituting the parameters of the non-informative priors in (5.12):

$$\begin{aligned}
\pi(\theta_0|\theta_1; H_1^c) &= \frac{1}{\theta_1}\mathcal{I}_{[0 \leq \theta_0 \leq \theta_1]} \\
\pi(\theta_1|H_1^c) &= 1 \times \mathcal{I}_{[0 \leq \theta_1 \leq 1]} \\
\pi(\theta_2|\theta_1; H_1^c) &= \frac{1}{1-\theta_1}\mathcal{I}_{[\theta_1 \leq \theta_2 \leq 1]} \\
\pi(\theta_3|\theta_2; H_1^c) &= \frac{1}{1-\theta_2}\mathcal{I}_{[\theta_2 \leq \theta_3 \leq 1]}
\end{aligned}$$

Substituting the above in (5.11) we obtain the result we want to prove:

$$\pi(\boldsymbol{\theta}|H_1^c) = \frac{1}{\theta_1}\frac{1}{1-\theta_1}\frac{1}{1-\theta_2}\mathcal{I}_{[\theta_0 < \theta_1 < \theta_2 < \theta_3]}$$

The Normalising Constant of the Joint Constrained-Uniform Distribution

We want to show thet $c \equiv c(M) = M!$, where,

$$c(M) = \left[\int_0^1 \int_0^{Z_M} \dots \int_0^{Z_2} dZ_1 \dots dZ_{M-1} dZ_M\right]^{-1} = \left[\int_0^1 \int_{Z_1}^1 \dots \int_{Z_{M-1}}^1 dZ_M dZ_{M-1} \dots dZ_1\right]^{-1} \quad \text{(A.5)}$$

We proceed to directly calculate the multiple integral:

$$\begin{aligned}
c(M) &= \left[\int_0^1 \int_0^{Z_M} \dots \int_0^{Z_2} dZ_1 \dots dZ_{M-1} dZ_M\right]^{-1} = \\
&\left[\int_0^1 \int_0^{Z_M} \dots \int_0^{Z_3} Z_2 \quad dZ_2 \dots dZ_{M-1} dZ_M\right]^{-1} = \\
&\left[\int_0^1 \int_0^{Z_M} \dots \int_0^{Z_4} \frac{(Z_3)^2}{2} dZ_3 \dots dZ_{M-1} dZ_M\right]^{-1} = \\
&\left[\int_0^1 \int_0^{Z_M} \dots \int_0^{Z_5} \frac{(Z_4)^3}{2 \cdot 3} dZ_4 \dots dZ_{M-1} dZ_M\right]^{-1} = \dots \\
&\dots \left[\int_0^1 \int_0^{Z_M} \dots \int_0^{Z_k} \frac{(Z_{k-1})^{k-2}}{(k-2)!} dZ_k \dots dZ_{M-1} dZ_M\right]^{-1} = \dots \\
&\dots \left[\int_0^1 \int_0^{Z_M} \frac{(Z_{M-1})^{M-2}}{(M-2)!} dZ_{M-1} dZ_M\right]^{-1} = \left[\int_0^1 \frac{(Z_M)^{M-1}}{(M-1)!} dZ_M\right]^{-1} = \\
&\left(\frac{1^M - 0^M}{M \times (M-1)!}\right)^{-1} = \left(\frac{1}{M!}\right)^{-1} = M! \qquad QED
\end{aligned}$$

Order of the Conditional MAFs for the four sub-populations in the SCCS design.

Figure A.1 shows some examples on how the conditional MAFs of the four sub-populations in the SCCS design are ordered when the rare allele is protective ($\beta < 0$; $\theta_0 > \theta_1 > \theta_2 > \theta_3$) and causal ($\beta > 0$; $\theta_0 < \theta_1 < \theta_2 < \theta_3$).

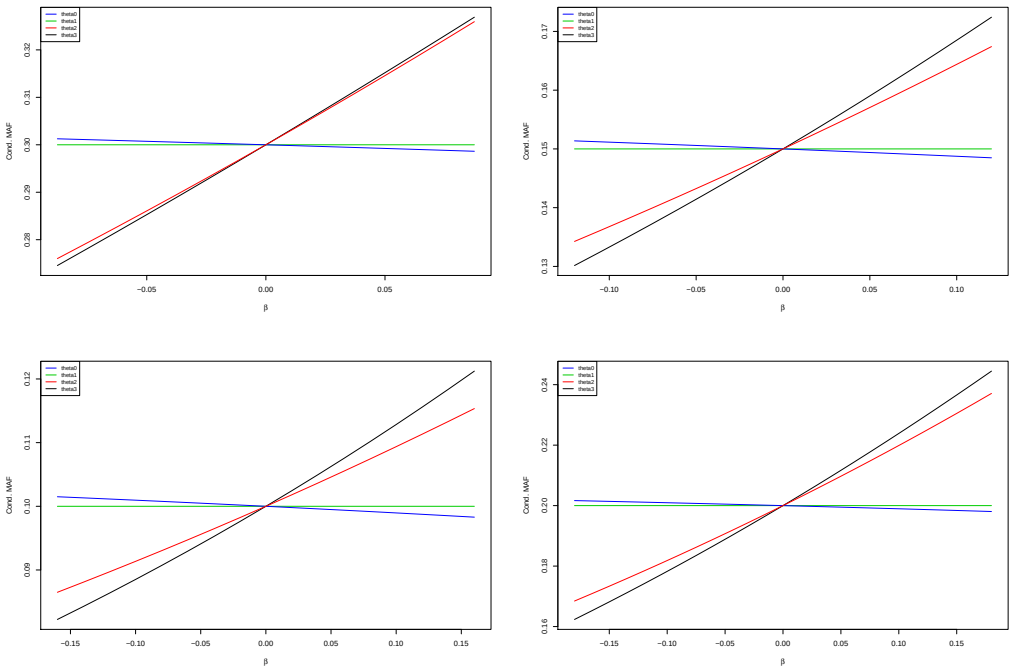


Figure A.1: Some examples of conditional Minor Allele Frequencies for the four sub-populations in the SCCS design. Blue: Super-controls; Green: Controls; Red: Cases; Black: Super-cases.

Appendix B

Additional Figures For Section 3.2

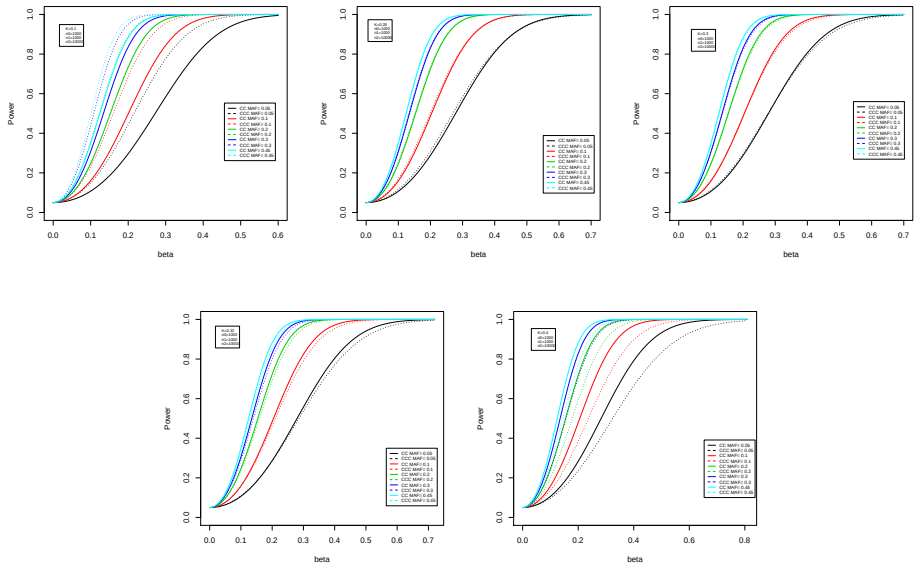


Figure B.1: Power-curves (logistic disease model)

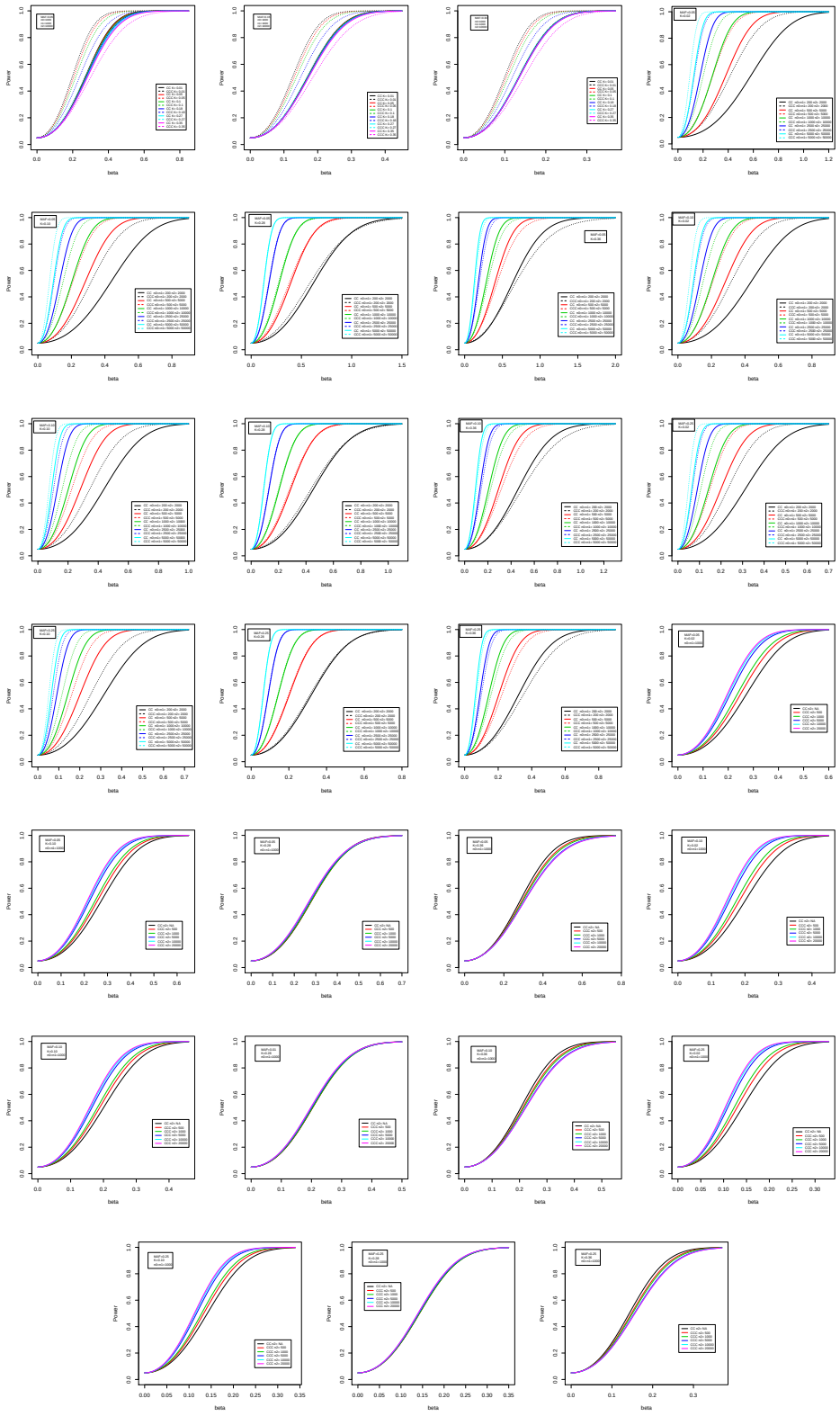


Figure B.2: Power-curves (logistic disease model - cont'd)

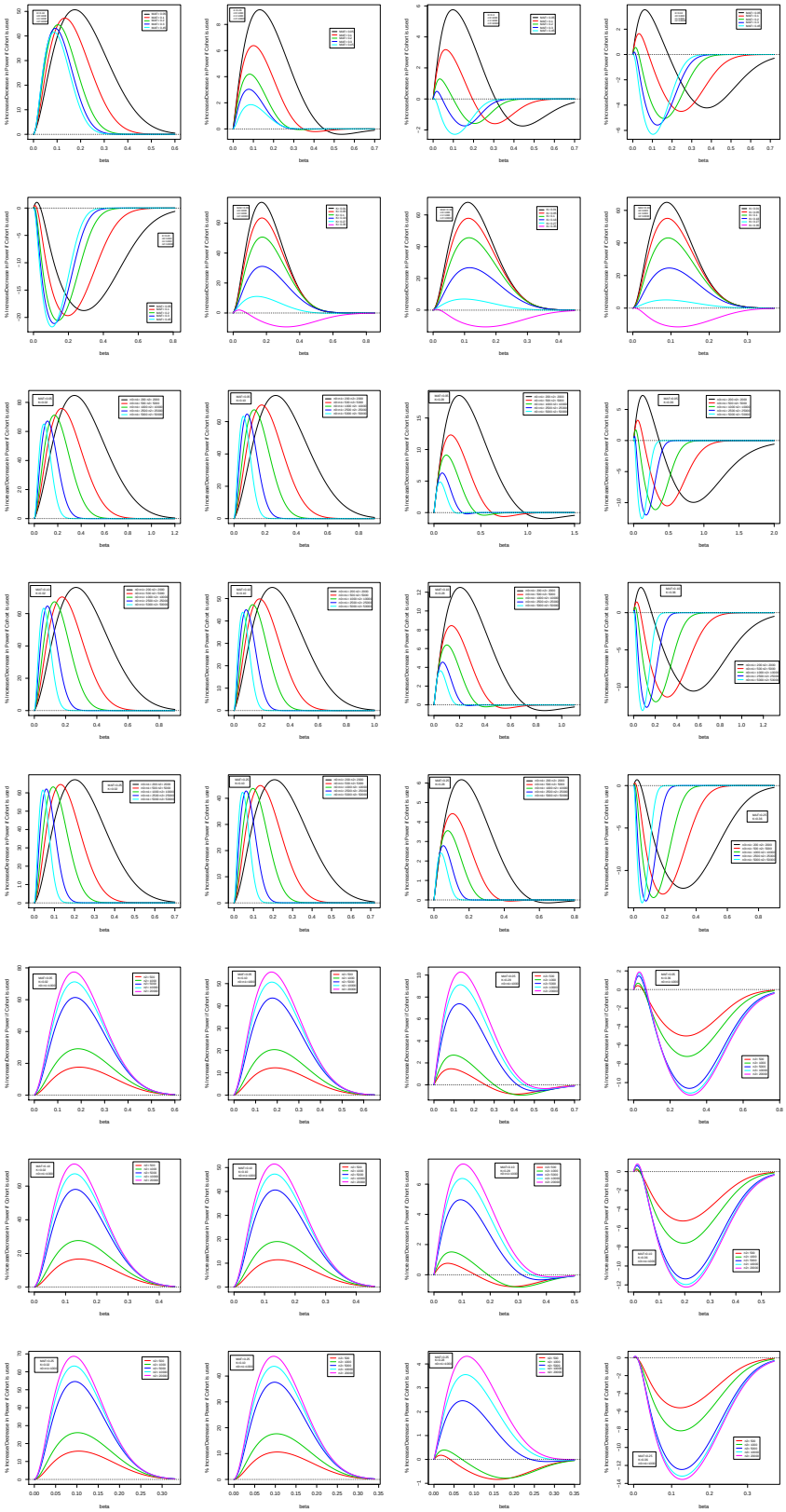


Figure B.3: $\gamma(\beta|\cdot)$-curves (logistic disease model)

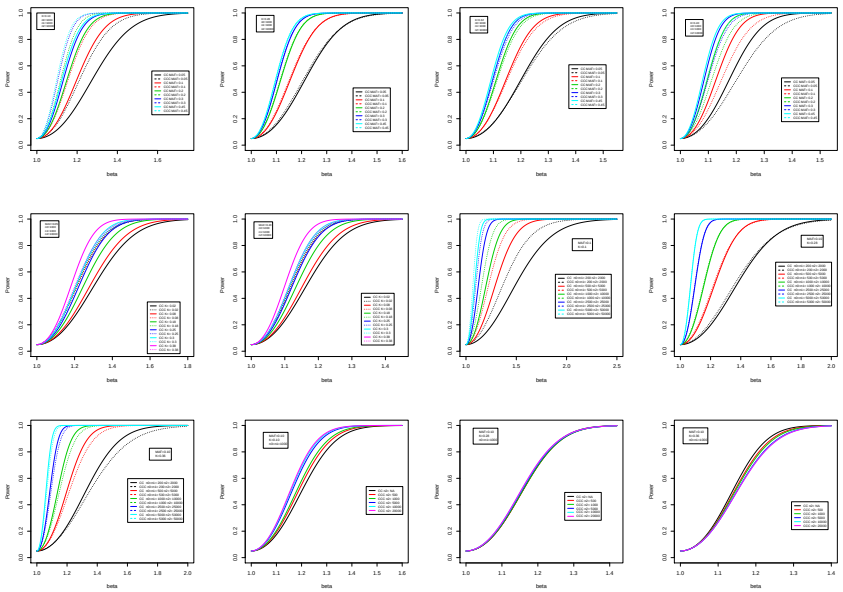


Figure B.4: Power-curves (additive disease model)

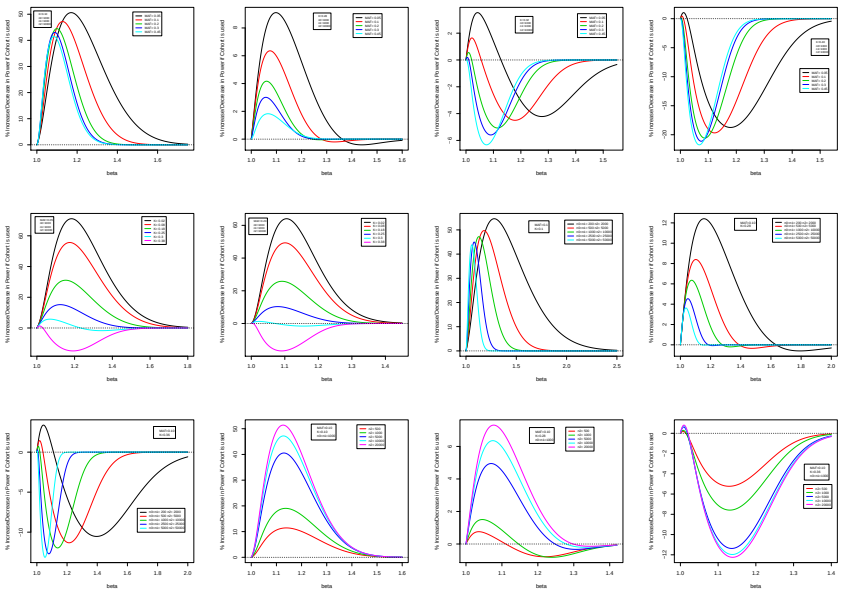


Figure B.5: $\gamma(\beta|\cdot)$-curves (additive disease model)

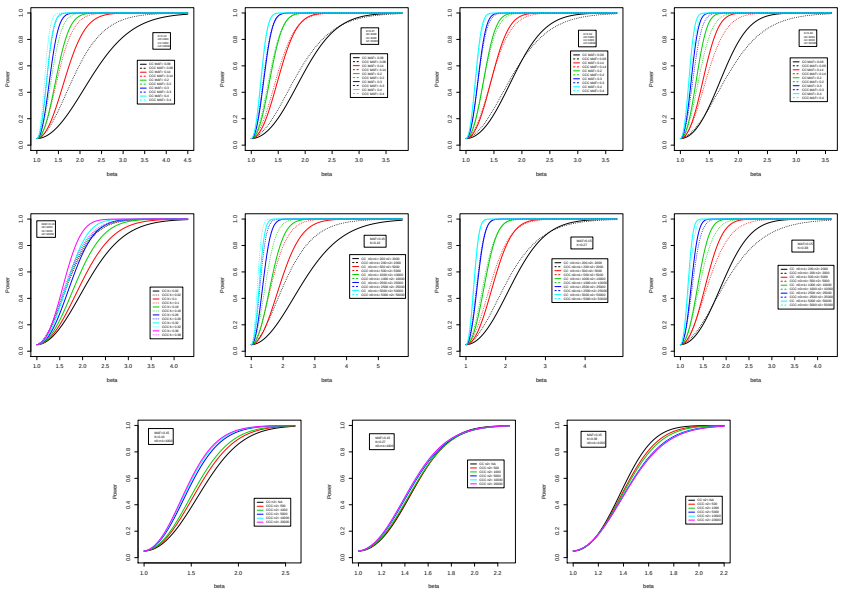


Figure B.6: Power-curves (recessive disease model)

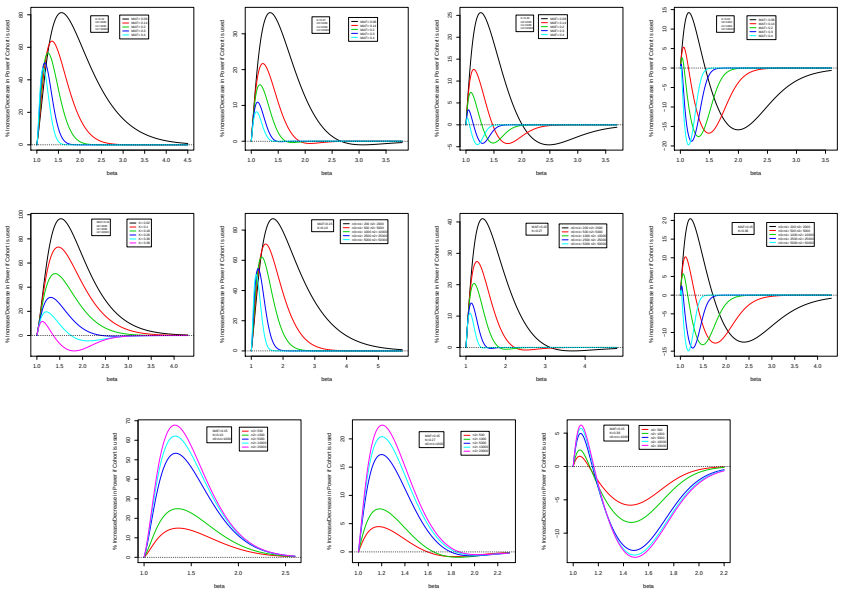


Figure B.7: $\gamma(\beta|\cdot)$-curves (recessive disease model)

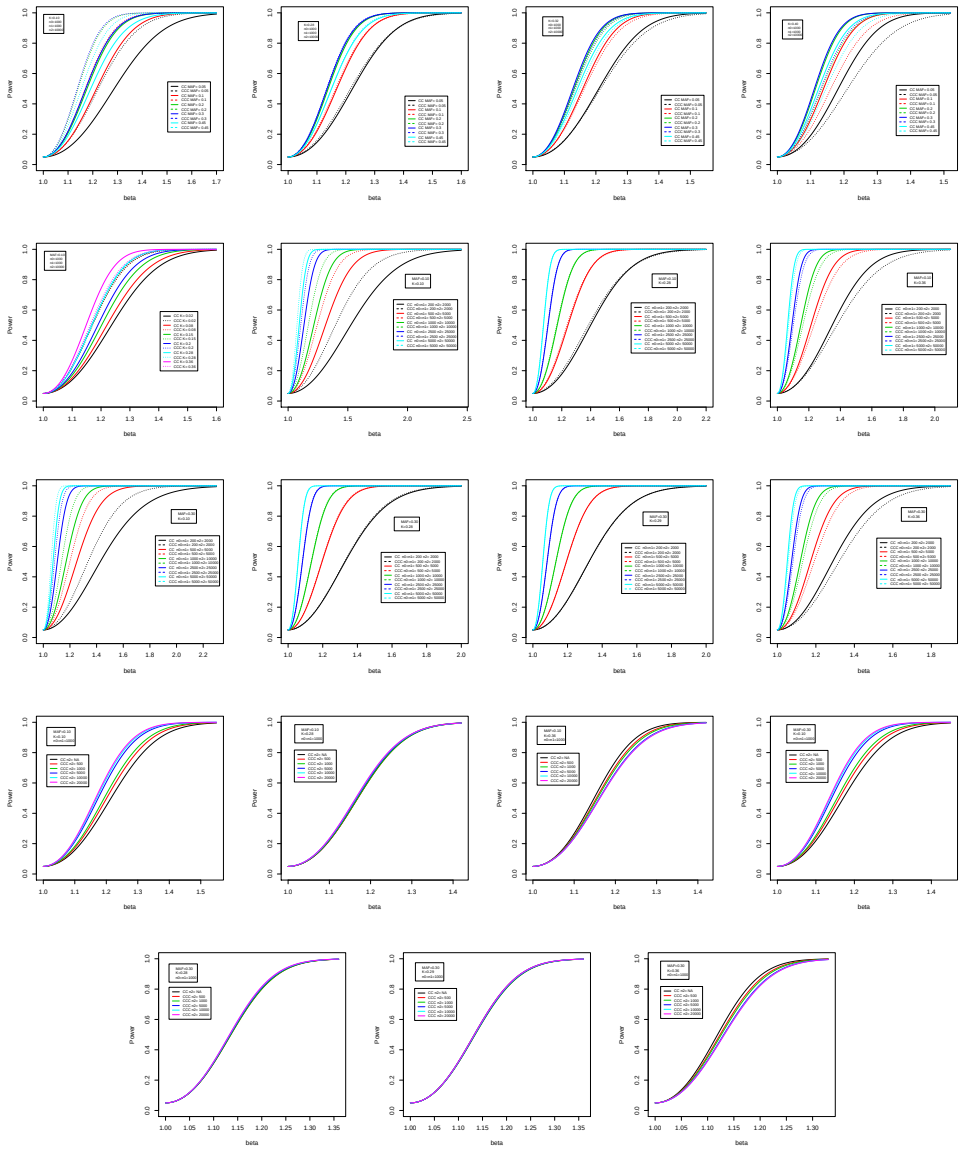


Figure B.8: Power-curves (dominant disease model)

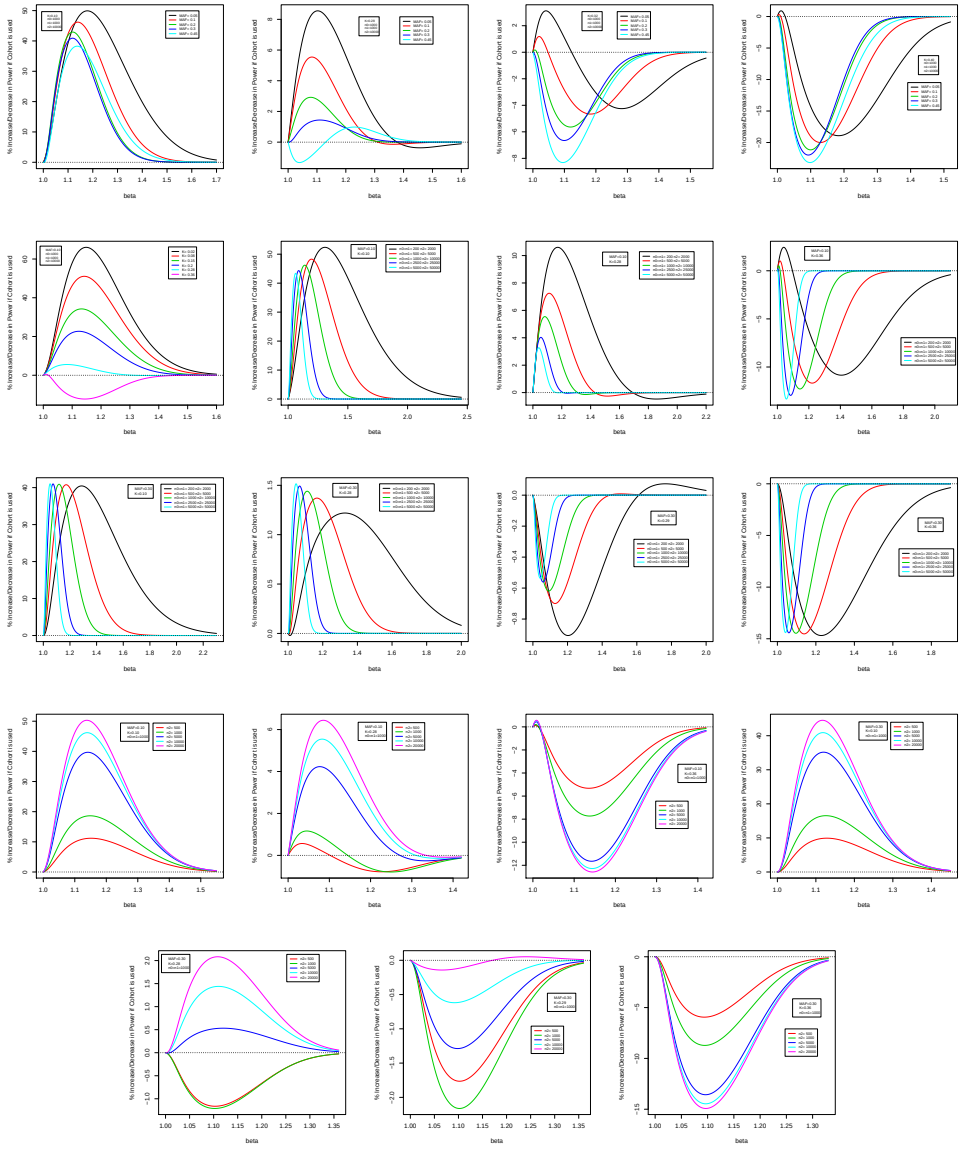


Figure B.9: $\gamma(\beta|\cdot)$-curves (dominant disease model)

Appendix C

Additional Material For Section 5.4.2

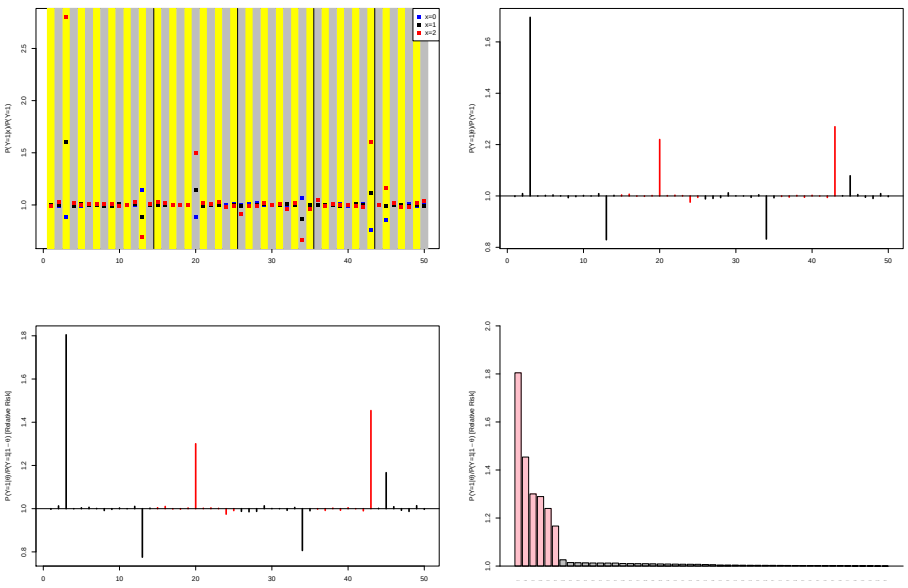


Figure C.1: Figures providing additional information for Example 1. Top: Genotypic (left) and allelic (right) risk. Bottom: Relative Allelic Risk: Genome order (left), Decreasing order (right)

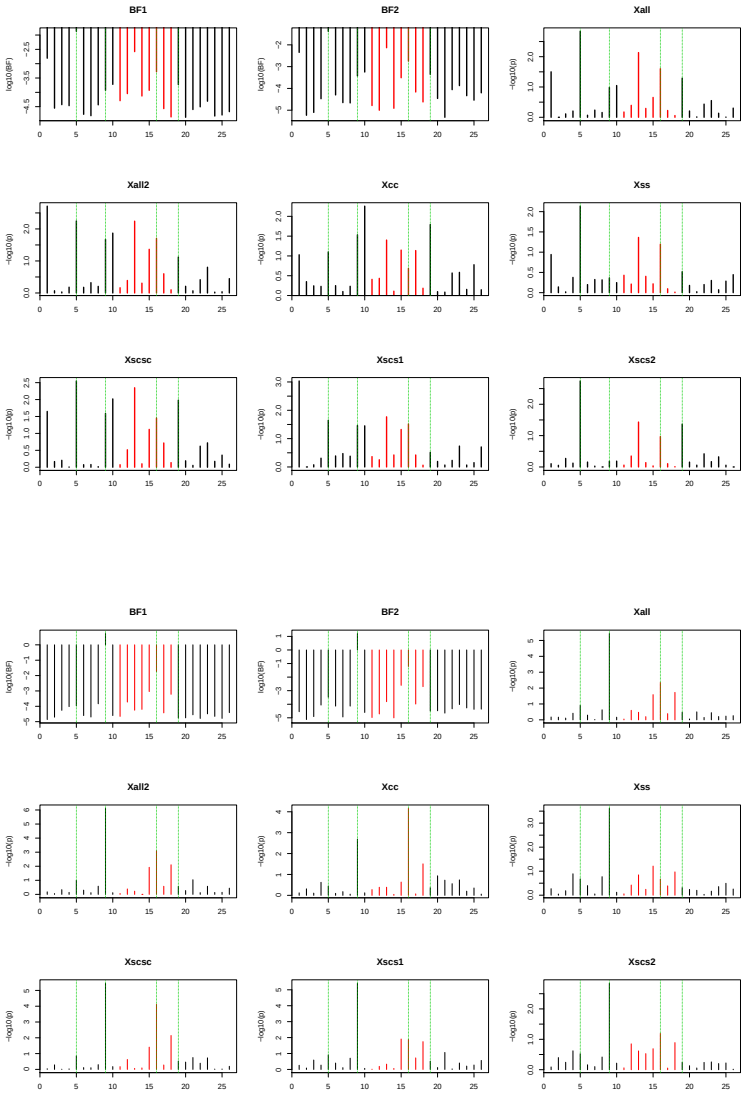


Figure C.2: The results for the two datasets that are not shown in figure 5.9

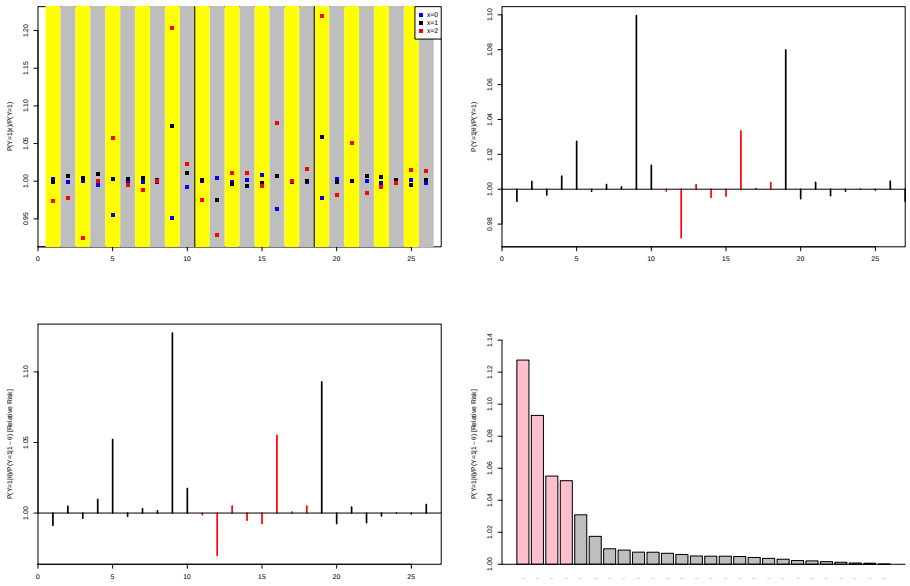


Figure C.3: Figures providing additional information for Example 2. Top: Genotypic (left) and allelic (right) risk. Bottom: Relative Allelic Risk: Genome order (left), Decreasing order (right)

Table C.1: The rankings of the 4 markers for each of the three simulated datasets (Example 2)

Actual Order	c1.9	c3.1	c2.6	c1.5
Dataset 1				
Bay. Mod. 1	8	6	4	1
Bay. Mod. 2	7	6	4	1
X^2_{all}	7	5	3	1
X^2_{all2}	6	8	5	2
CA_{cc}	3	2	10	7
CA_{ss}	10	5	3	1
CA_{scsc}	6	4	7	1
CA_{scs1}	5	10	4	3
CA_{scs2}	10	3	4	1
Dataset 2				
Bay. Mod. 1	2	24	8	6
Bay. Mod. 2	2	24	5	4
X^2_{all}	2	23	7	5
X^2_{all2}	1	22	4	7
CA_{cc}	6	11	1	2
CA_{ss}	2	26	15	11
CA_{scsc}	1	17	2	3
CA_{scs1}	2	23	12	15
CA_{scs2}	2	24	6	3
Dataset 3				
Bay. Mod. 1	1	23	2	7
Bay. Mod. 2	1	17	2	5
X^2_{all}	1	10	2	5
X^2_{all2}	1	10	2	6
CA_{cc}	2	13	1	10
CA_{ss}	1	14	8	7
CA_{scsc}	1	9	2	5
CA_{scs1}	1	11	3	6
CA_{scs2}	1	14	2	8

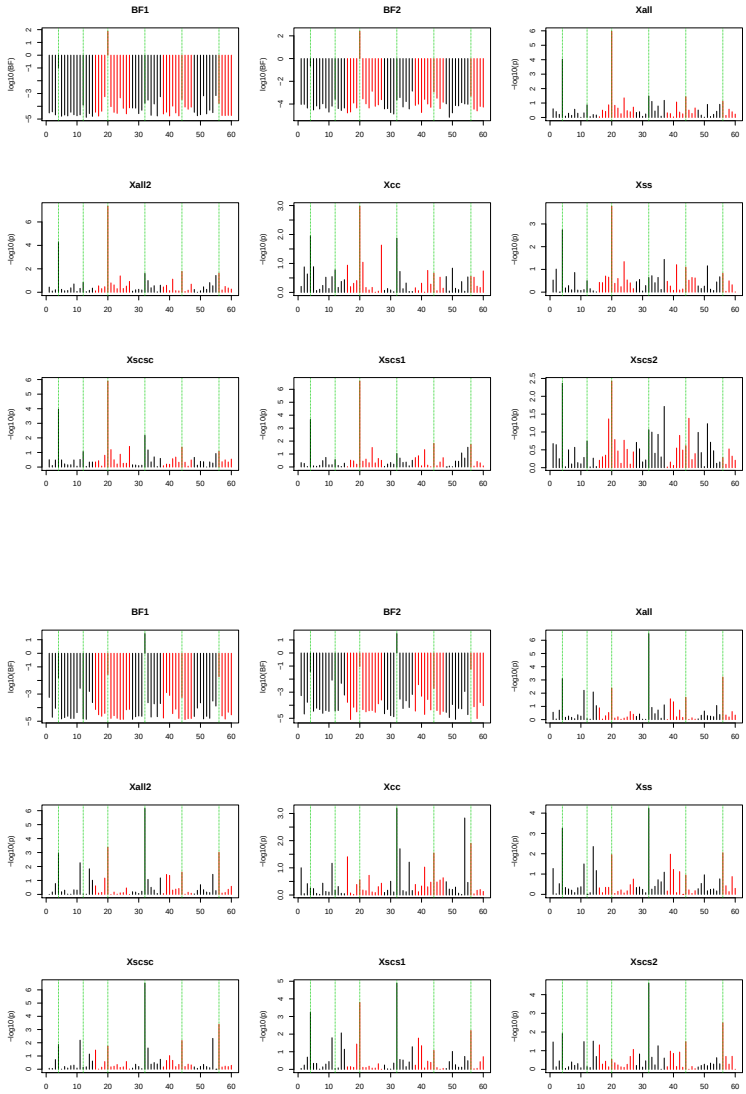


Figure C.4: The results for the two datasets that are not shown in figure 5.11

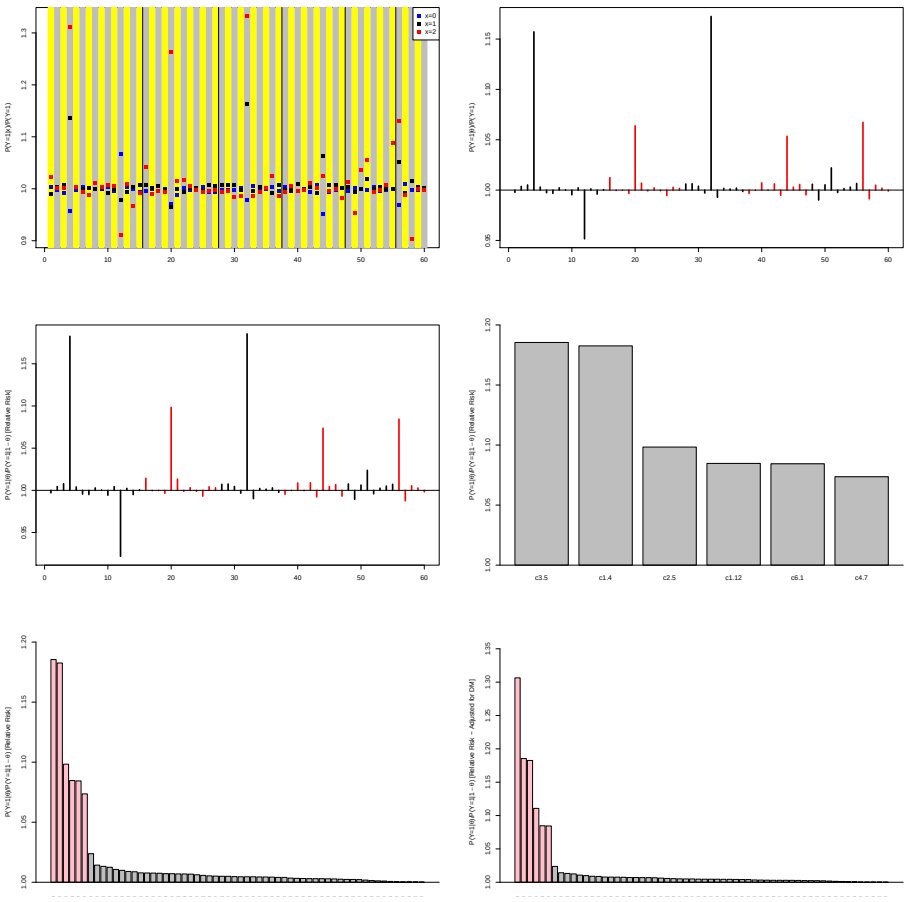


Figure C.5: Figures providing additional information for Example 3. Top: Genotypic (left) and allelic (right) risk. Middle: Relative Allelic Risk: Genome (left), Decreasing order for the 6 associated markers (right) Bottom: Relative allelic risk in dicreasing order: Non-adjusted (left) and adjusted (right) for different disease models.

Table C.2: The rankings of the 6 markers for each of the three simulated datasets (Example 3)

Actual Order	c2.5	c3.5	c1.4	c4.7	c1.12	c6.1
Sample (a)						
Bay. Mod. 1	1	10	2	8	13	11
Bay. Mod. 2	1	13	2	5	10	6
X^2_{all}	1	3	2	4	13	7
X^2_{all2}	1	5	2	3	11	4
CA_{cc}	1	3	2	14	10	16
CA_{ss}	1	17	2	7	24	10
CA_{scsc}	1	3	2	5	9	8
CA_{scs1}	1	9	2	3	18	4
CA_{scs2}	1	7	2	19	14	38
Sample (b)						
Bay. Mod. 1	1	4	2	55	15	3
Bay. Mod. 2	1	6	2	44	16	3
X^2_{all}	1	6	2	35	15	3
X^2_{all2}	1	7	2	30	21	3
CA_{cc}	3	25	10	30	7	1
CA_{ss}	2	3	1	53	28	8
CA_{scsc}	1	20	3	32	10	2
CA_{scs1}	2	4	1	34	35	3
CA_{scs2}	1	8	2	57	9	6
Sample (c)						
Bay. Mod. 1	2	1	4	10	54	3
Bay. Mod. 2	2	1	4	8	40	3
X^2_{all}	4	1	3	7	49	2
X^2_{all2}	2	1	4	7	52	3
CA_{cc}	13	1	30	5	36	3
CA_{ss}	6	1	2	14	59	4
CA_{scsc}	7	1	6	5	47	2
CA_{scs1}	2	1	3	12	54	4
CA_{scs2}	22	1	3	6	44	2

Bibliography

C. V. Ananth and D. G. Kleinbaum. Regression models for ordinal responses: a review of methods and applications. *Int J Epidemiol*, 26(6):1323–1333, Dec 1997.

Antonis C Antoniou and Douglas F Easton. Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol*, 25(3):190–202, Nov 2003. doi: 10.1002/gepi.10261. URL `http://dx.doi.org/10.1002/gepi.10261`.

P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955. ISSN 0006-341X.

David J Balding. A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 7(10):781–791, Oct 2006. doi: 10.1038/nrg1916. URL `http://dx.doi.org/10.1038/nrg1916`.

Mark A Beaumont and Bruce Rannala. The bayesian revolution in genetics. *Nat Rev Genet*, 5(4):251–261, Apr 2004. doi: 10.1038/nrg1318. URL `http://dx.doi.org/10.1038/nrg1318`.

Peter Broderick, Luis Carvajal-Carmona, Alan M Pittman, Emily Webb,

Kimberley Howarth, Andrew Rowan, Steven Lubbe, Sarah Spain, Kate Sullivan, Sarah Fielding, Emma Jaeger, Jayaram Vijayakrishnan, Zoe Kemp, Maggie Gorman, Ian Chandler, Elli Papaemmanuil, Steven Penegar, Wendy Wood, Gabrielle Sellick, Mobshra Qureshi, Ana Teixeira, Enric Domingo, Ella Barclay, Lynn Martin, Oliver Sieber, C. O. R. G. I. Consortium, David Kerr, Richard Gray, Julian Peto, Jean-Baptiste Cazier, Ian Tomlinson, and Richard S Houlston. A genome-wide association study shows that common alleles of smad7 influence colorectal cancer risk. *Nat Genet*, 39(11):1315–1317, Nov 2007. doi: 10.1038/ng.2007.18. URL `http://dx.doi.org/10.1038/ng.2007.18`.

R.M. Cantor, K. Lange, and J.S. Sinsheimer. Prioritizing gwas results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

RJ Carroll, S. Wang, and CY Wang. Prospective Analysis of Logistic Case-Control Studies. *Journal of the American Statistical Association*, 90(429), 1995.

G. Casella and R.L. Berger. Statistical inference. 2002.

W.G. Cochran. Some methods for strengthening the common χ 2 tests. *Biometrics*, 10(4):417–451, 1954. ISSN 0006-341X.

S.C. Collins. Talking glossary of genetic terms - single nucleotide polymorphisms (snps). National Human Genome Research Institute website http://www.genome.gov/glossary/index.cfm?id=185. Retrieved, 12 March 2011.

International Schizophrenia Consortium, Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, and Pamela Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, Aug 2009. doi: 10.1038/nature08185. URL `http://dx.doi.org/10.1038/nature08185`.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007. doi: 10.1038/nature05911. URL `http://dx.doi.org/10.1038/nature05911`.

Heather J Cordell and David G Clayton. Genetic association studies. *Lancet*, 366(9491):1121–1131, 2005. doi: 10.1016/S0140-6736(05)67424-7. URL `http://dx.doi.org/10.1016/S0140-6736(05)67424-7`.

P. David Wilson and C.L. Meinert. Linear regression analysis of binary response data with mixed covariates–a simulation study. *Controlled Clinical Trials*, 5(1):73–83, 1984.

N.R. Draper and H. Smith. Applied regression analysis (wiley series in probability and statistics). 1998.

Brian J Edwards, Chad Haynes, Mark A Levenstien, Stephen J Finch, and Derek Gordon. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genet*, 6:18, 2005. doi: 10.1186/1471-2156-6-18. URL `http://dx.doi.org/10.1186/1471-2156-6-18`.

J. Engel. Polytomous logistic regression. *Statistica Neerlandica*, 42(4):233–252, 1988. ISSN 1467-9574.

B. Freidlin, G. Zheng, Z. Li, and J. L. Gastwirth. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered*, 53(3):146–152, 2002.

M. H. Gail, D. Pee, and R. Carroll. Kin-cohort designs for gene characterization. *J Natl Cancer Inst Monogr*, (26):55–60, 1999.

D. Gordon, SJ Finch, M. Nothnagel, and J. Ott. Power and Sample Size Calculations for Case-Control Genetic Association Tests when Errors Are Present: Application to Single Nucleotide Polymorphisms. *Human Heredity*, 54(1):22–33, 2000. ISSN 0001-5652.

J.E. Griffin and P.J. Brown. Inference with Normal-Gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

J.W. Hardin, J.M. Hilbe, and J. Hilbe. *Generalized linear models and extensions*. Stata Corp, 2007.

G.H. Hardy. Mendelian proportions in a mixed population. *Science*, 28(706): 49, 1908.

John L Hopper, D. Timothy Bishop, and Douglas F Easton. Population-based family studies in genetic epidemiology. *Lancet*, 366(9494): 1397–1406, 2005. doi: 10.1016/S0140-6736(05)67570-8. URL `http://dx.doi.org/10.1016/S0140-6736(05)67570-8`.

F. Hu. The asymptotic properties of the maximum-relevance weighted likelihood estimators. *Canadian Journal of Statistics*, 25(1):45–59, 1997.

F. Hu, W.F. Rosenberger, and J.V. Zidek. Relevance weighted likelihood for dependent data. *Metrika*, 51(3):223–243, 2000.

Emma Jaeger, Emily Webb, Kimberley Howarth, Luis Carvajal-Carmona, Andrew Rowan, Peter Broderick, Axel Walther, Sarah Spain, Alan Pittman, Zoe Kemp, Kate Sullivan, Karl Heinimann, Steven Lubbe, Enric Domingo, Ella Barclay, Lynn Martin, Maggie Gorman, Ian Chandler, Jayaram Vijayakrishnan, Wendy Wood, Elli Papaemmanuil, Steven Penegar, Mobshra Qureshi, C. O. R. G. I. Consortium, Susan Farrington, Albert Tenesa, Jean-Baptiste Cazier, David Kerr, Richard Gray, Julian Peto, Malcolm Dunlop, Harry Campbell, Huw Thomas, Richard Houlston, and Ian Tomlinson. Common genetic variants at the crac1 (hmps) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet*, 40(1):26–28, Jan 2008. doi: 10.1038/ng.2007.41. URL `http://dx.doi.org/10.1038/ng.2007.41`.

Johanna Jakobsdottir, Michael B Gorin, Yvette P Conley, Robert E Ferrell, and Daniel E Weeks. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet*, 5(2):e1000337, Feb 2009. doi: 10.1371/journal.pgen.1000337. URL `http://dx.doi.org/10.1371/journal.pgen.1000337`.

R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. ISSN 0162-1459.

Teri A Manolio. Cohort studies and the genetics of complex disease. *Nat Genet*, 41(1):5–6, Jan 2009. doi: 10.1038/ng0109-5. URL `http://dx.doi.org/10.1038/ng0109-5`.

O. Mayo. A century of Hardy-Weinberg equilibrium. *Twin Research and Human Genetics*, 11(3):249–256, 2008. ISSN 1832-4274.

S. Nadarajah and S. Kotz. R Programs for Computing Truncated Distributions. *Journal of Statistical Software*, 16:1–8, 2006.

Melissa G Naylor, Scott T Weiss, and Christoph Lange. A bayesian approach to genetic association studies with family-based designs. *Genet Epidemiol*, 34(6):569–574, Sep 2010. doi: 10.1002/gepi.20513. URL `http://dx.doi.org/10.1002/gepi.20513`.

Miles Parkes, Jeffrey C Barrett, Natalie J Prescott, Mark Tremelling, Carl A Anderson, Sheila A Fisher, Roland G Roberts, Elaine R Nimmo, Fraser R Cummings, Dianne Soars, Hazel Drummond, Charlie W Lees, Saud A Khawaja, Richard Bagnall, Denis A Burke, Catherine E Todhunter, Tariq Ahmad, Clive M Onnie, Wendy McArdle, David Strachan, Graeme Bethel, Claire Bryan, Cathryn M Lewis, Panos Deloukas, Alastair Forbes, Jeremy Sanderson, Derek P Jewell, Jack Satsangi, John C Mansfield, Wellcome Trust Case Control Consortium, Lon Cardon, and Christopher G Mathew. Sequence variants in the autophagy gene irgm and multiple other replicating loci contribute to crohn's disease susceptibility. *Nat Genet*, 39(7):830–832, Jul 2007. doi: 10.1038/ng2061. URL `http://dx.doi.org/10.1038/ng2061`.

Alan M Pittman, Emily Webb, Luis Carvajal-Carmona, Kimberley Howarth, Maria Chiara Di Bernardo, Peter Broderick, Sarah Spain, Axel Walther, Amy Price, Kate Sullivan, Philip Twiss, Sarah Fielding, Andrew Rowan, Emma Jaeger, Jayaram Vijayakrishnan, Ian Chandler, Steven Penegar, Mobshra Qureshi, Steven Lubbe, Enric Domingo, Zoe Kemp, Ella Barclay, Wendy Wood, Lynn Martin, Maggie Gorman, Huw Thomas, Julian Peto, Timothy Bishop, Richard Gray, Eamonn R Maher, Anneke Lucassen, David Kerr, Gareth R Evans, C. O. R. G. I. Consortium, Tom van Wezel, Hans Morreau, Juul T Wijnen, John L Hopper, Melissa C Southey, Graham G Giles, Gianluca Severi, Sergi Castellv-Bel, Clara Ruiz-Ponte, Angel Carracedo, Antoni Castells, E. P. I. C. O. L. O. N. Consortium, Asta Frsti, Kari Hemminki, Pavel Vodicka, Alessio Naccarati, Lara Lipton, Judy W C Ho, K. K. Cheng, Pak C Sham, J. Luk, Jose A G Agndez, Jose M Ladero, Miguel de la Hoya, Trinidad Calds, Iina Niittymki, Sari Tuupanen, Auli Karhu, Lauri A Aaltonen, Jean-Baptiste Cazier, Ian P M Tomlinson, and Richard S Houlston. Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer. *Hum Mol Genet*, 17(23):3720–3727, Dec 2008. doi: 10.1093/hmg/ddn267. URL `http://dx.doi.org/10.1093/hmg/ddn267`.

R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403, 1979. ISSN 0006-3444.

C.P. Robert and J.M. Marin. On computational tools for Bayesian data analysis. *ArXiv e-prints*, stat.CO(arXiv:1002.2684v2), 2010.

Nilesh J Samani, Jeanette Erdmann, Alistair S Hall, Christian Hengstenberg,

Massimo Mangino, Bjoern Mayer, Richard J Dixon, Thomas Meitinger, Peter Braund, H-Erich Wichmann, Jennifer H Barrett, Inke R Knig, Suzanne E Stevens, Silke Szymczak, David-Alexandre Tregouet, Mark M Iles, Friedrich Pahlke, Helen Pollard, Wolfgang Lieb, Francois Cambien, Marcus Fischer, Willem Ouwehand, Stefan Blankenberg, Anthony J Balmforth, Andrea Baessler, Stephen G Ball, Tim M Strom, Ingrid Braenne, Christian Gieger, Panos Deloukas, Martin D Tobin, Andreas Ziegler, John R Thompson, Heribert Schunkert, W. T. C. C. C., and the Cardiogenics Consortium. Genomewide association analysis of coronary artery disease. *N Engl J Med*, 357(5):443–453, Aug 2007.

P. D. Sasieni. From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261, Dec 1997.

S.R. Seaman and S. Richardson. Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika*, 91(1): 15, 2004. ISSN 0006-3444.

N. Sepulveda, CD Paulino, J. Carneiro, and C. Penha-Goncalves. Allelic penetrance approach as a tool to model two-locus interaction in complex binary traits. *Heredity*, 99(2):173–184, 2007.

J. S. Shoemaker, I. S. Painter, and B. S. Weir. Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet*, 15(9):354–358, Sep 1999.

D. Siegmund. Upward bias in estimation of genetic effects. *Am J Hum Genet*, 71(5):1183–1188, Nov 2002. doi: 10.1086/343819. URL `http://dx.doi.org/10.1086/343819`.

S. L. Slager and D. J. Schaid. Case-control studies of genetic markers: power and sample size approximations for armitage's test for trend. *Hum Hered*, 52(3):149–153, 2001.

R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, 52(3):506–516, Mar 1993.

Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nat Rev Genet*, 10(10):681–690, Oct 2009. doi: 10.1038/nrg2615. URL `http://dx.doi.org/10.1038/nrg2615`.

Daniel O Stram. Tag snp selection for association studies. *Genet Epidemiol*, 27(4):365–374, Dec 2004. doi: 10.1002/gepi.20028. URL `http://dx.doi.org/10.1002/gepi.20028`.

D.C. Thomas. *Statistical methods in genetic epidemiology*. Oxford University Press, USA, 2004. ISBN 019515939X.

Duncan C Thomas, Graham Casey, David V Conti, Robert W Haile, Juan Pablo Lewinger, and Daniel O Stram. Methodological issues in multistage genome-wide association studies. *Stat Sci*, 24(4):414–429, Nov 2009.

Timothy Thornton and Mary Sara McPeek. Roadtrips: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet*, 86(2): 172–184, Feb 2010. doi: 10.1016/j.ajhg.2010.01.001. URL `http://dx.doi.org/10.1016/j.ajhg.2010.01.001`.

John A Todd, Neil M Walker, Jason D Cooper, Deborah J Smyth, Kate

Downes, Vincent Plagnol, Rebecca Bailey, Sergey Nejentsev, Sarah F Field, Felicity Payne, Christopher E Lowe, Jeffrey S Szeszko, Jason P Hafler, Lauren Zeitels, Jennie H M Yang, Adrian Vella, Sarah Nutland, Helen E Stevens, Helen Schuilenburg, Gillian Coleman, Meeta Maisuria, William Meadows, Luc J Smink, Barry Healy, Oliver S Burren, Alex A C Lam, Nigel R Ovington, James Allen, Ellen Adlem, Hin-Tak Leung, Chris Wallace, Joanna M M Howson, Cristian Guja, Constantin Ionescu-Trgovite, Genetics of Type 1 Diabetes in Finland, Matthew J Simmonds, Joanne M Heward, Stephen C L Gough, Wellcome Trust Case Control Consortium, David B Dunger, Linda S Wicker, and David G Clayton. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet*, 39(7):857–864, Jul 2007. doi: 10.1038/ng2068. URL `http://dx.doi.org/10.1038/ng2068`.

Ian Tomlinson, Emily Webb, Luis Carvajal-Carmona, Peter Broderick, Zoe Kemp, Sarah Spain, Steven Penegar, Ian Chandler, Maggie Gorman, Wendy Wood, Ella Barclay, Steven Lubbe, Lynn Martin, Gabrielle Sellick, Emma Jaeger, Richard Hubner, Ruth Wild, Andrew Rowan, Sarah Fielding, Kimberley Howarth, C. O. R. G. I. Consortium, Andrew Silver, Wendy Atkin, Kenneth Muir, Richard Logan, David Kerr, Elaine Johnstone, Oliver Sieber, Richard Gray, Huw Thomas, Julian Peto, Jean-Baptiste Cazier, and Richard Houlston. A genome-wide association scan of tag snps identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet*, 39(8):984–988, Aug 2007. doi: 10.1038/ng2085. URL `http://dx.doi.org/10.1038/ng2085`.

Ian P M Tomlinson, Emily Webb, Luis Carvajal-Carmona, Peter Broderick, Kimberley Howarth, Alan M Pittman, Sarah Spain, Steven Lubbe, Axel Walther, Kate Sullivan, Emma Jaeger, Sarah Fielding, Andrew Rowan, Jayaram Vijayakrishnan, Enric Domingo, Ian Chandler, Zoe Kemp, Mobshra Qureshi, Susan M Farrington, Albert Tenesa, James G D Prendergast, Rebecca A Barnetson, Steven Penegar, Ella Barclay, Wendy Wood, Lynn Martin, Maggie Gorman, Huw Thomas, Julian Peto, D. Timothy Bishop, Richard Gray, Eamonn R Maher, Anneke Lucassen, David Kerr, D. Gareth R Evans, C. O. R. G. I. Consortium, Clemens Schafmayer, Stephan Buch, Henry Vlzke, Jochen Hampe, Stefan Schreiber, Ulrich John, Thibaud Koessler, Paul Pharoah, Tom van Wezel, Hans Morreau, Juul T Wijnen, John L Hopper, Melissa C Southey, Graham G Giles, Gianluca Severi, Sergi Castellv-Bel, Clara Ruiz-Ponte, Angel Carracedo, Antoni Castells, E. P. I. C. O. L. O. N. Consortium, Asta Frsti, Kari Hemminki, Pavel Vodicka, Alessio Naccarati, Lara Lipton, Judy W C Ho, K. K. Cheng, Pak C Sham, J. Luk, Jose A G Agndez, Jose M Ladero, Miguel de la Hoya, Trinidad Calds, Iina Niittymki, Sari Tuupanen, Auli Karhu, Lauri Aaltonen, Jean-Baptiste Cazier, Harry Campbell, Malcolm G Dunlop, and Richard S Houlston. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet*, 40(5):623–630, May 2008. doi: 10.1038/ng.111. URL `http://dx.doi.org/10.1038/ng.111`.

J. Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology*, 33(1):79–86, 2009.

Eleftheria Zeggini, Michael N Weedon, Cecilia M Lindgren, Timothy M Frayling, Katherine S Elliott, Hana Lango, Nicholas J Timpson, John R B Perry, Nigel W Rayner, Rachel M Freathy, Jeffrey C Barrett, Beverley Shields, Andrew P Morris, Sian Ellard, Christopher J Groves, Lorna W Harries, Jonathan L Marchini, Katharine R Owen, Beatrice Knight, Lon R Cardon, Mark Walker, Graham A Hitman, Andrew D Morris, Alex S F Doney, Wellcome Trust Case Control Consortium (WTCCC), Mark I McCarthy, and Andrew T Hattersley. Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341, Jun 2007. doi: 10.1126/science.1142364. URL `http://dx.doi.org/10.1126/science.1142364`.

G. Zheng, B. Freidlin, Z. Li, and J.L. Gastwirth. Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical Journal*, 45(3):335–348, 2003. ISSN 1521-4036.

Gang Zheng and Xin Tian. The impact of diagnostic error on testing genetic association in case-control studies. *Stat Med*, 24(6):869–882, Mar 2005. doi: 10.1002/sim.1976. URL `http://dx.doi.org/10.1002/sim.1976`.