Comparing and developing statistical methods for fine-mapping genes in complex diseases

Amy Spencer

Submitted for the degree of Doctor of Philosophy

School of Mathematics and Statistics

February 2014

Supervisors: Dr Kevin Walters and Prof Angela Cox

University of Sheffield

Acknowledgements

I would like to express my gratitude to my supervisors for their teaching, guidance and patience. Angie, thank you for your encouragement and always being able to point me in the right direction. Kevin, you deserve more thanks than anyone, because you helped me with everything, so thanks for it all. Thanks, also, to the Engineering and Physical Sciences Research Council, which funded this project, and to the two examiners who carefully read the original draft and helped me to improve it.

There are more people at the University that helped me, though, particularly in the School of Mathematics and Statistics. I have found many friends among the staff, and particularly the PhD students, in the department. Thank you all for keeping me entertained so many lunchtimes in the common room and Friday evenings in the pub! Not wishing to miss anyone out by trying to list them all, the only person I will name is Elizabeth Boggis, who has put up with more than most! I was very fortunate to start at the same time as another student under the same supervisor and working on a related project. We were often able to give each other help and advice and I am sure I progressed much more quickly because of sharing an office with her.

I would also like to thank Douglas Easton, Kyriaki Michailidou and the other Collaborative Oncological Gene-Environment Study researchers, for allowing me to use their data in my work. Wei-Yu Lin also provided a lot of help, mainly with respect to Linux computing, downloading functional data and the iCOGS analysis.

Thanks must also go to my family. In particular, to my parents, David and Diane Baddeley, who have always encouraged me to do whatever I wanted with my life and always strived to find and provide activities that would keep me both interested and entertained. Also to my sisters, Lizzy and Louise, who, along with my parents (and although I am sure they cannot really be interested!), have continued to ask about my progress and listen to my rants when things when wrong. My in-laws and wider family have also been wonderfully supportive, and I know I am lucky to have such great relatives.

So finally I come to my amazing husband, Martin. I can't really express my gratitude to him in words, but I'm sure he knows how grateful I am, both for just being there always and for carrying me through the worst parts of this project. The last few years have been a whirlwind for us, because whilst I've been studying, we got engaged and then married. As well as the highs, there have also been lows, but we got here together and we're both ready for our next challenges! Thank you, my love.

Abstract

This project is an investigation of current and novel methods of analysing genotype data from fine-mapping association studies, where a single, unknown causal variant is present in a genomic region known to be associated with disease risk. The analysis methods are all univariate and can be used to filter the genetic variants in an association region, reducing them to a set of candidate causal variants.

Initially, a number of methods including p-value, likelihood, genetic map distance and linkage disequilibrium based analyses are compared using simulated data. A method that we call likelihood percentile is found to generally be the most effective, in various scenarios that may occur in fine-mapping studies. I also investigate the effects of varying such things as the causal variant, genomic region and genetic effect on the efficacy of likelihood percentile filtering. To explain the difference in filtering results using two different likelihood based methods, I consider the likelihood surface as a function of the numbers of case and control subjects with the risk genotype. In addition, I examine the effect of imputing missing genotype dosage, as is common practice in fine-mapping studies, on the efficacy of variant filtering.

The use of Bayes factors to filter genetic variants is investigated, assuming that a prior on the natural logarithm of the odds ratio of the form $N \sim (0, W)$ is considered appropriate. It is shown that filtering efficacy is sensitive to W, so several methods of informing the choice of W are compared. These include defining W as a function of minor allele frequency, an empirical Bayes method and several novel forms for the Bayes factor which put a prior on W, thus taking into account prior uncertainty about W.

Several appropriate methods are applied to a large dataset of a fine-mapped region from the Collaborative Oncological Gene-environment Study.

Publications

A. V. Spencer, A. Cox and K. Walters. Comparing the efficacy of SNP filtering methods for identifying a single causal SNP in a known association region. *Annals of Human Genetics*, 78(1):50-61, 2014.
A. V. Baddeley, K. Walters, A. Cox and W. Y. Lin. Using Bayes factors to analyse fine-mapped genotype data. *Genetic Epidemiology*, 36:763-764, 2012. (Poster abstract for *Annual meeting of the International Genetic Epidemiology Society (IGES)*, Oct 18-20 2012, Stevenson, WA, USA; as Amy Victoria Baddeley.)

Thesis summary

This thesis is a comparison of statistical methods for analysing genotype data on the fine-mapping level. Only methods which analyse variants individually are considered, so that only marginal effects and no interactions are taken into account. No attempt is made to classify variants as statistically ‘significant’ or otherwise. The methods are compared on their ability to rank the causal variant highly among all the variants in a region of interest.

Chapter 1 is the introduction and begins with the basic genetic background to which the statistical methods are applied. All the relevant genetic terminology is explained in this chapter. It also contains descriptions of genome-wide association studies and fine-mapping studies and of the current standard methods of analysis used in these studies. Genetic effects based on different modes of inheritance are described and it is explained how these may be modelled using logistic regression. Throughout the thesis, software is mentioned that has been used to aid analysis. This software is all commonly used in genetic analysis and its use is described in this chapter. There is also a brief introduction to the original research, with what is referred to as the filtering framework described, and some of the methods used to compare the efficacy of different filters, such as receiver operating characteristic curves, are explained. The genotype data from the Collaborative Oncological Gene-environment Study that is later used to illustrate the application of the methods is described in this chapter.

Chapter 2 is a thorough comparison of methods that fit into the filtering framework and use only the genotype data from a fine-mapping study to calculate the filtering statistics. The different filtering statistics used in this chapter are described, with some based on p-values, some on likelihoods and some on the structural relationship between variants. Simulated data is used to compare the overall efficacy of these methods. Certain methods have highly variable outcomes in terms of the number of false positive signals they generate and this is investigated in detail. The most efficacious method appears to be that labelled likelihood percentile, so a sensitivity analysis is carried out using this method. Its sensitivity to the causal variant, the size and type of effect of that variant and the sample size of the study are all investigated. The likelihood surface dependant on the number of case and control subjects with the causal variant is used to explain the difference in results of two likelihood-based methods, likelihood percentile and relative likelihood. It is now common practice to impute missing genotypes in genome-wide and fine-mapping association studies, so included in this chapter is a comparison of the same analysis on simulated genotype data and the same data but with the majority of the variants having only imputed genotype doses. It should be noted that much of the work contained in this chapter has been published in *Annals of Human Genetics* in an

article by Spencer, Cox and Walters, the author and supervisors of this thesis [46].

Chapter 3 is an investigation into using Bayes factors in filtering. These must be approximated and, initially, an approximation method developed by J. Wakefield is used. This requires a specific form of prior distribution to be put on the effect size (the log-odds ratio) of the variants. Using Bayes factors as the filtering statistic, the sensitivity of the results to the variance of this prior is investigated. Several methods which may be used to choose the variance of this prior are described, including choosing variant-specific variances based on MAF and choosing a universal variance based on empirical information from the genotype data. Several novel forms of Bayes factor are also described in this chapter which allow for added uncertainty in the variance by putting a hyperparametric distribution on it. Finally, a method is described which allows for the inclusion of prior functional information on the variant level. This is done by specifying a prior probability of association for each variant, based on the available information, and combining this with the Bayes factor for that variant to calculate a posterior probability of association, which is then used as the filtering statistic.

Chapter 4 is an illustration of the application of some of the methods included in the thesis to genotype data from a real fine-mapping study. The most appropriate methods investigated in the previous chapters are applied to the data from the Collaborative Oncological Gene-environment Study, and their results are compared and discussed. Because this study has a very large sample size, some of the analyses were repeated on a smaller subset of the study population. Some advantages and limitations to using the different filtering methods with smaller fine-mapping studies are demonstrated with this subset.

Chapter 5 contains a summary of the project outcomes and some conclusions from the work. Advice on choosing an appropriate sample size and filtering method, dependent on the available resources, is given, as well as limitations which may restrict the choice in some cases. The methods contained in this thesis are discussed in relation to other methods which have recently been published. Some possible avenues for further work are also considered.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Genetic background to the project

Genetic epidemiology is the study of the relationship between genetic factors and health outcomes. Using a combination of biological, chemical and statistical techniques, it is possible to uncover relationships between variations in DNA and variations in physical traits. These include our basic appearance, for example eye, skin and hair colour and, perhaps more importantly, susceptibility to many diseases.

Research began in the early 1900s and RA Fisher was a famous pioneer in this area. During the 1950s and 1960s, significant advances were made on linkage (genetic patterns in families) with NE Morton publishing one particular landmark paper [38]. With the completion of the Human Genome Project in 2003 [4] much of the current research into genetic association in populations was made possible. Several major discoveries were found through family-based (linkage) studies and population-based studies such as genome-wide association studies (GWAS). In particular, linkage was a powerful tool in the discovery of high penetrance variants (those with a large effect on disease risk), whereas GWAS enabled variants with lower penetrance to be identified. Many of the variants discovered using these methods result in the risk of developing a disease more than doubling when one allele (version of the variant) is present compared to when the other allele is present. However, it is likely that by now most genetic variants with such large effects have already been discovered. Such studies have also made huge progress in finding many genes and regions of chromosomes with very low penetrance with relation to a particular disease. However, within such a genetic region thousands of variants may be found, and identifying those which are causal is still a difficulty. Pinpointing the precise variants causing such associations could open up possibilities of medicine targeted at causal genetic loci or the downstream products (such as proteins) affected by those variants.

Fine-mapping is a branch of genetic epidemiology which aims to find the precise causal variants, usually in a region where there is already known to be an association with a disease.

1.1.1 DNA inheritance

Population based association studies rely on the knowledge that all humans, as a single species, share almost all of their DNA. However, the minute variations that do occur result in the differences in our natural physical appearance as well as how prone we are to different diseases, and these physical outcomes are known as our phenotypes.

DNA is present in human cell nuclei in the form of molecules called chromosomes. Almost everybody has 46 chromosomes in each of these cells; 22 pairs of autosomes (chromosomes 1 to 22) and a pair of sex chromosomes (two X chromosomes for females and an X and a Y chromosome for males). Humans also have small DNA molecules which are found in another part of the cell, the mitochondria. This complete set of DNA is known as the genome, but from now on we will consider only chromosomal DNA. The chromosomes are a double stranded string of nucleotide bases, carried on a sugar-phosphate backbone, which pair in a very specific way. There are 4 possible bases: adenine (**A**), cytosine (**C**), guanine (**G**) and thymine (**T**). Because of the types of chemical bonds they are able to form, an **A** base on one strand of DNA pairs with a **T** base on the other strand, and similarly, **C** pairs with **G**. Therefore, when carrying out any sort of analysis, it is only necessary to take into account the sequence of bases on one of the strands as the opposite sequence can be inferred from it. The chromosomes vary in length from about 47 million (Chr 21) to about 247 million (Chr 1) pairs of nucleotide bases. Identical copies of a person's DNA are present in the nucleus of almost all the cells in their body.

To understand the variation in DNA, it helps to understand the process of inheritance. Figure 1.1 illustrates, in simple terms, meiosis, which is the creation of the sex cells (ovum and sperm) in the female and male bodies. The creation of most other cells involves an identical copy of the DNA being made to be incorporated in the new cell, but the same is not true with meiosis. The full DNA is present in the nucleus of the original cell (1). In this figure, we consider only 2 pairs of chromosomes, but in reality there are 23 pairs. One copy of each pair will have been maternally inherited and these are represented in red, whereas those represented in blue were paternally inherited. These chromosomes are replicated, so that they are made up of two chromatids, joined together at the centromere and giving an X-shaped appearance. As meiosis begins, the homologous chromosomes (the matching maternally and paternally inherited ones) pair up and begin to exchange material during what is known as crossover or recombination (2). The cell then divides in two, with one copy of each partially divided chromosome going into each of the new cells (3). Finally, the pairs of chromatids split fully from one another and the new cells divide again, with one copy of each chromatid going to each new cell (4). As can be seen the in the diagram, the new sex cells (gametes) have only one copy of each chromosome (a total of 23), rather than pairs. They are thus known as haploid, rather than diploid like most human cells. Because of the recombination of DNA between the original copies of the chromosomes, each new sex cell has a different mix of DNA.

A new individual's DNA is created when they are conceived. Figure 1.2

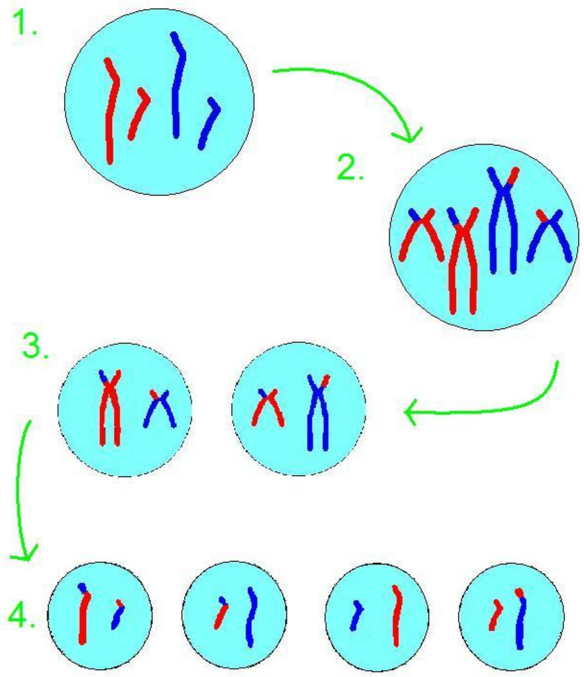


Figure 1.1: Meiosis. (1) Original cell; (2) splitting into chromotids (replication) and recombination; (3) cell division; (4) gametes (sex cells).

demonstrates how this occurs, using the person whose cells were shown in the previous example as one of the parents. The other parent's gametes will also have been produced by meiosis and parts of their maternal (purple) and paternal (orange) chromosomes will have also recombined to create new chromosomes. When the gametes, one ovum and one sperm, meet at fertilisation (5), the DNA from the two parent's cells combine to produce the offspring's diploid DNA (6). This means that the newly conceived child will have it's own different mix of DNA, including parts from all four grandparents.

1.1.2 DNA code

The sequence of nucleotide bases (**A**s, **T**s, **C**s and **G**s) can be thought of as a code that other structures in the cells are able to read and interpret. It is a triplet code, with 3 consecutive bases coding for a unique amino acid, and dependant on the overall sequence, different amino acids combine to make different proteins. In turn, these proteins build the structures in the body and carry out the chains of chemical reactions which are needed for life. A section of DNA that codes for a single protein is called a gene and each chromosome contains hundreds or thousands of genes.

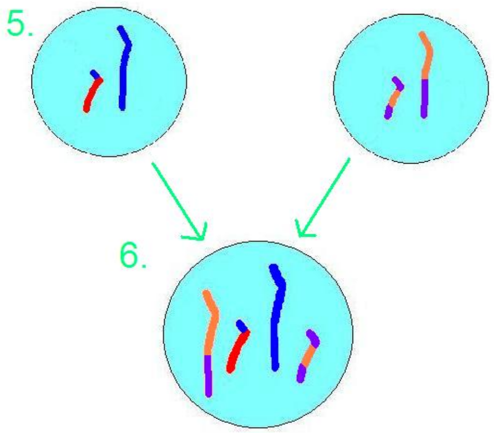


Figure 1.2: Conception. (5) Parents' haploid gametes; (6) offspring's diploid cell.

It is the few variations in DNA between human subjects that are of interest, and these variations can be passed on from parents to offspring through meiosis and conception. These come in several forms, for example, a minority of the population could have extra bases (insertions), missing bases (deletions) or some sort of rearrangement of bases at a particular position (locus) in the DNA. However, the majority of this work concerns only the most common type of DNA variant, single nucleotide polymorphisms (SNPs, termed *snips*).

A SNP occurs when a single nucleotide base at a specific locus of DNA varies between different people. For example, a part of a particular chromosome may have the bases TGTA**G**CTGGC in 80% of the population of that chromosome, but the bases TGTA**C**CTGGC in 20% of the population. The fifth locus in this sequence is a SNP with the two alleles **G** and **C**, where **C** is the minor allele with minor allele frequency (MAF) = 0.2. Generally, SNPs only have two alleles. As any individual has two copies of each chromosome, they will have two copies of the SNP. These make up their genotype, which can be homozygous (**GG** or **CC**) or heterozygous (**GC**) depending on which alleles they have inherited from their parents. A person's genotypes can be determined using biological and chemical processes known as genotyping. Ideally, we would like to know the two different haplotypes. In this context, haplotype refers to the specific ordered sequence of alleles on one of the two copies of a chromosome. However, genotyping is not usually able to provide this information, so that although we are able to discover that a person has a heterozygous genotype at one SNP, we won't know which copy of the chromosome each allele is located on.

Although a SNP is a change in the DNA code, a single one may make no difference to the resulting phenotype for several reasons. For example, different sections of DNA are used differently by the body. As previously mentioned, a section which codes for a protein is known as a gene, but only around 1% of DNA is genes, with the parts in between known as intergenic regions. Each gene is also split into several intronic and exonic regions, but only the exons code directly for the protein. Although we would expect a SNP to have a more noticeable effect if it is located in an exon, rather than an intron or an intergenic region, only a small proportion of SNPs are exonic. Also, the codes for amino acids are made up of sequences of 3 consecutive bases and, because there are 4 possible bases, there are $4^3 = 64$ possible sequences of 3 bases. However, there are only 20 different amino acids, so many of these have multiple 3-base codes relating to them (3 of these codes are also stop codons, which tell the cell to stop creating amino acids from the DNA code). This non-uniqueness is known as the code being degenerate. For example, all of the sequences **ATT**, **ATC** and **ATA** code for the amino acid *isoleucine*, so if there was a SNP at the last base in a sequence which normally coded for *isoleucine*, then a change in amino acid would only occur if the minor allele was a **G**. In that case, the SNP would be non-synonymous and *methionine* would be produced instead. Otherwise, it would be known as a synonymous SNP and *isoleucine* would still be produced. If a SNP does result in a different amino acid being produced, this may result in a change of phenotype in people with the different SNP genotypes. If the SNP is in an important gene, then it could result in a significant qualitative effect, but more often the effect is smaller and even SNPs in intergenic regions can affect the level at which a neighbouring gene is expressed, resulting in quantitative changes.

1.1.3 Genetic recombination and DNA structure

Recombination fractions and genetic distance

In §1.1.1, it was explained that one of the reasons for each person having unique DNA is the recombinations (or crossovers) which occur during meiosis in the formation of ovum and sperm cells. This was represented by (2) in Figure 1.1. There are usually only a small number of recombinations on each chromosome.

Two loci on the same chromosome can be thought to be related in terms of the number of recombinations that occur in the interval between them. Two loci that are very close together are very unlikely to have a recombination occur between them, but the further apart they lie, the larger the probability of a recombination becomes. Where the loci in question have different possible alleles (for example, if they are SNPs), families can be genotyped to determine whether or not a new combination of alleles occurred after meiosis. For example, take a person with two copies of Chromosome 1 who has genotype AB at loci A and B on one copy of the chromosome and ab at the same loci on the other copy. Any gametes they produce which have haplotypes AB or ab are known as non-recombinant, whereas any which have Ab or aB genotypes are recombinant. It should be emphasised, though, that non-recombinant gametes have not necessarily had no recombinations between these two loci. In fact they can have had any possible even number of recombinations, just as recombinant gametes can have had any possible odd number of recombinations. Because of this it is difficult to measure the probability of a single recombination between two loci, so what is usually recorded is the recombination fraction, the proportion of gametes which are recombinant for two loci.

Recombination fractions can be used to determine the genetic distance between two loci. Genetic map distance is measured in morgans (M) and centimorgans (cM, of which there are 100 in a morgan). Using Haldane's mapping function, 1cM is defined as the genetic distance between two loci with a recombination fraction of 0.01 [24]. This corresponds to approximately 1 million bases (one mega-base, Mb) in most of the human genome, but does vary to some extent. The variation is due to different recombination rates, for example, between the chromosomes and the sexes.

Although genetic map distances can be thought of in this way, they are not additive across large distances [49]. Because even numbers of crossovers between two loci result in a non-recombinant, the recombination fraction never exceeds 0.5. This is also the recombination fraction between two loci on different chromosomes, as it is pure chance whether or not they are passed on together. However, on an additive scale, this would equate to only 50cM, and many chromosomes are longer than this. Recombinations are also not independent. A phenomenon known as interference results in one recombination inhibiting another close by.

LD structure

As was explained above, two loci close together on a chromosome are unlikely to have a recombination between them. Therefore, in most cases, the SNP alleles that are close together are passed on together from one generation to the next. This means that certain SNP alleles are almost always found together (are highly correlated). This correlation is caused by genetic linkage and is termed linkage disequilibrium (LD). There are two common measures of LD which can be calculated from population data to get a good idea of which SNPs are commonly found together. These measures of LD are D' and r^2.

When the allele frequencies at two loci are completely independent, those loci are said to be in linkage equilibrium, and D' is a scaled value of deviation from this equilibrium. For two alleles at separate loci, A and B, p_A, p_B and p_{AB} are the frequencies of these alleles and the frequency at which they both occur on the same chromosome, respectively. $p_A p_B = p_{AB}$ would hold if A and B were in linkage equilibrium and deviance from this is measured as $D = p_{AB} - p_A p_B$. However, this value depends on allele frequency, so is scaled by dividing by $D_{max} = min(p_A p_b, p_a p_B)$ if $D \geq 0$ or $D_{max} = max(-p_A p_B, -p_a p_b)$ if $D < 0$ (where a and b are the alternative alleles at the A and B loci). This gives

$$D' = D/D_{max},$$

a measure of LD.

Alternatively, the squared correlation coefficient, r^2, can be used. Using the same notation as above, this is calculated

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}.$$

Although D' and r^2 are related, one cannot necessarily be calculated from the other. Therefore, in different circumstances, one may describe LD in a more appropriate way than the other, and when considering the LD structure of a region, it is often worth looking at both.

Haploview [10] is a piece of software which allows the input of genotype data and outputs the structure of the region in terms of LD. This helps to identify LD blocks of highly associated SNPs which are likely to be passed on together. The software outputs r^2 values as well as D' values for each pair of SNPs in the input dataset and can produce heat-map style diagrams showing which parts of the region of interest have more correlation.

Figure 1.3 is an LD plot produced in Haploview which shows the LD structure of some SNPs in chromosome 2. This plot has been generated using publicly available SNP data. The plot in Figure 1.3 represents single SNPs in order of location on the chromosome across the top, with diagonal lines coming down from them to the left and right. Where the lines from two SNPs intercept, the plot represents the D' value of that pair of SNPs as a shade of red, with darker shades representing higher values (NB: in this kind of plot, the blue also represents high D' but low LOD score, the $\log_{10}$ of the likelihood odds ratio [38], which Haploview uses as 'a measure of confidence in the value of D'').

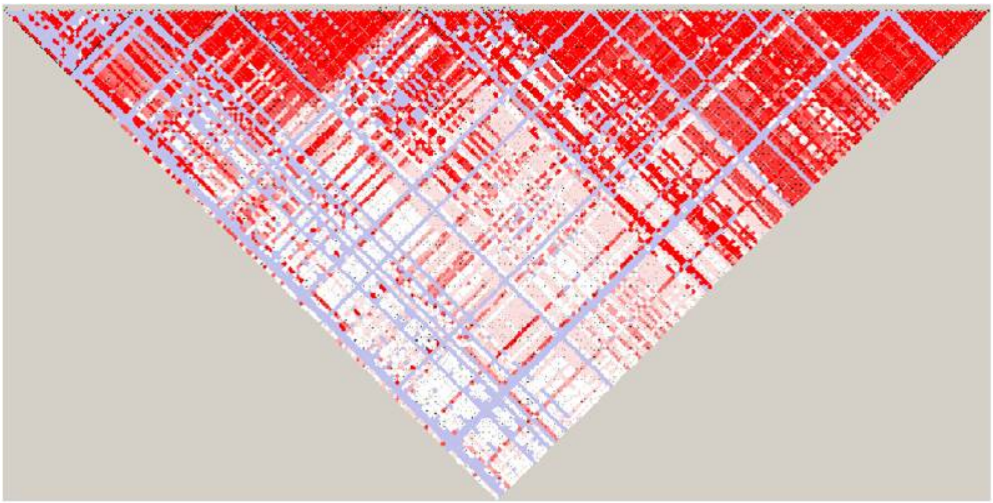


Figure 1.3: Linkage disequilibrium plot of part of chromosome 2. The size of the D' value between two SNPs is indicated by the depth of colour at the intercept of the diagonal lines between them.

1.2 Current statistical analysis in genetic association studies

Now that it is possible to genotype DNA (to read the bases at targeted chromosomal loci), studies can be carried out analysing the effect of known DNA variants on susceptibility to diseases. Determining the genotypes uses biological and chemical technologies, but the analysis must be carried out using statistical techniques. Historically, this was done largely through family based linkage studies, taking subjects from families in which the disease of interest occurs in multiple members. However, these kinds of studies do not measure LD or association with a disease in a whole population and are not so useful at identifying variants that only slightly affect disease risk, as in those cases it is less likely that the disease will affect multiple close family members.

With modern technology and ever-reducing costs, population based case-control studies have come to the forefront of genetic epidemiology. A case-control study is a kind of epidemiological study which takes a group of subjects with a disease and a group of subjects who are disease-free. Statistical analyses compare physical and environmental factors such as age, weight and exposure to cigarette smoke and look for significant differences between the two groups of subjects. Any factors which appear to be significant are associated with the disease. They may be having a causal effect on the disease themselves, or they may be associated in some way to another factor that is causal. Similar methods can be applied to genetic data by seeing whether genotypes tend to be different in cases compared to controls.

1.2.1 Genome-wide association studies and fine-mapping

The human genome has now been fully mapped (ie. the whole sequence has been read).As well as the Human Genome Project [4] (completed in 2003), which sequenced all of the DNA for one person, the 1000 Genomes Project [8], which aims to characterise genetic variation in populations, is well under way. In addition to this, smaller regions of DNA have been sequenced in more people for smaller projects. Therefore, we now know the location of many SNPs, in particular those with MAF >0.005, and information about these is widely available from online databases such as dbSNP [3].

A DNA microarray, also known as a chip, is a piece of equipment used to identify the bases at specific locations in the DNA. These can be targeted so that they are the locations of known SNPs, allowing testing for just those nucleotides, rather than the whole of a subject's DNA, saving time and money. n recent years this technology has been used to carry out many genome-wide association studies (GWAS). These are case-control studies, often population-based, in which association between the alleles of a large number of SNPs located throughout the genome and a particular disease are analysed.

In these kinds of studies we can also use the LD structure to our advantage. Software such as Haploview (§1.1.3) can be used to analyse the LD structure of the SNPs that may be of interest in the study. One of the things that the software does is to divide the SNPs into LD blocks. If these groups of highly correlated SNPs can be identified, then only one of them needs to be included in a GWAS and it represents or 'tags' the other SNPs as well. Plots of the type in Figure 1.3 can help with the choice of tagSNPs for genotyping studies, as large blocks of dark red represent LD blocks. From these blocks, the alleles of only one or two SNPs need to be genotyped to infer the rest with high accuracy. In most GWAS the arrays used collect information on between 500,000 and 1,500,000 SNPs, a small proportion of all the SNPs in human DNA.

A GWAS usually has a fairly large sample size, for example, 5000 cases and 5000 controls for a specified disease, and very often uses a standard, mass produced genotyping chip which tests for the alleles of SNPs thought to tag most of the common SNPs in the genome. The results of genotyping can be output as binary data (major/minor allele at each locus) which can then be analysed statistically. The main analysis often takes the form of a Cochran-Armitage test (a modified χ^2 test) for association between each SNP and the disease.

These tests output p-values, but because so much multiple testing occurs when analysing so many SNPs, only those with $p < 5 \times 10^{-8}$ are considered to have genome-wide significance [30].

Generally, a genomic region is only considered to be confirmed as associated with a disease when associations have been demonstrated in at least two study populations. It could be that a single SNP, or multiple SNPs that are located close together, are found to be be associated with the disease, but this does not necessarily mean that any of these SNPs cause the disease. The tagging, which is helpful in identifying a region of association, hinders the actual pin-pointing of the causal SNP. Because of the high correlation between SNPs that is likely to occur in the associated region of DNA, we can only be sure that there is a variant somewhere in the region which affects the risk of the disease.

The purpose of fine-mapping is to look in more detail at a known area of association, analysing more SNPs in that region. To get more comprehensive results, it is necessary to use even larger samples, perhaps a total sample size of 50,000 or more. Whilst more SNPs may be directly genotyped, it is unlikely that they all will be, so sophisticated software may be used to impute the alleles of any other known SNPs in the region [33]. Carrying out association analysis again will still not necessarily highlight the true causal SNP as the most significant, though. The SNPs being analysed are now going to be even closer together and more highly correlated. As well as this, the disease in question is likely to be complex with lots of risk factors, both genetic and environmental, with interactions between some of these risk factors. Therefore, the causal SNP may only have a small effect size (for example, a causal allele relative risk of 1.2 or smaller) and this will be very difficult to identify even in very large sample sizes.

Analysis is further complicated by the fact that causal effects may take several different forms, which are modelled differently. There is also the possibility that the disease may not be affected by a single SNP in the region of interest but by several, making the situation even more complex. These multiple SNPs may or may not be in LD with each other, and may or may not interact with each other. If they do interact, there are several possible types of interaction. Different analysis techniques will pick up different effects and may miss others.

This project reviews some statistical methods used for fine-mapping and attempts to develop novel methods for locating causal SNPs, or at least significantly reducing the set of candidate causal SNPs in a region. The main focus is methods for identifying a single causal SNP within a known association region.

1.2.2 Statistical modelling

Logistic regression modelling

One of the main techniques used to analyse the association between single SNPs and disease outcomes is logistic regression modelling. This kind of model is common in statistical analysis within many areas, and is used to analyse data where there are only two possible outcomes. Therefore, it is suited to genetic disease analysis where the outcome of interest is simply whether or not the disease is present. It would not be suitable if the outcome of interest was, for example, disease severity, or time from diagnosis to death. To model these, another method such as linear regression may be used.

The two possible outcomes are coded as 0 and 1 and the model predicts the probability of outcome 1 occurring under certain conditions, which are represented by independent model variables. The usual way to code a disease model is 0 for no disease and 1 for disease present, so the model predicts y, the probability (absolute risk) of the disease occurring dependent on conditions which may be genetic, environmental or a mixture of the two.

The model can be considered in two parts, the first being the logistic (or logit) link:

$$y = \frac{e^z}{1+e^z} \tag{1.1}$$

The value of z is calculated using the linear predictor:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k. \tag{1.2}$$

This kind of model allows any number and type of independent variables to be input as the values of the x_is and the logit link will always transform the output to a value between 0 and 1. The maximum likelihood estimates (MLEs) of the regression coefficients can be calculated by fitting the model in a statistical software program, such as R [41], and the MLE of β_i is written $\widehat{\beta_i}$. For any x_i, the corresponding β_i is the natural log of the odds ratio (OR) for a unit increase in that variable, when all other variables are kept the same. We refer to this as the logOR. Although this is quite difficult to interpret in itself, a positive β_i means that an increase in the variable x_i will increase the probability, y, of the outcome of interest, in this case the risk of the disease. Conversely, a negative β_i means that an increase in the variable x_i will decrease the probability of this outcome.

Using logistic regression to model marginal genetic effects

As discussed in §1.1.2, a person can carry 0, 1 or 2 copies of the minor allele at a SNP depending on whether they inherited it from neither, one or both parents. The genotype is homozygous if the alleles are the same (in the case of a causal SNP, homozygous wildtype indicates no risk alleles and homozygous risk indicates 2 risk alleles), and heterozygous if there is one of each allele present. Because different parts of the genome carry out biological processes in different ways, some causative alleles may have different kinds of effects to others. Although there are some exceptions, most effects can be classified as either additive, recessive or dominant. A SNP has an additive effect if each copy of the risk allele that is present increases the disease risk by the same magnitude. If, however, the risk only increases when there are two risk alleles present, the effect type is recessive and the odds, and therefore the ORs and logORs, are the same for homozygous wildtype and heterozygous genotypes. Finally, a dominant effect is one for which the increase in the risk is the same no matter whether there are one or two risk alleles present, because the risk allele is dominant over the wildtype allele.

The following form of logistic regression model is used throughout the thesis. It analyses the association between the disease and each SNP individually. This is done by modelling the probability (y_{ij}) of subject j having the disease dependent on their genotype at SNP i, as well as on any covariates, using a logistic regression model, with linear predictor

$$z_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \beta_{2i}v_{2j} + ... + \beta_{ni}v_{nj}. \tag{1.3}$$

In this model, x_{ij} is the number of copies (0, 1 or 2) of the allele coded '1' (usually the minor allele) for SNP i present in subject j. Therefore, the model is an additive model and β_{1i} is the per-allele log odds ratio (logOR) of disease for the allele coded '1' compared to the allele coded '0'. The covariates are given in the model by v_{hj} and may be measured environmental factors or principal components derived to account for population diversity in the study sample. When we use simulated data (see §1.2.4), there are no covariates to include in the model.

1.2.3 Statistics used to compare variants

Hypothesis testing and p-values

A common method of analysing data is to carry out a statistical hypothesis test. There are many such tests, but they all require the specification of a null (H_0)

Disease status	Genotype			Row totals
	Homozygous wildtype (0)	Heterozygous (1)	Homozygous risk (2)	
Control (0)	N_{00}	N_{01}	N_{02}	R_0
Case (1)	N_{10}	N_{11}	N_{12}	R_1
Column totals	C_0	C_1	C_2	N

Table 1.1: Contingency table of genotype and disease status, as required for Cochran-Armitage trend test.

and an alternative (H_1) hypothesis. In considering the association between a single genetic variant and a disease, a suitable null hypothesis would be that that there is no association, and the alternative would be that there is some association. Using the notation in Equation (1.3), $H_0 : \beta_{1i} = 0$; $H_1 : \beta_{1i} \neq 0$.

Typically, hypothesis tests output an observed value of a test statistic. How extreme this value is relates to how unusual the data is, if it is assumed that the null hypothesis is true. The probability of obtaining a value as extreme, or more extreme, than the one observed is called the p-value. A very small p-value is evidence that the data is unlikely to come from the population described by the null hypothesis and in a hypothesis test, a significance level will be pre-specified. If the p-value is below this level, the null hypothesis will be rejected.

In genetic association, p-values are often obtained from Wald tests or Cochran-Armitage tests. For the Wald test, the test statistic is

$$\frac{\widehat{\beta_{1i}} - \beta_N}{var(\widehat{\beta_{1i}}),} \tag{1.4}$$

where β_N is the value of β_{1i} given by the null hypothesis, in this case 0. The Cochran-Armitage test does not require the fitting of regression models, but summarises the different combinations of genotypes and disease status in the data. Using the values given in Table 1.1, the trend test statistic is

$$\sum_{g=0}^{2} w_g (N_{0g} R_1 - N_{1g} R_0) \tag{1.5}$$

where w_g are the weights of the different genotypes, so that for an additive effect the weights $(w_0, w_1, w_2) = (0, 1, 2)$ are used. The test statistics for both of these tests are compared to a χ^2 distribution to obtain the p-values. In this case the relevant distribution is that with 1 degree of freedom. A disadvantage of using p-values from Cochran-Armitage tests compared to statistics that come from regression models is that there is no way to adjust for covariates.

It is common in GWAS studies to rank SNPs based on how small their p-

values from one of these tests are, and these methods are also sometimes used in fine-mapping studies [36] [7]. Often, only p-values below a particular value are considered to be significant. In a single statistical test, it is common, for example, to consider the test significant if $p < 0.05$, because this indicates that the null hypothesis is only likely to be true for 1/20 samples with as extreme or more extreme results than were observed. However, consider testing around million SNPs in a GWAS. The number of observed data points which have $p < 0.05$, but for which the null hypothesis is in fact true, is increased by the number of SNPs tested. Therefore the level of significance is adjusted accordingly, often using a Bonferroni correction, such that $p < 0.05/s$ is considered significant, where s is the number of SNPs [30]. With a million SNPs, this threshold becomes $p = 5 \times 10^{-8}$. We do not, however, consider any particular significance level in the original work, but only rank the SNPs from the smallest p-value.

Likelihood

The likelihood of a set of parameter values in a logistic regression model ($\boldsymbol{\beta} = (\beta_{0i}, \beta_{1i}, ..., \beta_{ni})$) is the probability of the observed data having occurred giving these are the true parameter values. If $\boldsymbol{x}$ is the data, then the likelihood, $\mathcal{L}(\boldsymbol{\beta}|\boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{\beta})$. In the case where there are no covariates, this can be written

$$\mathcal{L}(\beta_{0i}, \beta_{1i}|\boldsymbol{y}_i) = \prod_{j=1}^{n} p(x_{ij})^{y_{ij}}(1 - p(x_{ij}))^{(1-y_{ij})}. \tag{1.6}$$

When regression models are fitted, the MLEs of the regression coefficients, $\widehat{\boldsymbol{\beta}}$, are chosen such that the likelihood of the model is as large as possible, according to an algorithm implemented by the program.

The SNPs in the models with the largest likelihoods can be considered as the most likely to be causal. A method which is sometimes used in fine-mapping [17] [52] [53] [19] to identify the SNPs to take forward for further analysis is to identify the largest of all likelihood values and take forward those SNPs for which the model parameters have a likelihood within a particular ratio of this value, for example within 1/100.

LD

As previously mentioned, it is unlikely that a SNP which is a 'hit' in a GWAS (usually based on it having a p-value below genome-wide significance), is actually causal. It is more likely that it is in high LD with the the causal variant. Therefore, when candidate causal SNPs are chosen, it is normally all those that have r^2 above a particular threshold with the tagSNP. Recently, a method was published that scores SNPs based on preferential LD (PLD) [62]. This not only takes into account the LD between each untyped (not genotyped) SNP and the hit tagSNP, but also between that SNP and all the other tagSNPs. If an untyped SNP is in fact causal, it is likely that all tagSNPs in high LD with it will be hits. This method allows for SNPs to be taken forward based on them being in higher LD with the hit than other tagSNPs, thus taking more available information into consideration.

Bayes factors

Bayesian methods of statistical analysis allow for the inclusion of data other than the data gained directly from the main study in the analysis. In a genetic association study, the main data are the genotypes but prior data from previous studies could also be incorporated.

The Bayes Factor is a Bayesian hypothesis testing method defined as the ratio of the probability of observed data having occurred under the alternative hypothesis to the probability of it having occurred under the alternative hypothesis:

$$\mathrm{BF} = \frac{P(data|H_1)}{P(data|H_0)}. \tag{1.7}$$

It is derived from Bayes theorem and is the factor by which a prior odds should be multiplied to obtain a posterior odds [28], where δ is defined as the prior probability and Δ as the posterior probability, $\Delta/(1-\Delta) = \delta/(1-\delta) \times \mathrm{BF}$. In the case of genetic association, these are the prior and posterior probabilities of a SNP being associated with the disease. However, when BF is used to prioritise SNPs in fine-mapping, it is common to assume that all SNPs have equal prior probability [48] [32].

Analysis of multiple variants simultaneously.

There are several methods used which analyse multiple variants together, for example by including them all in a single model, rather than one at a time in individual logistic regression models. This is particularly useful if is possible that there are multiple causal SNPs in a region and may allow for the modelling of interaction between variants. Several such methods have been reviewed by Ayers and Cordell [9] and Abraham *et al.* [6]. One popular method is penalised logistic regression and this is often implemented using HyperLASSO [27]. Bayesian methods can be implemented through the pi-MASS software [23].

1.2.4 Software and computing aids

Snptest, Hapgen2 and Impute2

These are the up to date versions of three pieces of software from the Oxford genome-wide analysis software suite. Snptest [33] is able to carry out a number different standard and more specialised single SNP analyses on genotype data.

Hapgen2 [47] [50] is a computer program that simulates haplotypes for a case-control study. It uses a reference set of haplotypes to determine the general structure of the region of interest in terms of MAFs of SNPs and LD structure and simulates new haplotypes based on this structure. The reference haplotypes used throughout this project are from the 1000 Genomes Study [8]. The software requires input in terms of the start and end base-pair numbers of the region to be simulated, the location of the selected disease causing SNP(s), the relative risks for heterozygous and homozygous causal genotypes and the number of case and control haplotypes to be simulated. The output includes binary haplotype data for the required case and control subjects as well as information on the SNPs included.

Impute2 [34] [33] is a program written by the same group of researchers which uses similar methodology to impute missing genotype data. For a given subject with some genotypes known, it uses these and the same sort of reference data to determine the probabilities of the 3 possible genotypes at the remaining loci.

These programs use the same file formats, making it convenient to use them in combination with one another.

Haploview

Haploview [10] is a piece of software which uses a reference set of haplotypes (again the 1000 Genomes data was used for this project) to calculate pairwise r^2 and D' values for all SNPs in a selected region. It also has features which generate various heat-map style diagrams to display this information visually (see §1.1.3).

R

R [41] was used for all statistical analyses not specific to any other package mentioned.

iceberg

Much of the computer intensive work was carried out on iceberg, the Linux based High Performance Computing Cluster at the University of Sheffield [5]. This includes the simulation and imputation carried out using Hapgen2 and Impute2, some of the analyses in R and some image generation using Haploview.

1.3 Introduction to the original research

Genome-wide association studies (GWAS), as described in §1.2.1, and candidate gene studies (which investigate a gene known to be involved in a mechanism which could be involved with the disease of interest) have highlighted regions of the genome containing variants affecting disease susceptibility. The identification of the causal variants in such regions is confounded by high correlation between variants so close together in the chromosome. Because of this correlation (LD) and the effect of sampling variation, when tests of association are carried out at a fine-mapping level, the causal variant will not necessarily be the variant with the largest likelihood or smallest p-value.

This project considers the use of a single general methodological framework for fine-mapping analysis, which is referred to as filtering. Rather than highlighting a single variant and suggesting that this is causal, filtering works by removing as many variants as possible to leave a smaller group of candidate causal variants. This project investigates several ways of carrying out statistical filtering and analyses how successful different statistics (called filters) are at reducing the initial set of variants as far as possible whilst still retaining the causal variant with high probability. Although filtering alone is unlikely to be able to identify a causal variant, the hope is that it will provide a vital step by identifying the best set of variants to take forward for further biological testing. These techniques, which may include the biological analysis of pathways in cell lines or animal models, are expensive, so it is highly important to reduce the number of variants to be tested to as few as possible. Single nucleotide polymorphisms (SNPs) are considered to be the variants of interest throughout this work, but any other type of genetic variant able to be modelled using logistic regression could be analysed in this way.

1.3.1 Statistical analysis of genotype data and the filtering framework

Using the `glm` command in R [41], we fit models as described in §1.2.2. This can be used to obtain, for $h \in (0, n)$, $\widehat{\beta_{hi}}$, the maximum likelihood estimate (MLE) of β_{hi}, the variance of this estimate and the likelihood of the model, which we refer to as the likelihood for SNP i, and denote $\mathcal{L}_i$. Statistical tests such as the Cochran-Armitage test and Wald test can also be carried out on the raw data or model parameters to obtain p-values.

Filtering is our terminology for the ranking of SNPs based on a chosen statistic and the removal of all SNPs with a value of that statistic below a pre-specified threshold. All remaining SNPs are considered causal candidates and will be taken forward to any future analysis, but any SNPs that have been removed are considered very unlikely to be causal. We investigate the use of several filtering statistics which all use the output from the described logistic regression models in their calculation. As filtering statistics are obtained by modelling the association between each SNP and the disease individually, this methodology will not always be applicable in fine-mapping. We assume that there is a single causal SNP in an association region, but if there are multiple interacting causal SNPs, using filtering may not be effective. It will still help to find the SNP with the largest marginal effect, but in a scenario where multiple causal SNPs have strong interactive effects, their marginal effects may only be small and in this case they have a high probability of being removed by filtering.

1.3.2 Assessing the efficacy of different methods

Simulating data with which to test methods

We used Hapgen2 [47] [50] in conjunction with the European haplotypes of the August 2010 release of the 1000 genomes data [8] to simulate case-control genotype data on which we were able to test different filtering methods (see §1.2.4), so they should give a good representation of real data and the problems associated with them, such as high levels of short range LD and low MAF. As the causal SNP is user-specified in simulation, it is possible to carry out filtering and know whether or not the causal SNP is retained. To evaluate the sensitivity of the results to the location of the causal SNP, its MAF and effect size as well as the sample size, we simulated data with a variety of combinations of these features, and we refer to each unique combination as a scenario. We focus on small effect sizes (usually, additive effects with per-allele ORs of 1.06 to 1.36) as it is expected that many yet to be investigated loci are likely to have such effects. Although results are not shown for every scenario, we considered many combinations of the ORs given above, causal SNPs with MAFs from 0.05 to 0.43 and sample sizes from 4000 to 50,000. We generally refer to the total sample size, with equal numbers of cases and controls. For each scenario multiple independent datasets (usually 1000) were simulated to take into account random sampling and to give a more complete assessment of filter efficacy.

The majority of simulated genotype scenarios were based on a region of chromosome 2 which evidence suggests may include variants which have a small effect on several complex diseases including breast cancer [14] [11] [39] [25] [16]. In the studies which found this evidence, possible association with tagSNPs only just, or nearly, reached genome-wide significance. This suggests that while there may be an associated SNP, it is likely to have a small effect on the risk of the disease, so current analytic techniques may not be powerful enough to find it. The simulated region is one mega-base in size (1Mb, one million bases), ranging from 201,566,128 to 202,566,128 bases in the Hg19 build of chromosome 2 and contains 2871 SNPs in the 1000 genomes data. This region includes the *CASP8* gene, as well as around 20 other known genes including *CASP8* homologues *CFLAR*, *CASP10* and several *ALS2CR* genes.

CASP8 appears to be a suitable candidate for a gene which could affect cancer susceptibility as it codes for caspase 8, a protein that is involved in the biological process of apoptosis, or programmed cell death. Cells in the body are constantly being replicated, so it is necessary that some should die. Also, if a cell is damaged in some way, it is much better that the cell dies, reducing the chances of any damage being passed on to other cells. If apoptosis does not happen then the replicating cells cannot be controlled and may form a tumour. If one of the biochemical substances involved in the normal apoptosis process, such as caspase 8, does not function correctly, it is feasible that a cancer could be caused in this way, and one reason for caspase 8 not to function correctly could be that it is encoded differently to normal due to a variant in the *CASP8* gene.

Output of simulation analyses

Successful filters will reduce the initial set of SNPs down to a much smaller group in which it is highly probable that the true causal variant remains. Equivalently, they will have a low false positive rate (FPR) and a high true positive rate (TPR). To display the FPR and TPR at each threshold graphically, we present the results of filtering for a single simulated scenario as a receiver operating characteristic (ROC) curve. The TPR at a single threshold is plotted on the

y-axis against the FPR at that same threshold on the x-axis. ROC curves are commonly used for plotting the outcome from a single dataset, but we have multiple simulated datasets of the same scenario. There is not a single standard method of combining such results into one ROC curve, and this is discussed in a paper by Fawcett [18]. Of the three possible methods he describes, we use 'threshold averaging'. The FPR is the proportion of non-causal SNPs retained after filtering, which, unless stated otherwise, will be the mean FPR across the datasets of one scenario. We also use the mean TPR, the proportion of datasets in which the known true causal SNP is retained. Fawcett highlights the fact that combining the results in such a way will result in variation around the means that are presented. In the case of the FPR, this variation is due to the range of FPR values in the set of datasets. On the other hand, the TPR is the proportion of a sample of n datasets which have a binomial outcome. Therefore, the variance of the TPR is given by $\text{TPR}(1 - \text{TPR})/n$.

On some plots, the TPR and FPR at a specific filtering threshold have been highlighted as a point on the ROC curve. A $y = x$ line is often included to indicate the outcomes that would be expected if the removal of SNPs using this filter was random. Some of the figures that are given are not the full ROC curves, but specific parts that are of particular interest. This is the case if the range of both axes is not (0,1).

Area under the curve (AUC) is a common measure of a classifying method, such as filtering. Where we state AUCs, they are given as a percentage of the total possible area. An AUC of 100% would indicate that at any given threshold, the FPR = 0 and the TPR = 1. An AUC greater than 50% indicates that a ROC curve must at some thresholds lie above the $y = x$ line and therefore the filtering method is better than the removal of random SNPs. However, AUC does not give any indication about how the filter performs at specific thresholds, or the TPR at a given FPR, and many different shapes of ROC curve may result in the same AUC. We have calculated mean and variances of the AUCs using the method described in a paper by Valdar *et al.* [55]. If a curve is based on the results of a set of simulated datasets, $D_1, ..., D_d$, then $\text{mean(AUC)} = d^{-1} \sum_{i=1}^{d} \text{AUC}(D_i)$ and $\text{var(AUC)} = (d-1)^{-1} \sum_{i=1}^{d} (\text{mean(AUC) - AUC}(D_i))^2$. The confidence interval (CI) around a mean(AUC) is calculated using $100(1 - \alpha)\%\ \text{CI} = \text{mean(AUC)} \pm \Phi^{-1}(1 - \alpha/2)\sqrt{\text{var(AUC)}}$.

1.3.3 Applying the methods to a real dataset

Chapter 4 consists of the application of the most appropriate methods investigated and developed during this project to a real dataset from the Collaborative

Oncological Gene-Environment Study (COGS) [2]. This study has developed an Illumina genotyping array (the iCOGS chip) concentrating a large number of target SNPs in regions of the genome thought to be associated with several types of cancer, including ovarian, breast and prostate cancers [35]. By being selective in this way, it allows regions of particular interest to be studied in more detail. In general, these association regions are analysed separately using fine-mapping techniques.

COGS is a collaborative study involving seven consortia all of which have been investigating the genetic contribution to the risk of the cancers of interest. Each consortium selected regions of the genome that they wished to include in the study and the SNPs which should be genotyped from those regions. One of the regions selected by the Breast Cancer Association Consortium (BCAC) [1] comprises base positions 201500074 to 202569992 of chromosome two, and contains the gene *CASP8*, which is described in §1.3.2.

BCAC originally selected 585 SNPs in the *CASP8* region to be genotyped using the iCOGS chip and 501 passed quality control checks. IMPUTE2 [33] was used to sucessfully impute the genotypes of a further 1232 SNPs, resulting in 1733 SNPs in total. The sample consisted of 89,050 subjects, with 46,450 cases and 42,600 controls.

Chapter 2

Filtering using only association study data

2.1 Comparison of filtering statistics

Consider a fine-mapping study of region that is assumed to have a single causal SNP. There is no current gold-standard analysis method, but a statistic which has been used in several studies [17] [52] [53] [19] is something we call relative likelihood (RL). It is often referred to as the "likelihood ratio", but we avoid using this term to remove any confusion with the ratio used in the likelihood ratio test, a standard statistical method. To determine how suitable RL analysis is, we have carried out a thorough simulation study comparing it to several other statistics. All of the statistics that are examined in this chapter could be used as filters to remove non-causal variants from the set of all candidate causal variants. We refer to these as filters which use the genotype data only, as they do not require the input of any external functional data. However, imputed genotype doses may also be used and this practice is scrutinised in §2.5. This set of methods has not previously been compared in a thorough simulation study such as this. Each variant is analysed separately and they are then ranked in some way based on their probability or likelihood of being causal. Filtering is simply the removal of all variants from a dataset with a value of the relevant statistic below a pre-specified threshold. So when carrying out RL filtering at threshold of 1/100, only variants with RL $> 1/100$ are retained to be included in any further investigation or analysis.

The filters were tested on simulated datasets generated using Hapgen2 as described in §1.3.2. The *CASP8* region of chromosome 2 which we describe in that section was used, with 2871 SNPs identified in the 1000 genomes data [8]. To assess the effect of LD structure on filter efficacy, we also used two other chromosomal regions, chosen to have very different structures to each other and to the *CASP8* region. This region of chromosome 2 has a mixture of both large and small LD blocks, with an average size of approximately 22kb in length. A study by Smith *et al.* [45] contained results which we were able to use to select regions which had particularly high and particularly low levels of LD. We chose a region of chromosome 11 from 55Mb-56Mb (11q11-12), which has LD blocks with an average size of approximately 130kb and contains 6247 SNPs. The other region we chose, due to it having minimal LD, is located in chromosome 16p13 (9-10Mb), and has LD blocks with an average size of 8kb and 6200 SNPs in total.

Most datasets were simulated using a causal SNP with an additive effect, varying the per-allele odds ratio (OR) between 1.06 and 1.36. Two regions which have been analysed using RL filtering in previously published studies are *FGFR2* and *16q12* [52] [53] [17]. The effect sizes in these regions are at the top of this range, but effect sizes for SNPs involved in complex diseases can be much smaller and are therefore generally more difficult to detect. Some scenarios were also simulated with causal SNPs with different types of effect (both dominant and recessive), to investigate whether such SNPs could also be identified using the filtering framework.

2.1.1 Definitions of filters

Likelihood filters

The relative likelihood (RL) [17] [52] [53] [19] for the i^{th} SNP is the ratio between the likelihood for that SNP and the largest of the maximised likelihoods over all p SNPs in the region:

$$RL_i = \frac{\mathcal{L}(\widehat{\beta}_{0i}, \widehat{\beta}_{1i}|data)}{\max\limits_{k \in \{1,p\}} \{\mathcal{L}(\widehat{\beta}_{0k}, \widehat{\beta}_{1k}|data)\}}. \tag{2.1}$$

This results in $RL_i \in (0,1], \forall i \in \{1,...,p\}$. The SNP which has an RL of 1 is referred to as the 'top hit' or SNP$_{max}$. SNPs are ranked by RL and a threshold of a pre-specified ratio is applied. All SNPs with likelihoods within this ratio of that of SNP$_{max}$ are retained in the set of candidate causal variants. Previously, an RL filter threshold of 1/100 had been used in published fine-mapping studies [52] [53], meaning that only SNPs with RL $\geq 1/100$ are retained. We consider the suitability of this and other RL thresholds for filtering.

We suggest an alternative method of filtering also based on likelihood to overcome a potential shortcoming of RL filtering. This is the fact that the number of SNPs retained is subject to a large amount of variation, dependent on how extreme the likelihood of SNP$_{max}$ is. Instead, we suggest ranking the likelihood values for each SNP and using a proportion of SNPs as the filter, and call this the likelihood percentile (LP) filter. We may specify a threshold of 95%, meaning that the top ranked 5% of SNPs will be retained. Therefore, the number of SNPs retained is not subject to variation and so any extreme effects of sampling variation are reduced.

p-value filters

An alternative to the likelihood is to use a p-value from a test of association. It is common in GWAS to rank SNPs by p-values either from Cochran-Armitage trend tests or from Wald tests and both of these methods have now also been used in the context of fine-mapping [36] [7]. For SNP i it is simple to carry out such tests and we refer to the p-value from the Cochran-Armitage test as p_{CA}

and that from the Wald test as p_W. These can be used as filtering statistics by choosing a threshold value which may based on a Bonferroni correction, for example.

Cochran-Armitage and Wald tests are both hypothesis tests, where the null hypothesis in this case is no association between a SNP and the disease. The p-value indicates the probability of encountering data as or more extreme than that which was observed, given that the null hypothesis is true. Whereas the likelihood is the probability of encountering the observed data, given that the regression coefficients in the model are correct, the Wald test is a test for whether these regression coefficients are statistically significantly different from 0. Another common hypothesis test is the likelihood ratio test, which compares the likelihoods of nested models. This test tends to give rankings very similar to those of the other hypothesis test we consider. However, the models we wish to rank (each containing one SNP) are not nested, so the RL and LP methods are not equivalent to a likelihood ratio test and it is possible that ranks based on p-value and likelihood will differ. These methods are described in more detail in §1.2.3.

Structural filters

The remaining methods relate to LD structure. Within a small chromosomal region, linkage disequilibrium (LD) can be high between SNPs. When the top hits from GWAS are found, these are not assumed to be the causal SNPs, but it is often postulated that the causal SNP lies within the same gene or LD block as the tagSNP. Alternatively, a handful of candidates may be suggested based on high LD with the tagSNP ($r^2 > 0.9$, for example).

We formalise three filtering methods based on these ideas. SNPs were ranked by either genetic map distance in centimorgans (cMs) from SNP$_{max}$ or by pairwise D' or r^2 values with SNP$_{max}$ (see §1.1.3). Genetic map distances were obtained from the 1000 genomes data [8] and pairwise LD (D' and r^2) values were calculated using the simulated haplotypes. Once again, thresholds were specified so that SNPs further away in distance or with lower LD values than those thresholds were filtered out.

The final filtering method comes from a paper by Zhu *et al.* [62] and was also based on r^2 between each SNP and SNP$_{max}$. For this method, rather than ranking based on that single r^2 value, a preferential LD (PLD) score was calculated for SNP$_i$. Although we use the analyses set out by Zhu *et al.* [62], we use it in a slightly different setting, as it is designed for use with GWAS data, making use of the panel of tagSNPs from the genotyping array. PLD_i

is the proportion of tagSNPs for which r^2 between them and SNP$_i$ is greater than between SNP$_{max}$ and SNP$_i$. For the simulated regions, since all SNPs have been 'genotyped', we chose to use those on the Illumina 300 array in the regions we were fine-mapping as our tagSNPs. To complete the Zhu method, a second filtering step is required, which involves calculating an empirical p-value testing the r^2 value between SNP$_i$ and SNP$_{max}$ [62]. Specifically, this p-value 'estimates the probability of observing the same or better r^2 value for two random variants with the same frequencies' [62]. This is done by permuting the genotypes 2000 times in each dataset.

2.1.2 Relative efficacy of different filtering methods

We give the results of filtering using the different methods on 1000 datasets simulated using the 11q11-12 region as ROC curves in Figures 2.1(a) and 2.1(b). This is the region with overall high levels of LD and the scenario simulated was a sample size of 20,000 and a causal SNP with OR 1.1 and MAF 0.08. We have split the methods into those that are p-value and likelihood based in Figure 2.1(a) and those that are proximity and LD based in Figure 2.1(b). Figures 2.1(c) and 2.1(d) are the ROC curves of the equivalent filtering outcomes in the mixed LD (*CASP8*) region and Figures 2.1(e) and 2.1(f) are those of the low LD (16q13) simulated datasets. Similar causal SNP scenarios were used in all simulated regions. Something to note is that only one of the p-value methods (specifically p_{CA}) is shown as this gave very similar results to the alternative, p_W. When analysing real data these will not be the same as the Wald test takes into account the effect of covariates, whereas the Cochran-Armitage test does not. Therefore, it may be more advantageous to use p_W in a real study. Also, the Zhu method (PLD) was only tested on a subset of 100 datasets. The second step of this method involves permuting genotypes 2000 times. However, this number of permutations was too computationally expensive when analysing 1000 simulated datasets, as were analysed by all the other methods. For this method there were 77 tagSNPs (from the Illumina 300 array) in both the *CASP8* and 11q11-12 (mixed and high LD) regions and 135 in the 16q13 (low LD) region.

Table 2.1 contains the mean area under the curve (AUC) values, and their 95% confidence intervals (CIs), for all of the ROC curves in Figure 2.1. It is explained how these are calculated in §1.3.2. When filtering a set of thousands of candidate causal SNPs, though, it is important to significantly reduce this set, so the parts of the ROC curves that are most of interest are those which result in the lowest FPRs. Therefore we examine these in more detail in Figure

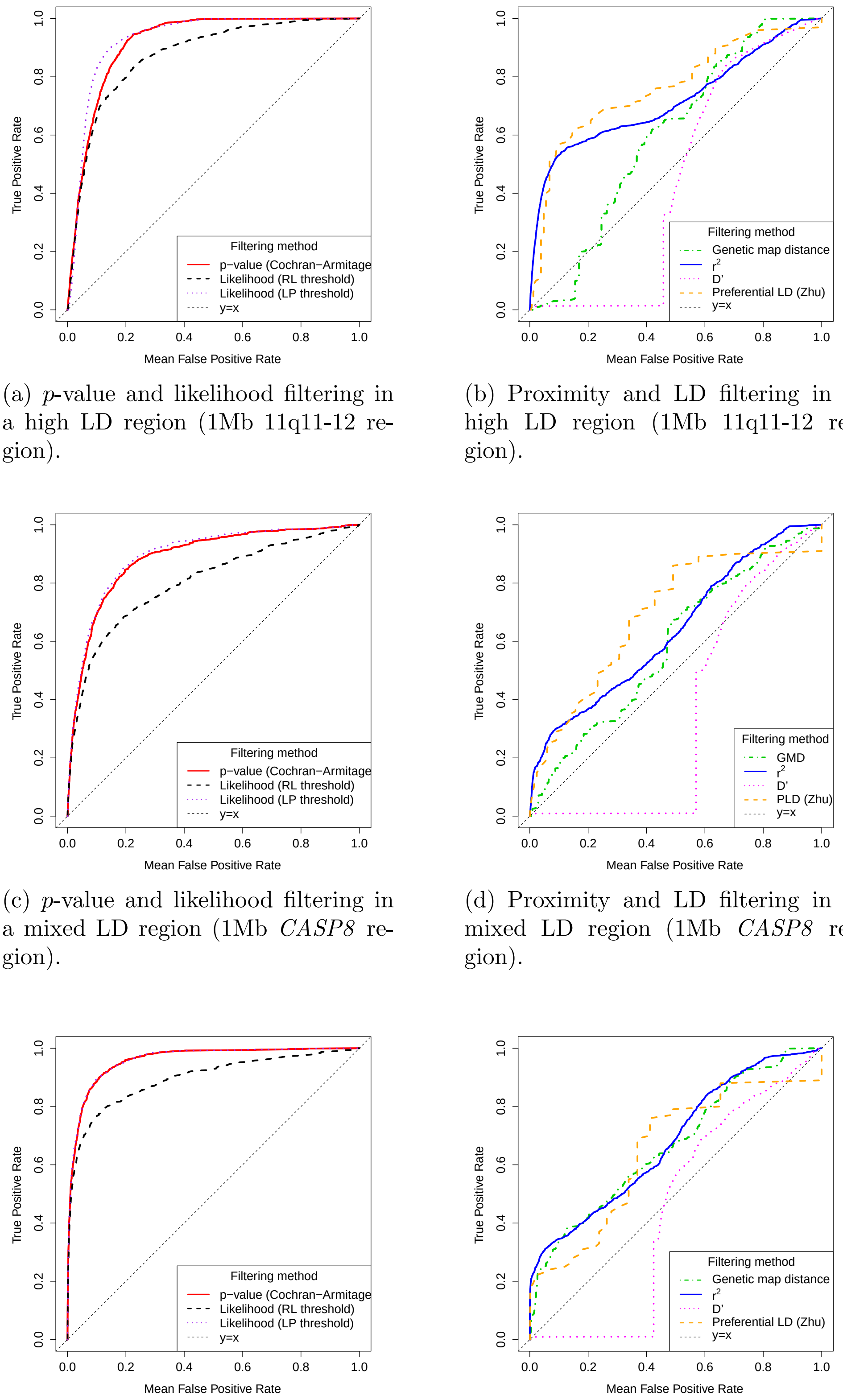
Filtering method	Genomic region High LD	 Mixed LD	 Low LD
Likelihood (LP threshold)	93% *(93%, 93%)*	90% *(90%, 90%)*	96% *(96%, 96%)*
p-value	92% *(83%, 100%)*	89% *(83%, 94%)*	96% *(94%, 97%)*
Likelihood (RL threshold)	87% *(69%, 100%)*	80% *(55%, 100%)*	90% *(75%, 100%)*
Preferential LD (Zhu)	76% *(60%, 93%)*	69% *(37%,100%)*	65% *(32%, 98%)*
r^2	72% *(53%, 91%)*	64% *(35%, 92%)*	68% *(48%, 88%)*
Genetic map distance (GMD)	60% *(53%, 67%)*	59% *(30%, 88%)*	67% *(54%, 80%)*
D'	44% *(6%, 82%)*	35% *(0%, 75%)*	44% *(7%, 81%)*

Table 2.1: Mean *(and 95% CI)* area under curve (AUC, given as a percentage) for ROC curves of different filters. Three different 1Mb regions of the genome were used but in each the causal SNP has an OR of 1.1, a MAF of 0.08 and the sample size is 20,000.

2.2 and Table 2.2, considering the part of the curves for which mean FPR$\leq$ 0.1. It should be noted that the maximum possible partial AUC as given in Table 2.2 is 10%.

It can clearly be seen that the likelihood and p-value based methods are generally more efficacious than the methods which filter based on proximity to, and LD with, SNP$_{max}$ for these scenarios. In particular, D' filtering is often not able to produce FPRs less than 0.4. Of the structural based methods, PLD usually results in the largest AUC overall, but r^2 appears to perform better when FPR $\leq$ 0.1 (Tables 2.1 and 2.2). The likelihood method using LP thresholds resulted in the ROC curves with the highest AUCs, with the AUCs for the p-value methods (p_{CA} and p_W) only slightly lower. The similar efficacies of these methods can also be seen by looking at specific FPRs of interest. For example, in the results of the *CASP8* simulation analyses, where these methods both have an FPR of 0.1 (so approximately 287 of the total 2871 of the total SNPs are retained) the corresponding TPRs are 0.695 for LP filtering and 0.694 for p_{CA} filtering.

Although these three regions were carefully chosen so that their LD structures were all very different, the results, including the AUCs are similar (Table 2.1). In particular, LP gave the best results in all three regions, although the results of p-value methods were similar to these, so if p-values were more readily available, it would be acceptable to use them for filtering. RL filtering was



(a) *p*-value and likelihood filtering in a high LD region (1Mb 11q11-12 region).

(b) Proximity and LD filtering in a high LD region (1Mb 11q11-12 region).

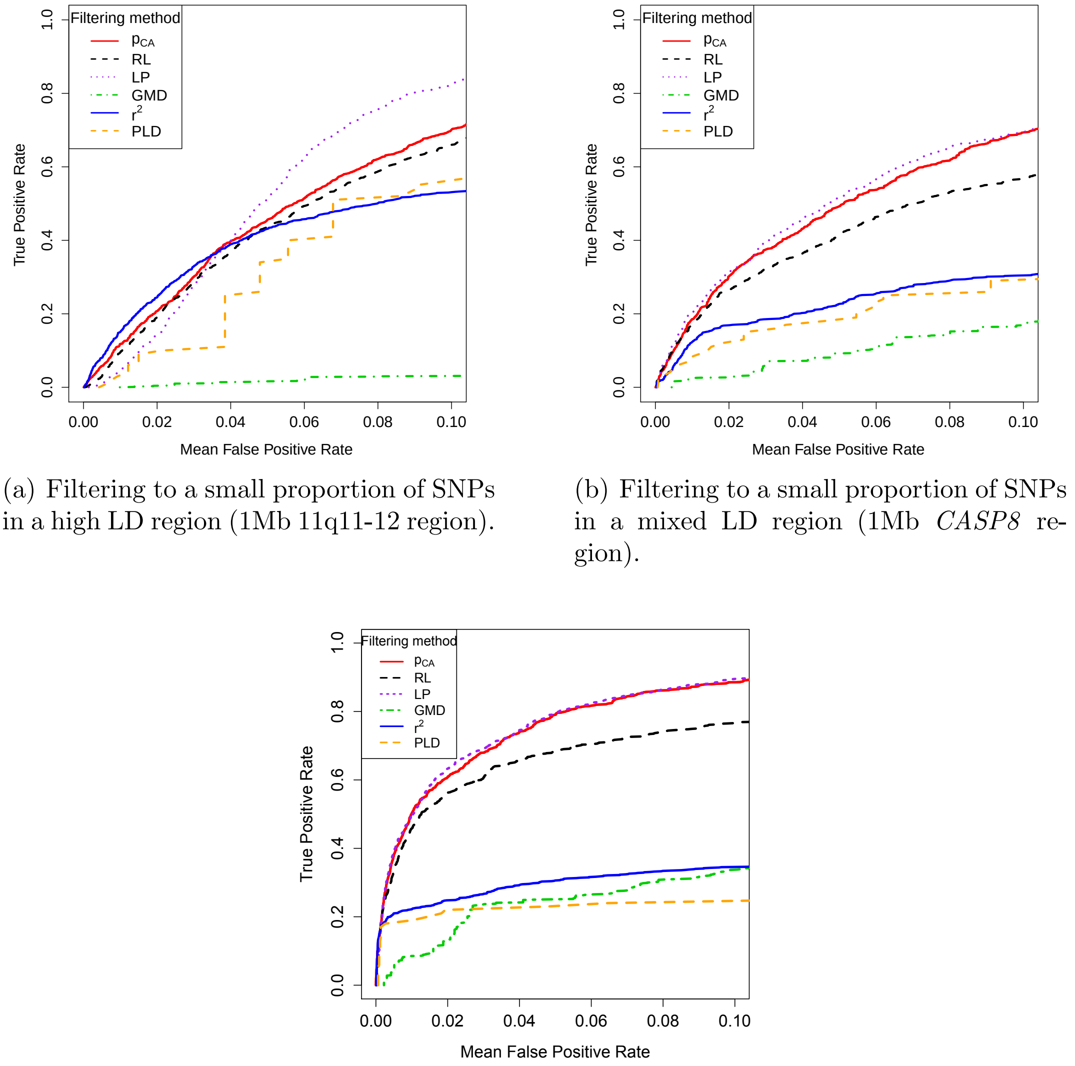
(c) *p*-value and likelihood filtering in a mixed LD region (1Mb *CASP8* region).

(d) Proximity and LD filtering in a mixed LD region (1Mb *CASP8* region).

(e) *p*-value and likelihood filtering in a low LD region (1Mb 16p13 region).

(f) Proximity and LD filtering in a low LD region (1Mb 16p13 region).

Figure 2.1: Comparing the effectiveness of filters for fine-mapped data in 3 regions of the genome. Using the LD structure of each region, 1000 datasets were simulated and then analysed using each method (only 100 were analysed using the Zhu method). Panels (a), (c) and (e) show the efficacy of filtering using ranks and thresholds based on *p*-values from Cochran-Armitage tests (p_{CA}), relative likelihoods (RLs) and likelihood percentile points (LPs). Panels (b), (d) and (f) show the results using genetic map distance (GMD) from and pairwise r^2 or D' values with the top hit and the Zhu method using preferential r^2. The causal SNPs all have an OR of 1.1, a MAF of 0.08 and the sample size is 20,000.



(a) Filtering to a small proportion of SNPs in a high LD region (1Mb 11q11-12 region).

(b) Filtering to a small proportion of SNPs in a mixed LD region (1Mb *CASP8* region).

(c) Filtering to a small proportion of SNPs in a low LD region (1Mb 16p13 region).

Figure 2.2: Comparing the effectiveness of filters for fine-mapped data in 3 regions of the genome, focussing on the results for which FPR$\leq$ 0.1. Using the LD structure of each region, 1000 datasets were simulated and then analysed using each method (only 100 were analysed using the Zhu method). Partial ROC curves show the efficacy of filtering using p-values from Cochran-Armitage tests (p_{CA}), relative likelihoods (RLs), likelihood percentile points (LPs), genetic map distance (GMD) from and pairwise r^2 values with the top hit and the Zhu method using preferential r^2 (PLD). The causal SNPs all have an OR of 1.1, a MAF of 0.08 and the sample size is 20,000.

Filtering method	Genomic region		
	High LD	Mixed LD	Low LD
Likelihood (LP threshold)	4.6% *(4.5%, 4.8%)*	4.7% *(4.4%, 5.0%)*	7.3% *(7.3%, 7.4%)*
p-value	4.7% *(0.4%, 9.0%)*	4.7% *(2.1%, 7.2%)*	7.2% *(6.2%, 8.1%)*
Likelihood (RL threshold)	4.8% *(0.6%, 9.0%)*	4.9% *(1.2%, 8.5%)*	6.9% *(4.3%, 9.5%)*
Preferential LD (Zhu)	2.7% *(0%, 6.8%)*	2.9% *(0%, 6.3%)*	2.5% *(0.4%, 4.7%)*
r^2	4.2% *(0.8%, 7.6%)*	2.8% *(0.2%, 5.3%)*	3.1% *(1.8%, 4.5%)*
Genetic map distance (GMD)	0.2% *(0%, 0.5%)*	1.4% *(0%, 4.1%)*	2.3% *(1.5%, 3.0%)*
D'	0.01% *(0%, 0.1%)*	0% *(0%, 0%)*	0.05% *(0%, 0.5%)*

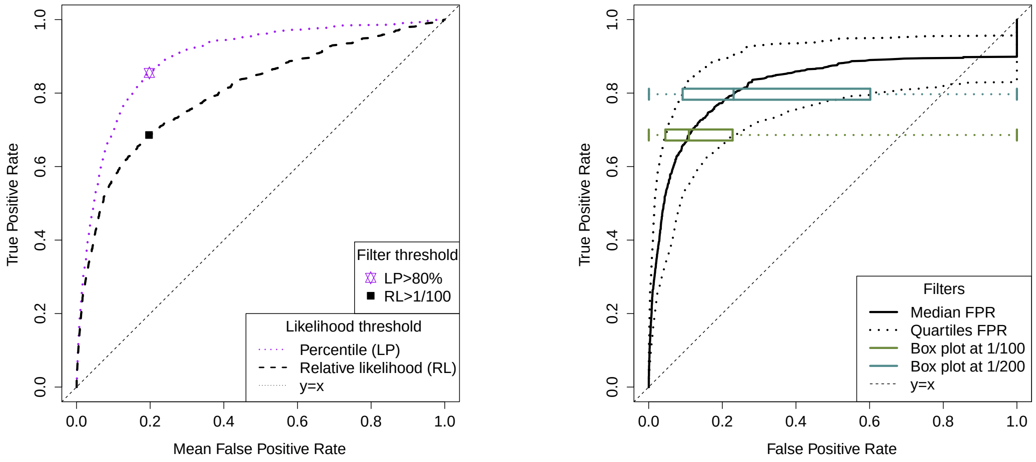
Table 2.2: Mean *(and 95% CI)* area under curve (AUC, given as a percentage) for portions of ROC curves of different filters for which FPR $\leq$ 0.1. Three different 1Mb regions of the genome were used but in each the causal SNP has an OR of 1.1, a MAF of 0.08 and the sample size is 20,000. The maximum percentage of AUC for such a portion is 10%.

always considerably worse than LP filtering, as were all of the LD-based methods, so these should not be used. The results shown in Figures 2.1 and 2.2 are all based on a single sample size, causal SNP OR and MAF. However, we also examined results for other scenarios (see the ranges specified in §1.3.2). The relative efficacies of the filters seem to apply generally within these scenarios, so we would recommend the use of LP filtering over the other methods examined here. We look in detail at the use of LP filtering in some of these scenarios in §2.3.

2.2 Variability in FPR

2.2.1 Relative likelihood filtering

Although we found several published fine-mapping studies [19] [52] [53] [17] using RL filtering, it appears that the somewhat simpler likelihood percentile (LP) filtering method is more efficacious. This is demonstrated by the larger AUCs for the averaged ROC curves and is illustrated in Figure 2.3(a). The common threshold of RL = 1/100 is marked on the RL ROC curve in this figure. For this particular scenario it results in a TPR of 0.686 and a mean FPR of 0.197. To retain the same number of SNPs by using LP filtering, we



(a) LP filtering compared to RL filtering with results highlighted at filtering thresholds which produce equivalent mean FPRs.

(b) RL filtering plotted using the median and quartiles of the FPR with the thresholds of 1/100 and 1/200 highlighted using boxplots.

Figure 2.3: Comparing the effectiveness of RL filtering and likelihood percentile (LP) filtering for fine-mapping data. The causal SNP has an OR of 1.1, a MAF of 0.08 and the sample size is 20,000. 1000 datasets were simulated using the LD structure of the *CASP8* region.

apply a filter of 80%. Therefore, using both methods, a mean number of 568 SNPs will be retained, but by using the LP method, the TPR increases to 0.855.

A further disadvantage to RL filtering is the large amount of variation in the FPR between simulated datasets when using a specified RL threshold, as shown in Figure 2.3(b). This figure shows the results of RL filtering on the 1000 *CASP8* simulations with 20,000 subjects and a causal SNP with an OR of 1.1 and a MAF of 0.08. However, instead of the mean FPR, ROC curves are plotted at the FPR quartiles. Box plots for filtering thresholds of 1/100 and 1/200 are marked on this plot. As previously mentioned, the TPR at an RL of 1/100 (the proportion of the 1000 datasets in which the causal SNP was retained) is 0.686. The median FPR (across the 1000 datasets) is 0.109 but the interquartile (IQ) range of the FPR is (0.045, 0.228) and the full range is (0.0003,1), indicating that as few as 1 or as many as all of the SNPs may be retained. At the threshold of RL = 1/200, which results in a more acceptable TPR of 0.797, the interquartile range can be seen to be even wider, at 265 to 1728 of the total SNPs.

The range of FPRs decreases for RL filtering as the OR increases. A per-allele OR of 1.24 is similar to the estimated effect sizes of the causal SNPs in the studies which have used this type of filtering before [17] [53] [52]. The sample size of 20,000 in the simulated datasets was also chosen to closely match the sample sizes used in these studies. The results for RL filtering for this scenario are not shown, but the AUC (with mean FPR) is very close to 1 and there is very little variability in FPR, suggesting that in general RL filtering was a suitable method to use in these studies. In particular, the mean FPR and TPR at a threshold of 1/100 are 0.031 and 0.986 respectively. The median and IQ range of the FPR are 0.015 and (0.009, 0.037).

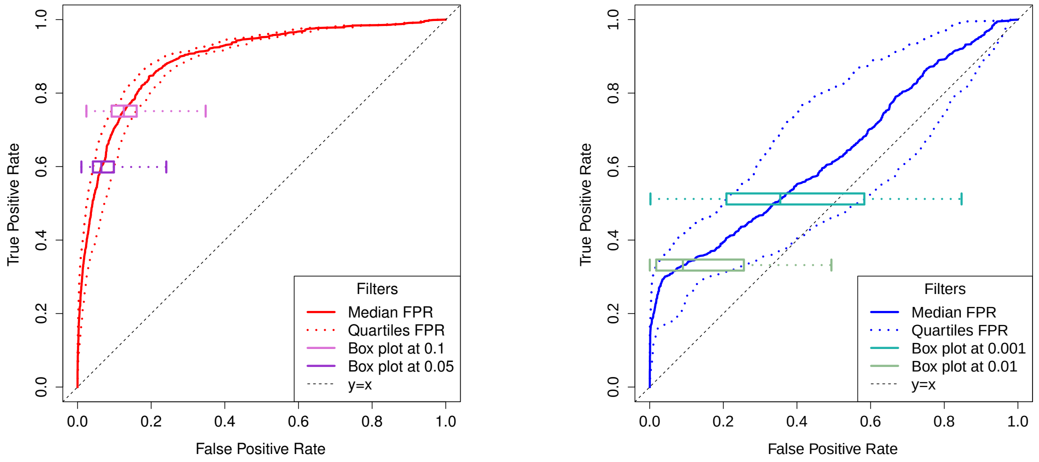
The variability between simulations is a clear limitation of RL filtering and we recommend filtering based on likelihood but using a percentile threshold chosen based on simulation analysis. To further understand the relative performance of RL and LP filters, we considered the log-likelihood surface as a function of the number of controls with the risk genotype and the number of cases with the risk genotype. This is detailed in §2.4.

2.2.2 Variability in other filtering methods

Likelihood percentile filtering specifies the proportion of SNPs to be retained, which is approximately equal to the FPR, so there is virtually no variability in FPR for this method. However, any other filtering method for which the retention threshold is not based on specific numbers or proportions of SNPs will result in some variability. Here we examine that variability for some of the alternative filtering methods.

Figure 2.4(a) displays the results of p_{CA} filtering in a similar way to how they were displayed for RL filtering in Figure 2.3(b). The results also come from filtering on the same datasets. When using p-values, a threshold is often specified based on a Bonferroni correction [36] [7]. In this case, that would result in a p-value threshold of $0.05/2871 = 1.74 \times 10^{-5}$. However, a Bonferroni correction results in a very conservative threshold and if this was used for this simulated data it would result in a TPR of 0.001 and a median FPR of 0. Therefore, we suggest using a higher threshold. The box plots on this figure illustrate the results at p-value thresholds of 0.05 and 0.1.

We have already seen that p_{CA} filtering gives very similar results to LP filtering, and it (or p_W filtering) was suggested as a reasonable alternative if likelihoods were not so readily available. We can now see that there is, in fact, relatively little variability in FPR for the p-value filtering methods (once again, p_W gives similar results). For example, for this scenario, a filter threshold of $p_{CA} = 0.1$ results in a TPR of 0.751 and a median FPR of 0.1257. The IQ range for FPR at this threshold is (0.0926, 0.1607) and the full range is (0.0240, 0.3480). These reasonably narrow ranges (especially the IQ ranges, which are shown on the figure for all filtering thresholds) reinforce the suitability of p-values as alternative filtering statistics.



(a) Results of p_{CA} filtering with the thresholds of $p = 0.05$ and $p = 0.1$ highlighted.

(b) Results of r^2 filtering with the thresholds of $r^2 = 0.01$ and $r^2 = 0.001$ highlighted.

Figure 2.4: The results of two methods of filtering plotted using the median and quartiles of the FPR with specific thresholds highlighted using boxplots. The same simulated data was analysed using both methods. The causal SNP has an OR of 1.1, a MAF of 0.08 and the sample size is 20,000. 1000 datasets were simulated using the LD structure of the *CASP8* region.

Similarly derived results are shown for r^2 filtering on the same datasets in Figure 2.4(b). This was one of the most efficacious LD-based methods. However, when plotted using mean FPR, the AUCs showed that it was not really a suitable alternative to LP filtering (for this scenario the AUC was 64% compared to 90% for LP filtering). Figure 2.4(b) illustrates further shortcomings of the use of this filtering method. As can be seen from the plotted FPR quantiles, the range of FPR values at most r^2 filtering thresholds is very large, and this is further illustrated by the box plots for the two r^2 thresholds (0.01 and 0.001) that are given. Giving the boxplots for these thresholds also demonstrates the sort of threshold values that would have to be used for r^2 filtering to work at all. Higher thresholds commonly result in very few SNPs being retained after filtering. For example, an r^2 threshold of 0.4 results in a TPR of 0.1290 and a median FPR of 0.0014 with this data.

2.3 The use of LP filters in different scenarios

In §2.1, we described several methods of filtering and compared their efficacy using a single causal SNP scenario in three different regions of the genome. Now we consider using SNP filtering in other causal SNP scenarios. We carried out analyses on simulated data covering various scenarios for all of the filtering methods. It was clear that the relative efficacies of the methods generally varied very little. LP filtering is generally the most efficacious of all these methods, so in this section, we focus on this method and we would recommend its usage over the other given statistics.

2.3.1 LP results for different causal SNP MAFs and LD structures

Figure 2.5 shows the how the results of LP filtering vary dependent on the local LD structure and MAF of the causal SNP. The effects of MAF and LD structure are not only important when carrying out genetic association studies in different regions of the genome, but also when studying the same region across multiple populations which may have different LD patterns.

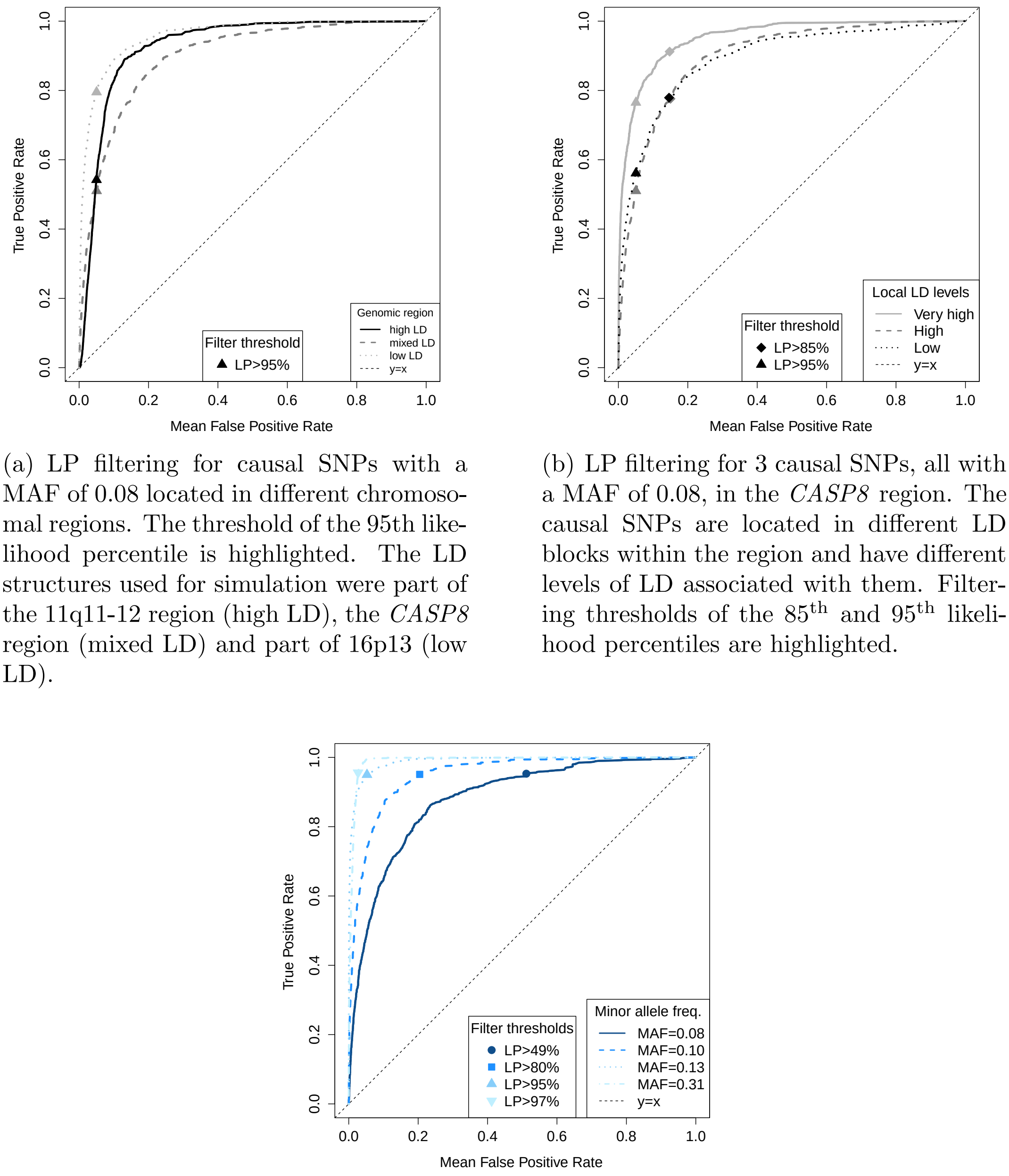
As well as the effects of LD and MAF, we use Figure 2.5 to examine the effect of filtering at different LP thresholds. Figure 2.5(a) compares the LP filtering results for the three different chromosomal regions that were examined in §2.1.2. There are two outcomes of interest, the true and false positive rates (TPR and FPR). With LP filtering, we fix the total proportion of SNPs retained, and as there is only one causal SNP, this proportion is almost identical to the FPR. For example, at a filtering threshold of 95% (as given in this figure), 5% of all SNPs will be retained and the FPR will be approximately 0.05. If there is a fixed proportion of SNPs that can be taken forward (due to experimental costs, for example) a threshold may be chosen based on this. Figure 2.5(b) shows the results from filtering for causal SNPs that are located in the same 1Mb chromosomal region but different LD blocks within that region. We use this figure to demonstrate that if a particular FPR does not yield a high enough TPR, then the filter threshold could be relaxed from the 95th to the 85th percentile, say. It is perhaps more relevant to focus on what threshold is required to achieve a particular TPR. Therefore, the thresholds given in Figure 2.5(c) (filtering for SNPs with different MAFs within the same LD block) are those that result in a TPR of 0.95. We focus on these thresholds as we examine the separate plots in more detail.

Figures 2.5(a) and 2.5(b) show that when filtering for different causal SNPs, even if the MAFs and ORs of these SNPs are the same, there will still be some variability in the results due to the unique LD pattern of each SNP. Every SNP is related to the SNPs around it through the amount of LD between it and those SNPs. We hypothesised that there may be a relationship between the levels of LD that a causal SNP has associated with it and the outcome of filtering when that SNP is causal. We therefore examined filtering for ‘similar’ causal SNPs (with the same MAF and per-allele OR) in chromosomal regions with different overall levels of LD (as examined in 2.1.2 and given in Figure 2.5(a)) and located within different sized LD blocks in the same 1Mb region of the genome (Figure 2.5(b)). These figures do not indicate any clear, simple relationship between local LD and filtering results, but do give an idea of the levels of variation. Overall, the AUCs are fairly similar. If we consider the filtering thresholds that would be required to give a TPR of 0.95, these vary between 61% and 90% for the five causal SNPs examined in these figures (3 within one chromosomal region and 2 in other regions).

For the scenarios for which the analysis results are given in Figure 2.5(c), causal SNPs were carefully selected from a single LD block within the *CASP8* region such that they had a range of MAFs. Datasets were then simulated using each of the causal SNPs with a sample size of 20,000 subjects and an OR of 1.1. We hoped that by comparing the results using SNPs with different MAFs in the same LD block we would reduce as far as possible any confounding effects of LD structure. Figure 2.5(c) clearly demonstrates the general result that LP filtering is more efficacious the larger the MAF of the causal SNP. At smaller MAFs this effect is more profound, with the AUCs for causal SNPs with MAFs 0.08, 0.10 and 0.13 equal to 0.8809, 0.9501 and 0.9890 respectively. Filtering of datasets with causal SNPs with MAFs larger than 0.13 results in ROC curves with only slightly larger AUCs. We have also marked on each ROC curve in this figure a point at the threshold which results in a TPR of 0.95. For the scenarios in which the SNPs have MAF 0.08, 0.1, 0.13 and 0.31, these thresholds are 49%, 80%, 95% and 97% respectively. For any given FPR, the TPR increases as MAF (within a single LD block) increases. This is also the case as causal SNP OR and sample size increase, as we go on to examine in more detail.

2.3.2 LP results for different causal SNP ORs

Figure 2.6(a) shows the results of applying LP filtering as the per-allele odds ratio of the causal SNP varies. The data has been simulated under an additive model, as with most of the examples shown throughout this work. For completeness, we decided to test whether LP filtering would produce similar results if the causal SNP was inherited under a different model. Figures 2.6(b) and 2.6(c) show ROC curves for LP filtering applied to data simulated under recessive and dominant models, respectively (see §1.2.2). Different effect sizes were also examined for these types of inheritance models. In all the simulations, the sample size was 20,000 subjects and the same causal SNP with a MAF of 0.08 was used. In a fine-mapping study it is likely that the mode of inheritance



(a) LP filtering for causal SNPs with a MAF of 0.08 located in different chromosomal regions. The threshold of the 95th likelihood percentile is highlighted. The LD structures used for simulation were part of the 11q11-12 region (high LD), the *CASP8* region (mixed LD) and part of 16p13 (low LD).

(b) LP filtering for 3 causal SNPs, all with a MAF of 0.08, in the *CASP8* region. The causal SNPs are located in different LD blocks within the region and have different levels of LD associated with them. Filtering thresholds of the 85^{th} and 95^{th} likelihood percentiles are highlighted.

(c) LP filtering for causal SNPs with MAFs of 0.08-0.31 located within the same LD block in the *CASP8* region. Each causal SNP has a per-allele OR of 1.1 and for each scenario, an LP threshold which results in a TPR of approximately 0.95 is highlighted.

Figure 2.5: Receiver Operating Characteristic (ROC) curves showing the effectiveness of likelihood percentile (LP) as a fine-mapping filter dependent on the chromosomal region, the local LD structure within a single region and MAF of the causal SNP for which filtering is carried out. 1000 datasets were simulated using a sample size of 20,000 for each scenario and the results of filtering at specific thresholds are highlighted. All causal SNPs have per-allele ORs of 1.1.

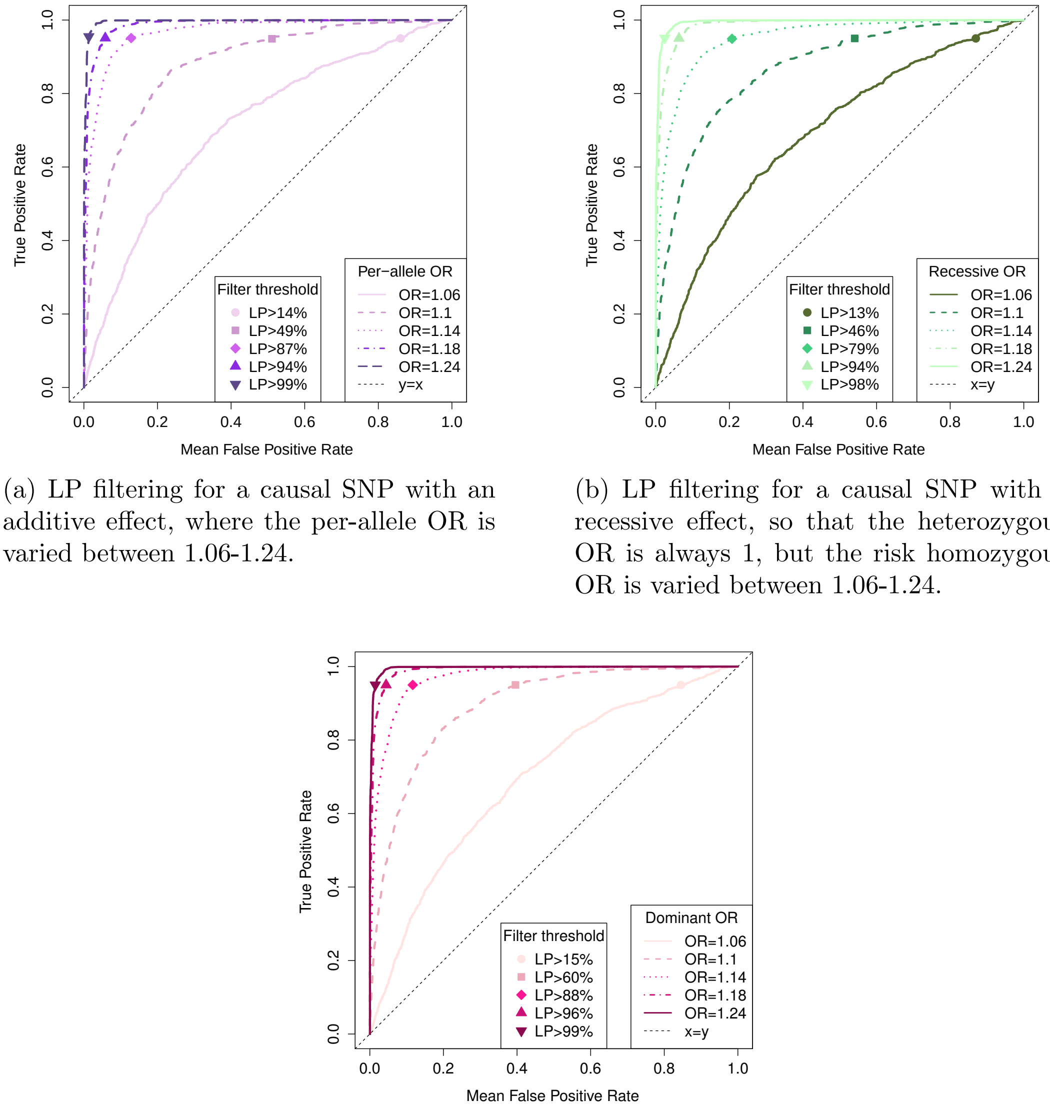
of the casual SNP will be unknown. Therefore, in all cases, we test the efficacy of filtering using likelihoods from fitted logistic regression models with additive SNP effects, as described in §1.3.1.

For the data simulated with an additive effect (Figure 2.6(a)), the per-allele ORs took values between 1.06 and 1.24. This results in heterozygous ORs of 1.06-1.24 and risk homozygous ORs of 1.12-1.54. Although we can see that when the causal SNP has a small OR, for example 1.06, most of the SNPs in the region would need to be retained to achieve a high TPR, it is also clear that efficacy improves with effect size. To illustrate this point, we consider the filtering thresholds required to achieve a TPR of 0.9. When the causal SNP has per-allele OR 1.06, this threshold is 27%, but when it is 1.14, a threshold of 93% can be used to achieve this same TPR. Similarly, if we consider a TPR of 0.95 for these two ORs, thresholds of 14% and 87%, respectively, are required, equivalent to retaining approximately 2469 or 373 SNPs of the 2871 in this dataset.

We used similar effect sizes for the data simulated under different inheritance models. For recessive effects, the heterozygous OR was fixed at 1, but the risk homozygous OR was varied between 1.06 and 1.24. For dominant effects, the risk homozygous and heterozygous ORs were always equal, and these were also varied between 1.06 and 1.24. The different models are quite distinct, so using the same ORs for different types of effect is not really equivalent. However, we can see from Figure 2.6 that using different models with the same range of ORs results in similar ROC curves. The filter thresholds needed to retain the causal SNP with a TPR of 0.95 are given on each figure and can be seen to be close, ranging from 13% to 15% for the smallest effect sizes and from 98% to 99% for the largest effect sizes. The AUCs are also very similar. For the data simulated under additive models, the AUCs range from 71.1% to 99.6%, while for dominant models the range is 69.6% to 99.6% and for recessive models it is 69.1% to 99.4%.

2.3.3 LP results for different sample sizes

We use Figure 2.7 to demonstrate the effect of sample size on LP filter efficacy. We simulated 1000 datasets with a causal SNP with an OR of 1.1 and a MAF of 0.08, each with a sample size of 50,000. We then analysed these full datasets, as well as subsets of them, using different numbers of the samples (always with equal numbers of cases and controls). As in previous figures, we have marked the points on the ROC curve at which the TPR is 0.95. The filtering threshold required to achieve this TPR is 15% with a sample size of 10,000 (retaining 85%



(a) LP filtering for a causal SNP with an additive effect, where the per-allele OR is varied between 1.06-1.24.

(b) LP filtering for a causal SNP with a recessive effect, so that the heterozygous OR is always 1, but the risk homozygous OR is varied between 1.06-1.24.

(c) LP filtering for a causal SNP with a dominant effect, so that the heterozygous OR and the risk homozygous OR are equal, and this value is varied between 1.06-1.24.

Figure 2.6: Receiver Operating Characteristic (ROC) curves showing the effectiveness of likelihood percentile (LP) as a fine-mapping filter dependent on the effect size of the causal SNP. Analyses were carried out on data simulated under 3 different types of inheritance model. 1000 datasets were simulated for each scenario using the LD structure of the *CASP8* region and the results of filtering at the thresholds which result in a TPR of 0.95 are highlighted. The same causal SNP with MAF of 0.08 and a sample size of 20,000 was used for all simulations.

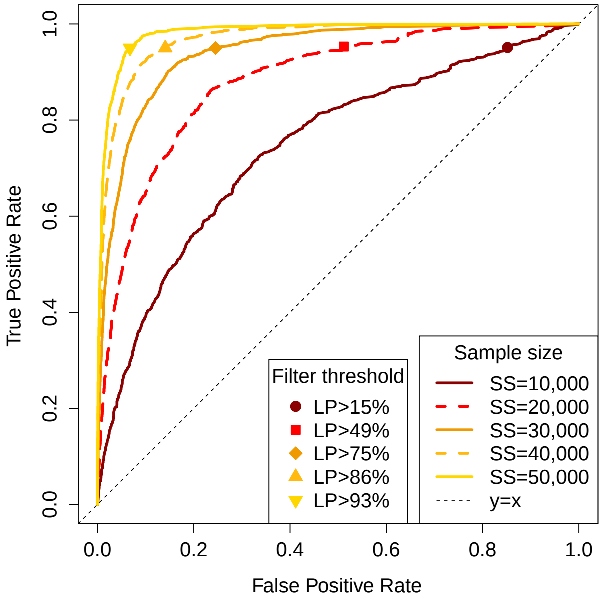


Figure 2.7: Receiver Operating Characteristic (ROC) curves showing the effectiveness of likelihood percentile (LP) as a fine-mapping filter dependent on the sample size used. Sample sizes of 10,000 to 50,000 were used and the thresholds required to retain the causal SNP with 95% probability are highlighted. 1000 datasets were simulated using the LD structure of the *CASP8* region and a causal SNP with a per-allele OR of 1.1 and a MAF of 0.08.

of the total SNPs), but this increases to 49% at a sample size of 20,000 and thresholds of 75%, 86% and 93% at sample sizes of 30,000, 40,000 and 50,000, respectively. With this particular scenario, to be 95% sure of capturing the causal SNP whilst retaining less than 5% of all SNPs, sample sizes larger than 50,000 would be required.

2.4 Using likelihood surfaces to understand results

The outcome of both LP and RL filtering is determined by maximised likelihoods from the individual SNP models. By examining the likelihood we are able to consider the effect that other quantities have on it. Although most of the simulations described in this work are based on an additive model, we concentrate now on a causal SNP with a dominant effect. This simplifies calculations and yields explicit expressions for the maximum likelihood estimates (MLEs), but demonstrates the general effects of different relevant quantities on the likelihood. In §2.3.2, we showed that for the range of ORs we are interested in, LP filtering for a dominant causal SNP with a particular OR (both heterozygous and risk homozygous) will give similar results to LP filtering for an additive causal SNP with the same per-allele OR.

Consider a causal SNP for which the disease risk follows a dominant model. For this scenario, take a sample of n_1 cases and n_0 controls in which it is assumed that the true causal SNP has an OR of λ, a MAF of f. Table 2.3 shows some further notation used throughout this section. The sampling distribution of D_0 and D_1 is derived in terms of λ and f and the likelihood for any realisation of D_0 and D_1 is also derived.

	Risk genotype AA/Aa ($x_j = 1$)	Wildtype genotype aa ($x_j = 0$)	
cases ($y_j = 1$)	D_1	E_1	n_1
controls ($y_j = 0$)	D_0	E_0	n_0
	D	E	n

Table 2.3: The number of cases and controls with the risk and wildtype genotypes for the causal SNP, where a is the wildtype allele and A is the risk allele.

2.4.1 Deriving the likelihood

The association of the causal SNP with the phenotype can be modelled using a logistic regression model with linear predictor $\beta_0 + \beta_1 x_j$. The likelihood of β_0 and β_1 given the data is

$$\mathcal{L}(\beta_0, \beta_1 | \boldsymbol{y}) = \prod_{j=1}^{n} p(x_j)^{y_j}(1 - p(x_j))^{(1-y_j)}. \tag{2.2}$$

The likelihood that is of primarily of interest here is the likelihood based on sampling randomly and not determined by disease status (the prospective likelihood). As the type of data considered here is case-control study data, the disease status is known prior to sampling, so the only likelihood that can be derived is the retrospective likelihood. However, this is not problematic because, as is shown by Prentice and Pike [40], the prospective and retrospective likelihoods are in fact the same.

When using the logistic regression analysis described above to model the probability of disease (being a case) given the genotype for the jth subject, this likelihood ($\mathcal{L}$) can be written as

$$\mathcal{L}(\beta_0,\beta_1|\boldsymbol{y}) = \left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^{E_1}\left(\frac{1}{1+e^{\beta_0}}\right)^{E_0}\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)^{D_1}\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)^{D_0}. \tag{2.3}$$

Taking the natural logarithm of this gives the log-likelihood (ℓ), which after some simplification can be shown to be

$$\begin{aligned}\ell(\beta_0,\beta_1|\boldsymbol{y}) &= E_1ln(e^{\beta_0}) - E_1ln(1+e^{\beta_0}) - E_0ln(1+e^{\beta_0}) + D_1ln(e^{\beta_0+\beta_1})\\ &\quad - D_1ln(1+e^{\beta_0+\beta_1}) - D_0ln(1+e^{\beta_0+\beta_1}) \qquad (2.4)\\ &= E_1\beta_0 - Eln(1+e^{\beta_0}) + D_1(\beta_0+\beta_1) - Dln(1+e^{\beta_0+\beta_1}). \qquad (2.5)\end{aligned}$$

This can then be used to find the maximum likelihood estimators (MLEs) of β_0 and β_1. For additive models this log-likelihood would be maximised using an iterative numerical algorithm, but for a dominant model it is possible to derive the MLEs analytically using differentiation. This gives $\widehat{\beta_0} = ln(E_1/E_0)$ and $\widehat{\beta_1} = ln(D_1E_0/D_0E_1)$. These are intuitive results, as $\widehat{\beta_0}$ is the log-odds of disease when there are no risk alleles present ($x_j = 0$) and $\widehat{\beta_1}$ is the logOR of disease for a risk genotype compared to a wildtype genotype ($x_j = 1$ compared to $x_j = 0$). Substituting these MLEs back into Equation (2.3) and simplifying further gives

$$\mathcal{L}(\widehat{\beta_0},\widehat{\beta_1}|\boldsymbol{y}) = \left(\frac{E_1}{E}\right)^{E_1}\left(\frac{E_0}{E}\right)^{E_0}\left(\frac{D_1}{D}\right)^{D_1}\left(\frac{D_0}{D}\right)^{D_0}. \tag{2.6}$$

This can simply be written in term of the total numbers of cases and controls, n_1 and n_0, and the numbers with the risk genotypes, D_1 and D_0:

$$\mathcal{L}(\widehat{\beta_0},\widehat{\beta_1}|\boldsymbol{y}) = \left(\frac{n_1-D_1}{n-D}\right)^{n_1-D_1}\left(\frac{n_0-D_0}{n-D}\right)^{n_0-D_0}\left(\frac{D_1}{D}\right)^{D_1}\left(\frac{D_0}{D}\right)^{D_0}. \tag{2.7}$$

Figure 2.8 shows a contour plot of the maximised log-likelihood as a function of D_0 and D_1.

2.4.2 Deriving the sampling distributions of D_0 and D_1

The allele frequencies in the general population are f for allele A and $1-f$ for allele a. Assuming Hardy-Weinberg equilibrium, the risk and wildtype genotype

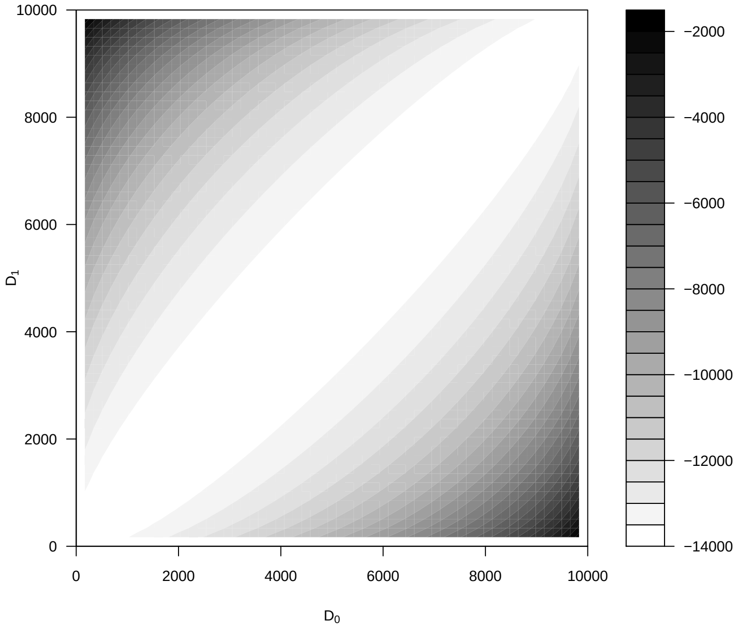


Figure 2.8: A contour plot showing maximised log-likelihood based on the number of controls (D_0) and cases (D_1) with the risk genotype when the sample size is $n_0 = n_1 = 10,000$.

frequencies can be written as

$$p(x_j = 1) = f^2 + 2f(1-f) = 2f - f^2, \tag{2.8}$$

$$p(x_j = 0) = (1-f)^2. \tag{2.9}$$

If the disease is rare so that there is only a small proportion of cases in the general population, then the sample of controls will approximate a completely random sample from the whole population. Therefore, these probabilities will be approximately the same in the controls as they are in the general population: $p_{D_0} = p(x_j = 1|d_j = 0) \approx p(x_j = 1) = 2f - f^2$. The probability distribution for D_0, the number of controls with a risk genotype is binomial with parameters n_0 and p_{D_0}:

$$P(D_0 = k) = \left(\begin{array}{c} n_0 \\ k \end{array} \right) (2f - f^2)^k (1-f)^{2(n_0-k)}. \tag{2.10}$$

The odds ratio (λ) of having a risk genotype ($x_j = 1$) in cases compared to controls can be written as

$$OR = \lambda = \frac{p_{D_1}/(1-p_{D_1})}{p_{D_0}/(1-p_{D_0})}. \tag{2.11}$$

where $p_{D_1} = P(x_j = 1|case)$. With $p_{D_0} = 2f - f^2$, this can be rearranged to give p_{D_1} in terms of λ and f:

$$p_{D_1} = \frac{\lambda(2f-f^2)}{\lambda(2f-f^2)+(1-f)^2}. \tag{2.12}$$

The probability distribution for D_1, the number of cases with a risk genotype is binomial with parameters n_1 and p_{D_1}, so

$$P(D_1 = k) = \left(\begin{array}{c} n_1 \\ k \end{array} \right) \left(\frac{\lambda(2f-f^2)}{\lambda(2f-f^2)+(1-f)^2} \right)^k \left(\frac{(1-f)^2}{\lambda(2f-f^2)+(1-f)^2} \right)^{n_1-k}. \tag{2.13}$$

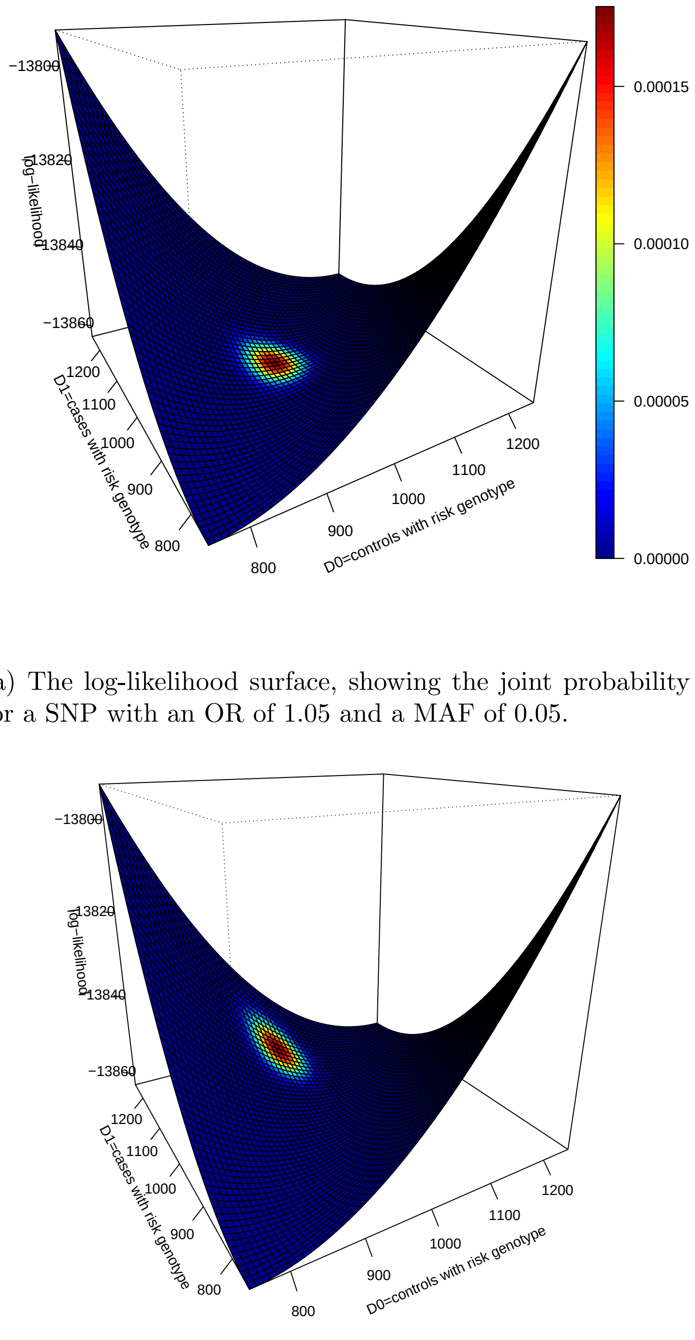
2.4.3 An illustration

As in Figure 2.8, the diagrams in Figure 2.9 show part of the log-likelihood surface as a function of D_0 and D_1, this time as a 3-dimensional surface. Figure 2.9(a) illustrates the scenario of a sample size of 20,000 and a causal SNP with a MAF of 0.05 and an OR of 1.05. The joint probability distribution of D_0 and D_1 is shown as a coloured circular region, with red for the most likely and blue for the least likely values. Figure 2.9(b) illustrates a similar situation but where the SNP has an OR of 1.2. This surface can help to explain how changing the OR of the causal SNP affects the (log-) likelihood and hence the filtering results.

$D_1 - D_0$ is a measure of the distance from the $D_0 = D_1$ line and since D_0 and D_1 are independent and have known probability distributions, it is straightforward to derive an approximation for the expectation of $D_1 - D_0$ and to show that this expectation increases with the OR:

$$\begin{aligned} E(D_1 - D_0) &= E(D_1) - E(D_0) && (2.14) \\ &= \frac{\lambda(2f-f^2)n_1}{\lambda(2f-f^2)+(1-f)^2} - (2f-f^2)n_0. && (2.15) \end{aligned}$$

The fraction part of this is the only part that changes with the OR (λ). As λ increases, the numerator increases, but only part of the denominator increases. Therefore, this fraction, and the whole of the expectation increases with in-



(a) The log-likelihood surface, showing the joint probability for a SNP with an OR of 1.05 and a MAF of 0.05.

(b) The log-likelihood surface, showing the joint probability for a SNP with an OR of 1.2 and a MAF of 0.05.

Figure 2.9: Part of the log-likelihood surface for SNPs, as a function of the number of cases and controls with the risk genotype for a sample with 10,000 cases and 10,000 controls. In both figures, the joint probability of controls with the risk genotype and cases with the risk genotype is shown as a coloured circular region with red representing highly probable values and blue unprobable values.

creasing λ. Figures 2.8 and 2.9 show that the likelihood is approximately a function of $D_1 - D_0$. So increasing the OR yields larger likelihoods on average. If the causal SNP has an OR of 1.05 then the joint probability distributions of D_0 and D_1 for SNPs in strong LD with it will have a large overlap with the joint probability distributions for SNPs with ORs of 1. A likelihood-based filter will capture many of the SNPs including those with ORs of or close to 1 in this case. If the causal SNP has an OR of 2, the joint probability distributions for SNPs in high LD with it will not overlap those for SNPs with ORs of 1 very much. In this case the filter will mainly capture SNPs in high LD with the causal SNP and generally less SNPs in total. When there is some overlap in these joint probabilities, the variability in FPR will also be higher, as high FPRs will result if the causal SNP happens to lie in the overlapping part of the probabilities, with likelihoods close to many other SNPs. There is some probability, though, that it will have a higher likelihood, producing low FPRs in these cases.

The shape of the 3-dimensional likelihood surface also helps us to understand the reason why LP filtering can be advantageous over RL filtering. Figure 2.9(a) demonstrates that the majority of the probability for a causal SNP with an OR close 1 lies close to the lowest possible likelihoods (in this realisation, log-likelihoods of around -13,860 to -13,855). When carrying out RL filtering at an RL threshold of 1/100, all SNPs with RL $\leq 1/100$ compared to SNP$_{max}$ are retained. This is the SNP with the highest log-likelihood, and this RL is equivalent to a difference in log-likelihoods of approximately 4.61. Due to repeated tests (thousands of SNPs) it often occurs that a very small number will result in an unusually high likelihood, one of these being SNP$_{max}$. Therefore, very few SNPs will have RL $\leq 1/100$, say, and be retained after RL filtering. The causal SNP's log-likelihood could occur anywhere in the range of probable values, so may be retained, but unless it is at the top of this range that is unlikely. However, if LP filtering is used instead, the number of SNPs retained is fixed, so the effect of a single extreme likelihood value will be reduced significantly. In LP filtering, only the rank of the SNPs based on their likelihood is important, rather than the specific values of the likelihoods, as in RL filtering.

2.5 Imputation

Although fine-mapping studies will genotype a large number of SNPs in a region, difficulties such as the high costs mean that not every SNP will be targeted, and the causal SNP may be missed. To reduce the probability of this, and make fine-mapping analysis as thorough as possible, it is common to impute the unknown genotypes of SNPs, as described in §1.2.4. To determine whether the methods we have considered so far would be effective under these circumstances, we have compared the analysis results of some of our simulated datasets to the same analysis results when a subset of the genotypes were removed and then re-imputed into these datasets.

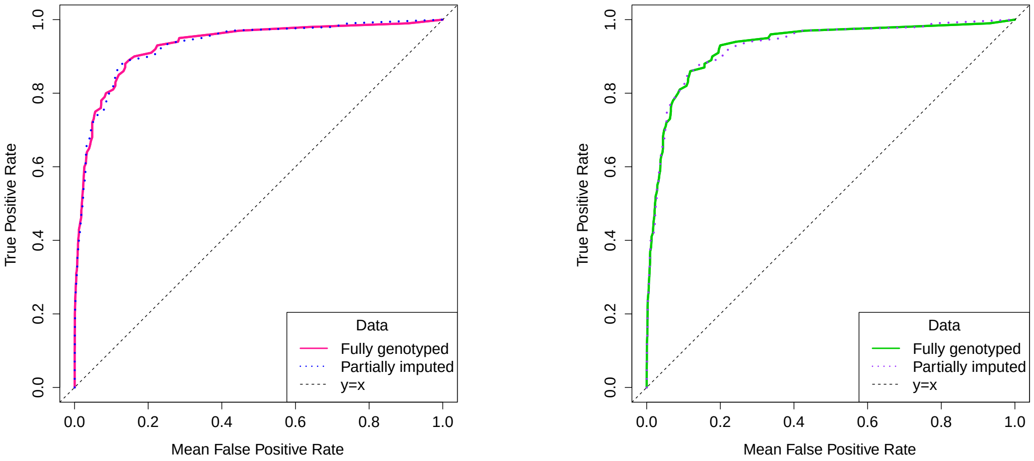
Using the *CASP8* data, for several scenarios, we simulated 100 datasets with the same 2871 SNPs as previously and carried out filtering on these. For the iCOGS fine-mapping study of this region, which is described in §1.3.3, a panel of tagSNPs were chosen which included all SNPs with $r^2 \geq 0.1$ with two previous hits in the region and further SNPs to ensure that all known SNPs were tagged at $r^2 \geq 0.9$. The set of SNPs we had simulated included 469 of these tagSNPs, so we reduced the simulated datasets down to the genotypes for these 469 SNPs alone. We then used IMPUTE2 [33] (§1.2.4) to impute the missing genotype doses based on the MAFs and LD of the region and re-analysed these new datasets. We had purposefully chosen a causal SNP for each scenario that was not one of the tagSNPs, so would always be imputed.

2.5.1 Robustness of filters when imputation is used

The best performing filters so far have been the LP and p_W filters, and we wished to see if these would perform as well on datasets where the majority of SNPs had been imputed. The results of filtering on fully genotyped data and partially imputed data using both of these filters are compared in Figure 2.10. These results are for a causal SNP scenario with an OR of 1.1, a MAF of 0.13 and a sample size of 10,000 and are given in the form of ROC curves. The ROC curves demonstrate the similarity of the outcomes for this particular scenario, whether all SNPs are genotyped or only an informative subset are genotyped and the causal SNP is among the majority of SNPs in having their genotype doses imputed. In fact they are so similar that for both methods, the AUCs agree to the nearest 1%. This similarity was also observed when testing the methods on other scenarios, leading us to conclude that these methods should be suitable to analyse fine-mapped data when a suitable panel of tagSNPs has been used to impute the genotype doses of further SNPs.

2.6 Summary of filtering using only genotype data

This chapter has considered the efficacy of a number of currently-used and novel methods as fine-mapping filters to significantly reduce the number of



(a) Filtering using likelihood percentile points (LP). The AUC using the fully genotyped data is 93.08% and the AUC using the partially imputed data is 93.02%.

(b) Filtering using p-values from Wald tests (p_W). The AUC using the fully genotyped data is 92.87% and the AUC using the partially imputed data is 92.80%.

Figure 2.10: The effectiveness of LP and p_W filtering for fine-mapping data which has been partially imputed compared to its effectiveness for data which is fully genotyped. The causal SNP has an OR of 1.1, a MAF of 0.13 and a sample size of 10,000. A set of 100 datasets were simulated using the LD structure of the *CASP8* region containing 2871 fully genotyped SNPs. These were then reduced to contain 469 genotyped informative SNPs and the remaining 2402 SNPs were imputed.

SNPs considered as causal candidates. The statistics we suggest for filters are all easily computed using univariate logistic regression models and require only the genotype data. We compared the true and false positive rates of these filters when used on a number of simulated datasets. We tried to replicate data similar to what would be expected with currently undiscovered causal SNPs, giving our simulated causal SNPs ORs of less than 1.3.

The motivation for this work was to examine how efficacious relative likelihood (RL) filtering is as a method for choosing the SNPs to take forward from a fine-mapping study. RL filtering with a threshold of 1/100 has been used for fine-mapping the *FGFR2* and the *16q12* loci [53], [52]. We have shown that with a moderate effect size, such as at these loci, and a similar sample size to these studies of 20,000, that this method works well. With smaller ORs, however, the RL filtering is less effective, and with a similar sample size it would be difficult to detect causal SNPs with per-allele ORs of less than 1.2 (see §2.1.2 and §2.2.1). A limitation of this method is the high levels of uncertainty in the number of SNPs that will be retained after filtering, demonstrated by the variability in FPRs across analyses of multiple realisations of the same causal SNP scenario. The variability in FPR increases as sample size and causal SNP OR and MAF decrease. An alternative likelihood filtering method using likelihood percentiles avoids this variability. In this case the number of SNPs retained will be set in advance, which is particularly useful if the experimental design of future stages of investigation limits the number of SNPs which can be taken forward.

In §2.1 and §2.3 we were able to demonstrate that whilst both likelihood and p-value based filters can be efficacious in the analysis of fine-mapped data, likelihood percentile (LP) is generally the most effective of these methods, and we therefore recommend its use over other methods. We showed this using simulated data for scenarios covering the very different LD structures of three regions of the genome, but observed similar results in each of these regions. We hope that our results will therefore generalise to a variety of different genomic regions for which fine-mapping is required. The investigation into the effect of short range LD on the results within the *CASP8* region examined causal SNPs with the same MAF and OR, but located in different LD blocks (with approximate sizes of 10kb, 72kb and 225kb). As with the results from different regions, these filtering results did not change in a uniform way but there was some variation between the results for the causal SNPs in the different LD blocks. In addition, we were able to demonstrate that filtering may be used on data for which only an informative subset of SNPs in a region are genotyped, with the genotype doses of the rest being imputed using IMPUTE2 [33]. We observed similar true and false positive rates when fully 'genotyped' data were simulated and analysed to when partially imputed data were analysed.

The filters based on the structural relationships between variants that were also investigated in §2.1 produced less encouraging results. For the scenarios we considered, the relatively complex PLD score developed by Zhu *et al.* [62] appeared to be only slightly more efficacious as a filter than simply filtering using r^2 with or genetic map distance from the SNP with the highest maximised likelihood. None of these methods produced ROC curves with AUCs as high as those for the LP or p-value filters. It may be that these or similar methods, such as taking into account LD with the top few hits, rather than just the single top hit, would be useful under certain circumstances.

Chapter 3

Bayes Factor based methods of filtering

3.1 Bayes Factors and the Wakefield approximation

3.1.1 Using a Bayesian method for SNP filtering

So far, we have examined a number of methods for filtering SNPs in fine-mapping analysis. These methods all assume a single causal SNP model and are based on genotype data alone. We were able to show that some methods, in particular the likelihood percentile (LP) method were able to appreciably reduce the set of candidate causal SNPs whilst retaining the causal SNP with a high probability. However, we expect such causal SNPs to have small effect sizes and these effects to be confounded by extremely high levels of short range linkage disequilibrium (LD). The simulated data we tested the methods on reflected these issues and we found that even in very large samples of tens of thousands of subjects, as are currently being employed by international consortia, it is not possible to identify a single SNP alone as having the causal effect.

Functional information is now widely available for much of the genome, largely from the ENCODE project [15], including some at SNP level. Incorporating information such as this in a filtering statistic should help to distinguish between variants that are likely and those that are unlikely to be causal. This could result in an increase in the probability of retaining of the causal SNP at a given false positive rate. The filtering statistics we considered previously were all based on frequentist statistical methods. Frequentist methods focus on the fact that statistical tests are performed on a sample of the whole population, and that there are many possible permutations of the population. Parameters are assumed to be fixed, whereas the data itself is variable. Therefore, a frequentist test is based on the frequency with which the observed data could have occurred under the null hypothesis. The other major approach to statistical analysis, which is sometimes used with genetic data, is Bayesian analysis. Using this methodology, the focus shifts to the fact that the only thing that is completely known is the collected data, so it is this that is presumed fixed and parameters are unknown. The likely values of the parameters are described probabilistically. Priors, which are often subjective, are specified, and inference is based on posterior distributions.

Bayesian analysis readily lends itself to the combination of multiple sources of data. This may include expert opinion, in that an investigator will have a prior degree of belief that a particular feature describes the true situation, and this belief could be formulated into a prior distribution. After observing new data, we would expect their degree of belief to change to take this evidence into account, and we can combine priors with observed data using formal methodology derived from Bayes' theorem. Information that could be used in this way may come from experts such as geneticists, genetic epidemiologists and biochemists. By using Bayesian methods in genetic fine-mapping, it is also possible to take into account known features of the SNPs, such as whether they occur in sequences that are well conserved across different species. Another important feature of Bayesian analysis is that it also allows for uncertainty to be modelled.

The Bayes Factor (BF) [28] is a Bayesian statistic which is already being used in genetic analysis [59] [60] [48]. The calculation of BFs is now implemented in genetic analysis software such as SNPTEST2 [34], and their use is becoming increasingly more popular as a filter in fine-mapping studies [32]. A BF is the value which can be multiplied with a prior odds to calculate a posterior odds, according to the formula $\Delta/(1-\Delta) = \delta/(1-\delta) \times \mathrm{BF}$, where Δ and δ are the posterior and prior probabilities, respectively. Derived from Bayes Theorem [28], the Bayes Factor is the ratio of the probabilities of observed data occurring under two differing hypotheses:

$$\mathrm{BF} = \frac{P(data|H_1)}{P(data|H_0)}. \tag{3.1}$$

We continue to focus on methods that analyse the association between each SNP and the disease separately, without taking any interactions into consideration. To calculate a BF for each SNP, consider the hypotheses that the SNP has no association, or alternatively some association, with the disease. This methodology can be applied to fine-mapped genotypes by once again using information from the same fitted single-SNP logistic regression models that we used previously,

$$y_{ij} = \frac{e^{\beta_{0i}+\beta_{1i}x_{ij}}}{1+e^{\beta_{0i}+\beta_{1i}x_{ij}}}. \tag{3.2}$$

For SNP i, $H_0 : \beta_{1i} = 0$ and $H_1 : \beta_{1i} \neq 0$, such that the BF formally compares the evidence for association to the evidence for no association. BF_i may then be combined with a prior probability of association specific to SNP i, δ_i, to determine a posterior probability of association, Δ_i [48].

3.1.2 Wakefield's approximate Bayes Factor

The BF, as given in Equation (3.1), includes marginal likelihoods which lead to intractable integrals in most cases. It is common to use a Laplace approximation [28], which has been shown to work well and this method is integrated into several pieces of software, including SNPTEST2 [34]. We have instead chosen to use an approximation derived by Wakefield [59] [60]. This is much easier to compute and we have shown it agrees closely with Laplace approximations from SNPTEST2 for sample sizes $\geq$ 10,000 and a variety of MAFs and effect sizes (data not included). In their review of *Bayesian statistical methods for genetic association studies*, Stephens and Balding [48] also comment that the Wakefield approximation is an appropriate and convenient alternative to the Laplace approximation. Both methods are based on asymptotics and all the datasets we use have very large sample sizes, so the approximations are expected to be good.

The Wakefield approximate Bayes Factor (WBF) is derived by writing Equation (3.1) as

$$\text{WBF} = \frac{\int p(\widehat{\beta_1}|\beta_1)\pi(\beta_1)d\beta_1}{p(\widehat{\beta_1}|\beta_1 = 0)}, \tag{3.3}$$

where $\pi(\beta_1)$ is the prior over β_1, the logOR of the SNP, and $\widehat{\beta_1}$ is the maximum likelihood estimate (MLE) of β_1. This can be simplified by considering a prior of the form $\beta_1 \sim N(0, W)$, and the fact that, asymptotically, $\widehat{\beta_1} \sim N(\beta_1, V)$. We use V estimated from the data. Using these distributions, Wakefield showed that

$$\text{WBF} = \sqrt{\frac{V}{V+W}} exp\left(\frac{\widehat{\beta_1}^2 W}{2V(V+W)}\right). \tag{3.4}$$

The full derivation of the WBF approximation is given in Appendix A. Note that in his papers [59] [60], Wakefield considers the inverse of this BF. Here the evidence in favour of the association model over the null model is of interest, so we use the form in Equation (3.4), such that values >1 signify more evidence in favour of a model with an association term. In general, the greater the value, the more evidence there is of an association.

3.2 SNP filtering using Wakefield Bayes Factors

We consider whether the posterior probability of association, Δ, (or equivalently the odds) is an efficacious filtering statistic. To calculate this, a prior probability, δ, must be specified for each SNP and this can be problematic. However, in many cases, investigators may have little or no prior information about the region and therefore it can be appropriate to assign equal δ to each SNP. In this case the rankings of SNPs using Δ will be the same as those using BF. For now we assume this scenario and therefore do not need to specify δ values, but simply filter using WBF approximations.

3.2.1 Wakefield Bayes factor sensitivity analysis

The WBF has a $N(0, W)$ prior on β_1. As we are assuming that there is no prior information available with which to differentiate between individual SNPs, the same value of W can be assigned to all SNPs. A suitable value of W must be chosen. We decided to carry out analysis to determine how sensitive the results are to this choice. W is a variance and therefore must be positive. We believe that fine-mapping is only likely to be necessary for causal SNPs with ORs < 2.25, so priors with 99.5^{th} percentiles below 2.25 should be appropriate. This corresponds to $0 < W \leq 0.1$.

Figure 3.1 contains threshold averaged receiver operating characteristic (ROC) curves [18] showing the results of filtering using WBF. Each curve illustrates the results for filtering using a different value of W (the prior variance of the logOR) on the same 1000 simulated datasets. The methods used for simulation and plotting ROC curves are the same as previously. We have also included in this figure the results for likelihood percentile (LP) filtering on the same datasets for comparison. This figure suggests that BFs are a promising tool for filtering, giving comparable ROC curves to LP filtering when certain values of W are used. In particular, when $W = 0.01$, WBF filtering is able to produce higher TPRs at FPRs < 0.16 than LP filtering. When combining BFs with appropriate prior information through δ, if the causal SNP is given a large value of δ, the TPR will increase. If most of the other SNPs are given smaller values, the FPR will decrease, potentially resulting in ROC curves with higher AUCs than those for LP filtering. If WBF is used for genetic analysis, W should be chosen carefully, as we can see from Figure 3.1 that the results are highly sensitive to this choice, even within the viable range we specified. Of the four WBF analyses carried out on these specific datasets we observed AUCs in the range 0.82 (when $W = 0.1$) to 0.87 (when $W = 0.01$), but the value of W which gives the largest AUC differs dependent on the causal SNP scenario. One approach for choosing an appropriate W is to carry out elicitation with an expert on the particular problem that is being investigated.

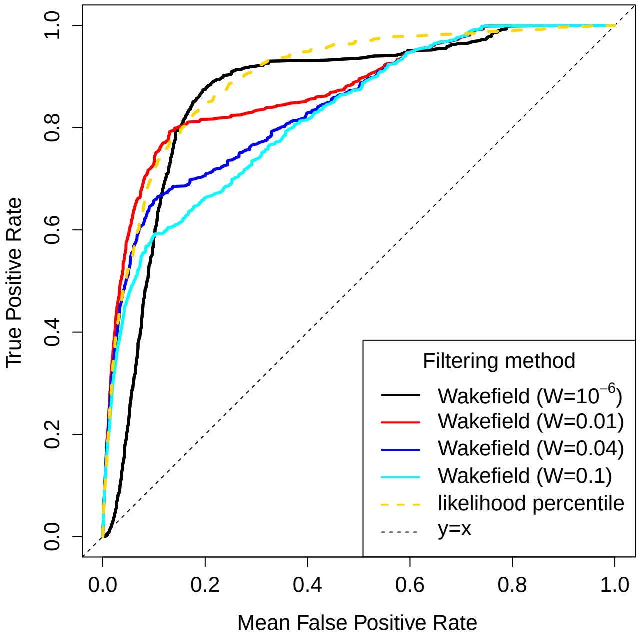


Figure 3.1: Receiver operating characteristic (ROC) curves of BF filtering results, each using the Wakefield approximation and $N(0, W)$ prior for the logOR with a different value of W. The filtering was carried out on 1000 datasets simulated using the LD structure of the *CASP8* region for a scenario with a causal SNP that has an OR of 1.1 and a MAF of 0.08 and a total sample size of 20,000. The results for likelihood percentile filtering on the same datasets are also shown.

3.2.2 A prior for log(OR) dependent on MAF

So far we have assigned the same value of W, the prior variance of the logOR, to all SNPs. However, if we have some prior information about individual SNPs, we may be able to assign different values. There is some evidence to suggest that causal SNPs with lower MAFs are more likely to have larger effects [61], so in his paper [60], Wakefield suggests a method of allowing W to vary dependent on MAF. For a SNP with MAF M and prior $\beta_1 \sim N(0, W(M))$, Wakefield gives the formula $W(M) = \alpha_0 \exp(-\alpha_1 \times M)$. To test this in the setting of breast cancer genetics, elicitation was employed with a breast cancer geneticist to determine suitable values for α_0 and α_1, following the guidelines set out by Wakefield. Subject-specific 99% centralised probability intervals (PIs) for ORs were elicited for SNP association with breast cancer at 4 different MAFs. However, there was no possible way to closely fit a $W(M)$ equation of the form given by Wakefield to the data elicited from the expert.

A new formula for W(MAF)

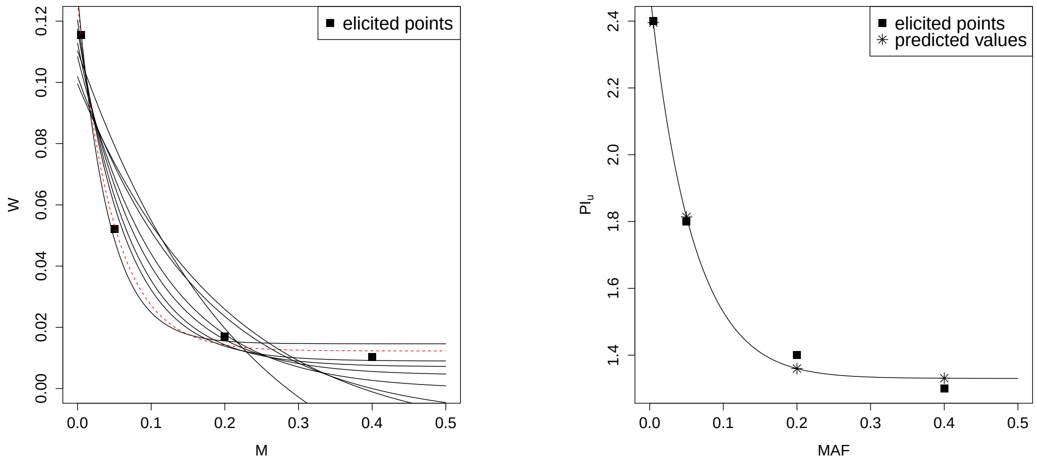
To get a better fit to the elicited percentiles, more parameters were added to give the form:

$$W(M) = \alpha_0 + \alpha_1 \exp(\alpha_2 + \alpha_3 \times M). \tag{3.5}$$

To fit an equation of this form to the data, the R package `nleqslv` was employed. Many combinations of different starting values for the α parameters were used to estimate appropriate values fitting the equation to the elicited points as closely as possible. Figure 3.2(a) shows the results of the output from 9 of these combinations in terms of W plotted against MAF. The "best fitting" equation was chosen as the one with the least sum of the squared distances between predicted W and elicited W at the 4 given MAFs. This is shown as a dashed red line in Figure 3.2(a) and the four elicited values of W are plotted as squares. The line has the equation

$$W(M) = 0.0123 + 0.172 \exp(-0.451 - 20.3M), \tag{3.6}$$

and this relationship is also plotted in Figure 3.2(b) in terms of the upper value of the 99% centralised probability interval (PI$_u$) plotted against MAF. The elicited points are plotted as well as the predicted values of PI$_u$ at the same MAFs according to the fitted relationship, showing how closely they agree.



(a) Nine fitted equations obtained by solving non-linear equations using different starting values and the four elicited points shown on the figure. The "best fitting" equation is shown as a red, dashed line.

(b) The dependence of PI_u, the upper value of the 99% probability interval for the per allele OR of a SNP, on MAF, derived from the "best fitting" $W(M)$ equation. The elicited points are shown, as well as the predicted values of PI_u at the same MAFs.

Figure 3.2: Fitted equations for dependence of W on MAF (M) according to $W(M) = \alpha_0 + \alpha_1 \exp(\alpha_2 + \alpha_3 \times M)$, where W is the prior variance of the logOR for a SNP with the relevant MAF. The "best fitting" equation based on least squares has the form $W(M) = 0.0123 + 0.172 \exp(-0.451 - 20.3M)$.

Filtering using Bayes Factors with W(MAF)

The use of BF filtering with SNP-specific logOR priors of the form $N(0, W(M))$ was tested on simulated datasets using different causal SNP and sample size scenarios. The results from these analyses were compared to the results of using $N(0, W)$ with the same value of W for all SNPs. Figure 3.3 shows that when the expert's beliefs fit the causal SNP scenario reasonably well, this method produces good results. In this case, the causal SNP had a small effect size (OR of 1.06) and large MAF (0.31), and the AUC of the ROC curve when $W(M)$ was employed was 93%. This is larger than the AUCs of three of the ROC curves produced using a fixed W, but slightly smaller than when $W = 10^{-6}$, which has an AUC of 96%. As might be expected, the results are not so positive when the causal SNP scenario differs somewhat from the prior belief. Therefore, this seems like it may be a good method to use if it is not possible to determine a prior W that an expert is confident about, but it is possible to elicit such a relationship between W and MAF. Because of the sensitivity of results to the choice of W, the strengths of an experts beliefs should be considered before a decision is made on the prior to use, whether it is a fixed value or dependent on MAF. We go on to explore the possibility of including such uncertainty of

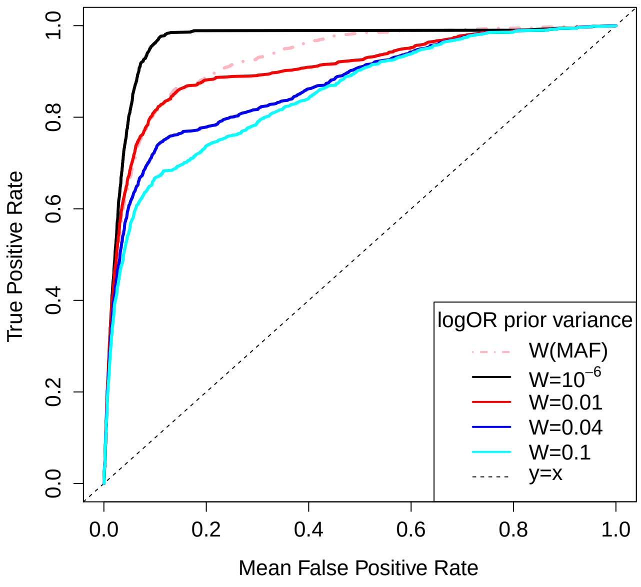


Figure 3.3: Comparison of BF filtering using $W(\text{MAF}) = 0.0123 + 0.172\exp(-0.451 - 20.3\text{MAF})$ to using fixed values of W. 1000 datasets were simulated with a causal SNP with a per-allele OR of 1.06, a MAF of 0.31 and 20,000 subjects. They were simulated using the LD structure of the *CASP8* region. A prior on the logOR of the form $N(0, W)$ and the Wakefield approximation were used.

beliefs in §3.3.

3.2.3 Choosing the prior on the log(OR) based on the data

Although standard Bayesian methods require a prior distribution to be fixed before the data is obtained, an alternative is to estimate a prior from the data, in what is called an empirical Bayes method. We hypothesised that a way to optimise filtering results would be to use the data to attempt to choose a prior which would maximise the BF approximation for the causal SNP, whilst not maximising those of other SNPs. To do this, we must first find the value of W which maximises a WBF approximation.

The numerator of the WBF can be re-written as follows:

$$\begin{aligned}
\int p(\widehat{\beta_1}|\beta_1)\pi(\beta_1)d\beta_1 =& \frac{1}{\sqrt{2\pi(V+W)}}\exp\left(-\frac{\widehat{\beta_1}^2}{2V}+\frac{\widehat{\beta_1}^2W}{2V(V+W)}\right) && (3.7)\\
=& \frac{1}{\sqrt{2\pi(V+W)}}\exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) && (3.8)\\
=& p(\widehat{\beta_1}|\widehat{\beta_1} \sim N(0,V+W)). && (3.9)
\end{aligned}$$

Therefore the WBF itself can be written as

$$\text{WBF} = \frac{p(\widehat{\beta_1}|\widehat{\beta_1} \sim N(0,V+W))}{p(\widehat{\beta_1}|\widehat{\beta_1} \sim N(0,V))}. \quad (3.10)$$

If $\widehat{\beta_1}$ and V are held constant, the denominator is constant, so to maximise the WBF, only the numerator must be maximised. Calculus is employed to find the value which maximises WBF with respect to (w.r.t.) W. Where the numerator is $f(W)$,

$$\begin{aligned}
f(W) =& \frac{1}{\sqrt{2\pi(V+W)}}\exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) && (3.11)\\
=& g(W) \cdot exp(h(W)). && (3.12)
\end{aligned}$$

Using the product and chain rules we find

$$\begin{aligned}
f'(W) =& g'(W) \cdot exp(h(W)) + g(W) \cdot exp(h(W)) \cdot h'(W) && (3.13)\\
=& \frac{-1}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}(V+W)^{\frac{3}{2}}}\exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right)\\
&+ \frac{1}{\sqrt{2\pi(V+W)}}\exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) \cdot \frac{2\widehat{\beta_1}^2}{(2(V+W))^2} && (3.14)\\
=& \left(\frac{\widehat{\beta_1}^2}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}(V+W)^{\frac{5}{2}}} - \frac{1}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}(V+W)^{\frac{3}{2}}}\right)\exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) && (3.15)
\end{aligned}$$

All stationary points of $f(W)$ (and therefore the WBF) will be found at values of W for which $f'(W) = 0$, and we denote such values $\widehat{W}$. In this case, either the exponential or its multiplier must be equal to 0 at any such points. As there are no exponentials which are equal to 0, if there are any stationary points they can be found by solving:

$$
\begin{aligned}
\frac{\widehat{\beta_1}^2}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}(V+\widehat{W})^{\frac{5}{2}}} - \frac{1}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}(V+\widehat{W})^{\frac{3}{2}}} &= 0 && (3.16)\\
\Rightarrow \frac{\widehat{\beta_1}^2}{V+\widehat{W}} &= 1 && (3.17)\\
\Rightarrow \widehat{W} = \widehat{\beta_1}^2 &- V && (3.18)
\end{aligned}
$$

By finding the second order differential of the numerator of the WBF w.r.t. W, we can show that $f(\widehat{W})$ is in fact a maximum rather than another type of stationary point:

$$f''(W) = \frac{-6\widehat{\beta_1}^2 + 3(V+W) + \widehat{\beta_1}^4(V+W)^{-1}}{2^{\frac{5}{2}}\pi^{\frac{1}{2}}(V+W)^{\frac{7}{2}}} exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right). \quad (3.19)$$

Substituting in $\widehat{W}$, we find:

$$
\begin{aligned}
f''(\widehat{W}) =& \frac{-6\widehat{\beta_1}^2 + 3\widehat{\beta_1}^2 + \widehat{\beta_1}^4\widehat{\beta_1}^{-2}}{2^{\frac{5}{2}}\pi^{\frac{1}{2}}|\widehat{\beta_1}|^7} exp\left(-\frac{\widehat{\beta_1}^2}{2\widehat{\beta_1}^2}\right) && (3.20)\\
=& \frac{-1}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}|\widehat{\beta_1}|^5} exp\left(-\frac{1}{2}\right) && (3.21)\\
\approx& \frac{-0.121}{|\widehat{\beta_1}|^5}. && (3.22)
\end{aligned}
$$

The powers which are multiples of $\frac{1}{2}$ are derived from the square root in the normal distribution, which is taken to be a positive square root, therefore, these will also be positive. This is why $|\widehat{\beta_1}|$, rather than $\widehat{\beta_1}$, applies in the denominator. Thus, $f''(\widehat{W})$ will always be negative, proving that the unique stationary point of WBF that occurs at $\widehat{W}$ is a maximum point.

We wish to use WBF with prior logOR $\sim N(0, \widehat{W})$. This is problematic if $V > \widehat{\beta_1}^2$, as $\widehat{W}$ would then be negative. As WBF must be strictly decreasing after its unique maximum at $\widehat{W}$, we can see that in this case, the variance which maximises WBF is the smallest positive value possible. In practice, we use

$$W_{EB} = \max(\widehat{\beta_1}^2 - V, 10^{-12}), \quad (3.23)$$

where the subscript, EB, indicates the empirical Bayes nature of such a value.

Filtering using Bayes Factors with W_{EB}

In practice using W_{EB} is not as simple as it may seem. To maximise W for the causal SNP, one would need to calculate W_{EB} using the $\widehat{\beta_1}$ and V of the causal SNP. The causal SNP is not known, so we need to find good approximations to its SNP-specific $\widehat{\beta_1}$ and V values. V is dependent on several quantities, including MAF and sample size. There is no way to pre-determine the MAF of the causal SNP, but the sample size is equal for all SNPs. When cases and controls are equal, V is inversely proportional to sample size [44] and we suggest using the median V from all SNPs in the dataset, as a representative value for that sample size, and we denote this V_m. As the SNP that produces the model with the largest likelihood (SNP$_{max}$) is likely to be in high LD with the causal SNP, we hypothesised that this SNP may have a $\widehat{\beta_1}$ value close to that of the causal SNP. We chose to use $\widehat{\beta_1}$ of SNP$_{max}$ ($\widehat{\beta_1}_{max}$) on a number of simulated scenarios, but soon discovered that it is not an effective estimate of $\widehat{\beta_1}$ of the causal SNP.

We already discussed in Chapter 2 how this can occur, especially with very small effect sizes. However, we would expect the SNPs in very high LD with the causal SNP to at least have higher ranks on average than the rest of the SNPs in the region. This led to the next method for estimating $\widehat{\beta_1}$ of the causal SNP, which is to take the top p% of SNPs ranked by likelihood and take the median value of $|\widehat{\beta_1}|$ for this group, denoted $\widehat{\beta_1}_p$. The choice of the median was due to the lower bound of zero and the skewed distribution. We investigated using different values of p in various causal SNP scenarios, and found that, in general, values around $p = 30$ work well for the region we have simulated.

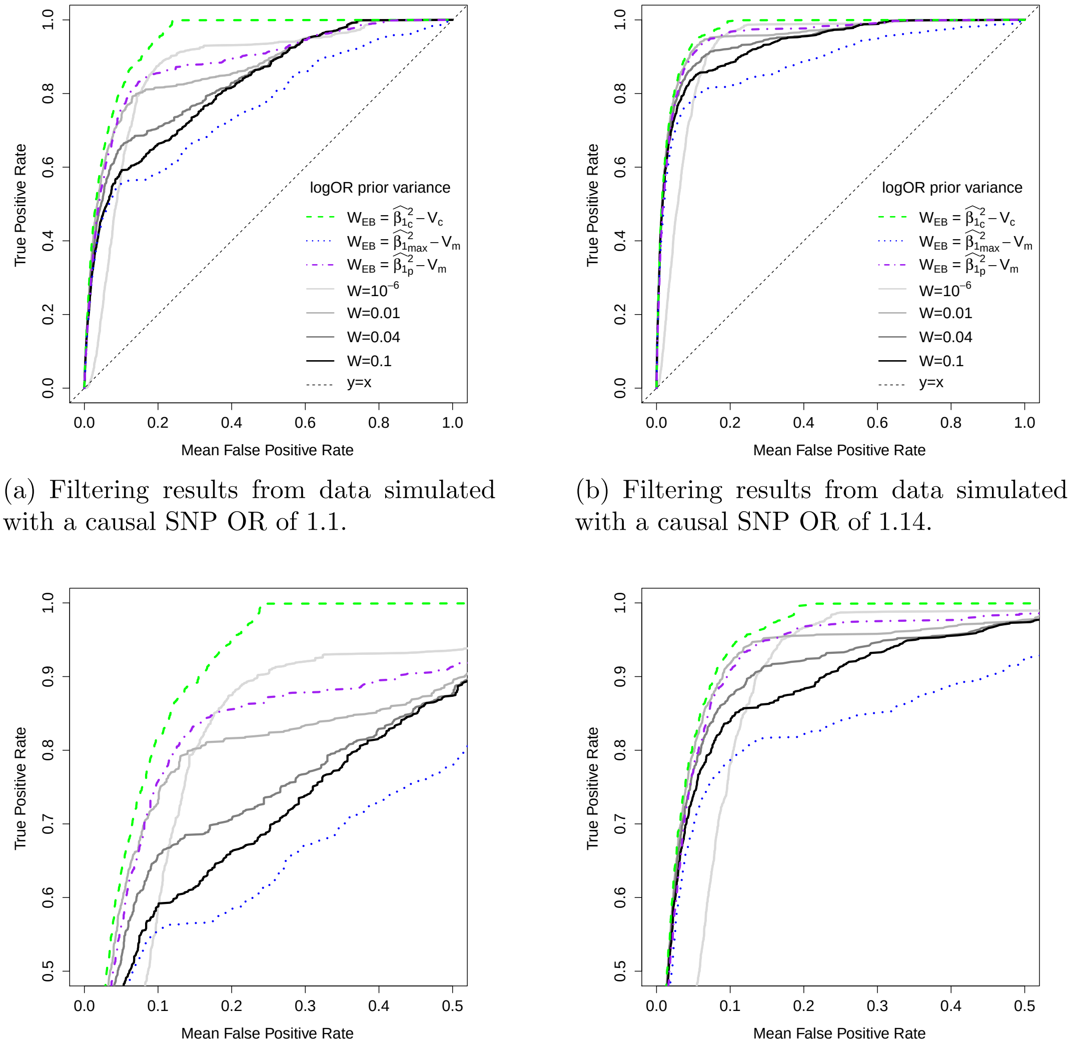
Figure 3.4 shows simulation filtering analysis results for two scenarios using the Wakefield Bayes factor where W, the prior variance of the logOR, is equal for all SNPs in a dataset. The causal SNP used in both scenarios has a MAF of 0.08 and the sample size is 20,000, but we have simulated data for an OR of 1.1 and also an OR of 1.14. The solid ROC curves show the results for a range of pre-specified Ws, whilst the dashed lines use W_{EB} values calculated in different ways. Figures 3.4(a) and 3.4(b) give the full ROC curves for the two scenarios, and Figures 3.4(c) and 3.4(d) focus on the parts of these curves for which FPR ≤ 0.5 and TPR ≥ 0.5. For both scenarios, the curves which clearly have the largest AUCs (94% and 97%) are from the analysis where W_{EB} is calculated using the values of $\widehat{\beta_1}$ and V of the true causal SNP ($\widehat{\beta_1}_c$ and V_c). This represents the ideal upper bound. For the analyses that give the other two dashed lines, the median of V across all SNPs was used (V_m). The figure shows that using $\widehat{\beta_1}_{max}$ in ineffective, giving the ROC curves with the lowest

AUCs (76% and 89%) in both of these plots. However, we can see that using a $\widehat{\beta_1}$ value estimated using the top 30% of SNPs ranked using likelihood ($\widehat{\beta}_{1p}$) improves filtering efficacy noticeably. In fact, for both of the scenarios shown, the AUCs for the ROC curves using these values of W_{EB} (89% and 96%) are just slightly higher than the largest AUCs of the ROC curves using pre-specified W values (87% and 95%).

We tested this method on a variety of fine-mapping scenarios and it seems to be generally effective. We therefore recommend empirically choosing a value of W for the prior logOR $\sim N(0, W)$ using the formula $W_{EB} = \widehat{\beta}_{1p=30}^{2} - V_m$; where $\widehat{\beta}_{1p=30}$ is the median $|\widehat{\beta_1}|$ of the top 30% of SNPs ranked by likelihood, and V_m is the median V of all SNPs, obtained by fitting logistic regression models. The value of 30% was chosen through investigation with simulated data using the set of 2871 SNPs that we are interested in, but a slightly different value may produce better results in a study concerned with a set of SNPs with a different MAF and LD structure. Therefore, we suggest investigating this with relevant simulated datasets prior to using W_{EB} on the real data. The utility of such a method is also likely to be restricted to scenarios with a single causal SNP.

3.3 Bayes factor approximations incorporating uncertainty

A limitation of the WBF is that the prior distribution of the logOR, β_1, must take the form $N(0, W)$. We suggest that elicitation is performed with an expert in order to identify the most appropriate value of W to use, by finding $p(\beta_1 < \beta_{1,p} | \beta_1 \sim N(0, W)) = p$; where $\beta_{1,p}$ is the p^{th} percentile of the distribution function of β_1. W is calculated using Φ, the distribution function of the standard normal distribution: $W = \{\beta_{1,p}/\Phi^{-1}(p)\}^2$ [60]. With values elicited with our breast cancer expert, we compared the closeness of their fit to a normal distribution, as described, and their fit to a variety of Student's t-distributions, allowing for heavier tails. We found that the normal distribution generally fit best, but that there was uncertainty in the expert's beliefs. We expect this expert uncertainty about the value of W to be a common occurrence. This may be problematic, as we have already shown that the results of fine-mapping analysis using WBF are highly dependent on the choice of W. In a situation where the expert is only confidently able to specify that the 80th percentile of the prior distribution for the OR is likely to be between 1.05 and 1.3, we have that $0.003 \leq W \leq 0.1$. We therefore wanted to allow for this uncertainty about W in the BF calculations. We have been able to determine a number



(a) Filtering results from data simulated with a causal SNP OR of 1.1.

(b) Filtering results from data simulated with a causal SNP OR of 1.14.

(c) Filtering results from data simulated with a causal SNP OR of 1.1, focussing on the parts of the ROC curves for which FPR ≤ 0.5 and TPR ≥ 0.5.

(d) Filtering results from data simulated with a causal SNP OR of 1.14, focussing on the parts of the ROC curves for which FPR ≤ 0.5 and TPR ≥ 0.5.

Figure 3.4: Receiver operating characteristic (ROC) curves of BF filtering results, each using the Wakefield approximation and $N(0, W)$ prior for the logOR with a different value of W, some based on empirical information (W_{EB}). Those which use empirical Bayes methods have subscripts denoting whether they are based on the causal SNP (c), the likelihood-based top hit (max), the median across all SNPs (m) or the median across the top $p\%$ of SNPs (p). The filtering was carried out on 1000 datasets simulated using the LD structure of the *CASP8* region for two scenarios with a causal SNP that has a MAF of 0.08 and a total sample size of 20,000.

of prior distributions for W which lead to BF approximations that can be easily calculated. Although the ease of calculation informed our choice of priors, they are flexible enough to accommodate a wide range of beliefs about expert uncertainty.

3.3.1 Novel Bayes factors allowing for uncertainty in W

By using similar methodology to Wakefield [60], we were able to develop 4 novel forms of Bayes factor approximation which still have the prior $\beta_1 \sim N(0, W)$, but each also puts a different prior on W. Theoretically, W could have a prior distribution of any form, but we wished to determine tractable Bayes factor approximations. The Bayes factor approximation includes integration over the variable parameters, in this case β_1 and W. As Wakefield integrated out β_1 using the normal density, W can also be integrated out if the integral is in the form of a standard probability density for which the cumulative distribution function can be expressed exactly. We were able to work backwards from some of these probability densities to come up with four prior forms for W. Three of these priors take the form of parametric families and the fourth is less flexible but may be useful in some scenarios. The range of distributions that these prior forms yield make them able to allow for a variety of different kinds of expert uncertainty. Table 3.1 contains these forms for the prior on W (up to proportionality) and the derivations of the BF approximations are demonstrated below. All 4 new BFs can be easily calculated in **R** [41] (code is provided in Appendix B). In all cases $0 < a \leq W \leq b$.

Name of prior	$f(W) \propto$	Restrictions on hyperparameters
Power	$(V+W)^k$	$k < -\frac{1}{2}$
Exponential	$\exp(-cW/2)$	$c > 0$
Hybrid	$(V+W)^k \exp\left(-\frac{d}{2(V+W)}\right)$	$d > -\widehat{\beta}_1, k < -1$
Reciprocal	$\frac{1}{(V+W)} \exp\left(-\frac{(V+W)}{2}\right)$	

Table 3.1: Density functions for each of the four prior forms (applies for $0 < a \leq W \leq b$).

Deriving the Bayes factor when $f(W) = q(V+W)^k$ for $k \leq -1/2$

The power prior Bayes factor (PPBF) approximation can be written

$$\begin{aligned}\text{PPBF} =& \frac{\int_W \int_{\beta_1} p(\widehat{\beta_1}|\beta_1) f(\beta_1|W) f(W)\, d\beta_1\, dW}{p(\widehat{\beta_1}|\beta_1 = 0)} \\ =& \frac{1}{Q} \int_W \int_{\beta_1} \frac{1}{\sqrt{2\pi V}} \exp\left(-\frac{(\widehat{\beta_1} - \beta_1)^2}{2V}\right) \frac{1}{\sqrt{2\pi W}} \exp\left(-\frac{\beta_1^2}{2W}\right) q(V+W)^k\, d\beta_1\, dW,\end{aligned} \tag{3.24}$$

where

$$q = \begin{cases} (k+1)[(V+b)^{k+1} - (V+a)^{k+1}]^{-1} & k \neq -1 \\ [\ln(V+b) - \ln(V+a)]^{-1} & k = -1 \end{cases}$$

is the normalising constant of the prior, and $Q = p(\widehat{\beta_1}|\beta_1 = 0) = \frac{1}{\sqrt{2\pi V}} \exp\left(-\frac{\widehat{\beta_1}^2}{2V}\right)$. As demonstrated by Wakefield in his approximation, β_1 can be integrated out by rearranging the integrand into the density of a normal distribution in β_1, giving

$$\begin{aligned}\text{PPBF} &= \frac{q}{Q} \int_W \frac{1}{\sqrt{2\pi(V+W)}} \exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) (V+W)^k\, dW \\ &= \frac{q}{Q\sqrt{2\pi}} \int_W (V+W)^{k-\frac{1}{2}} \exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) dW.\end{aligned} \tag{3.25}$$

Similarly, the integrand above takes the form of an inverse gamma density $f(y; \alpha, \gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} \exp\left(-\frac{\gamma}{y}\right)$ with shape and scale parameters $-(k + \frac{1}{2})$ and $\widehat{\beta_1}^2/2$, respectively, although for this to apply, we must restrict $k < -\frac{1}{2}$.

$$\begin{aligned}\text{PPBF} =& \frac{q}{Q\sqrt{2\pi}} \frac{\Gamma(-k-\frac{1}{2})}{(\widehat{\beta_1}^2/2)^{-k-\frac{1}{2}}} \int_{W=a}^{W=b} \frac{(\widehat{\beta_1}^2/2)^{-k-\frac{1}{2}}}{\Gamma(-k-\frac{1}{2})} (V+W)^{k-\frac{1}{2}} \exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) dW \\ =& \frac{q}{Q\sqrt{2\pi}} \frac{\Gamma(-k-\frac{1}{2})}{(\widehat{\beta_1}^2/2)^{-k-\frac{1}{2}}} \int_{W=a+V}^{W=b+V} \frac{(\widehat{\beta_1}^2/2)^{-k-\frac{1}{2}}}{\Gamma(-k-\frac{1}{2})} W^{k-\frac{1}{2}} \exp\left(-\frac{\widehat{\beta_1}^2}{2W}\right) dW \\ =& \frac{q}{Q\sqrt{2\pi}} \frac{\Gamma(-k-\frac{1}{2})}{(\widehat{\beta_1}^2/2)^{-k-\frac{1}{2}}} \left[\frac{\Gamma(-k-\frac{1}{2}, \frac{\widehat{\beta_1}^2}{2W})}{\Gamma(-k-\frac{1}{2})}\right]_{W=a+V}^{W=b+V},\end{aligned}$$

where $\Gamma(s, x) = \int_x^\infty t^{s-1} \exp(-t)\, dt$ is the upper incomplete gamma function. Substituting in Q and the normalising constant q in the forms given above results in

$$\text{PPBF} = \begin{cases} \dfrac{(k+1)\sqrt{V}\left[\Gamma\left(-k-\frac{1}{2},\frac{\widehat{\beta_1}^2}{2(b+V)}\right)-\Gamma\left(-k-\frac{1}{2},\frac{\widehat{\beta_1}^2}{2(a+V)}\right)\right]}{\left(\widehat{\beta_1}^2/2\right)^{-k-\frac{1}{2}}\exp\left(-\frac{\widehat{\beta_1}^2}{2V}\right)\left[(V+b)^{k+1}-(V+a)^{k+1}\right]} & k \neq -1 \\ \\ \dfrac{\sqrt{2V}\left[\Gamma\left(\frac{1}{2},\frac{\widehat{\beta_1}^2}{2(b+V)}\right)-\Gamma\left(\frac{1}{2},\frac{\widehat{\beta_1}^2}{2(a+V)}\right)\right]}{\widehat{\beta_1}\exp\left(-\frac{\widehat{\beta_1}^2}{2V}\right)\ln\left(\frac{V+b}{V+a}\right)} & k = -1. \end{cases}$$

Deriving the Bayes factor when $f(W) = r\exp\left(-\frac{cW}{2}\right)$ for $c > 0$

The normalising constant for the exponential prior is

$$r = \frac{c}{2}\left[\exp\left(-\frac{ca}{2}\right) - \exp\left(-\frac{cb}{2}\right)\right]^{-1}.$$

To derive the exponential prior Bayes factor (EPBF), the initial steps are the same as those that lead to Equation (3.25) for the PPBF. The equivalent result with this prior is

$$\text{EPBF} = \frac{r}{Q}\int_a^b \frac{1}{\sqrt{2\pi(V+W)}}\exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right)\exp\left(-\frac{cW}{2}\right)dW,$$

using the same definition of Q as previously. In this case, the integrand can be re-written in the form of a generalized inverse Gaussian density with parameters $c > 0$, $\widehat{\beta_1}^2$ and $\frac{1}{2}$:

$$\begin{aligned} \text{EPBF} =& \frac{2r\exp(\frac{cV}{2})K_{\frac{1}{2}}\left(\sqrt{c\widehat{\beta_1}^2}\right)}{Q\sqrt{2\pi}(c/\widehat{\beta_1}^2)^{\frac{1}{4}}} \\ &\times\int_{W=a}^{W=b}\frac{(c/\widehat{\beta_1}^2)^{\frac{1}{4}}(V+W)^{-\frac{1}{2}}}{2K_{\frac{1}{2}}\left(\sqrt{c\widehat{\beta_1}^2}\right)}\exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}-\frac{c(V+W)}{2}\right)dW \\ =& \frac{2r\exp(\frac{cV}{2})K_{\frac{1}{2}}\left(\sqrt{c\widehat{\beta_1}^2}\right)}{Q\sqrt{2\pi}(c/\widehat{\beta_1}^2)^{\frac{1}{4}}}\int_{W=a+V}^{W=b+V}\frac{(c/\widehat{\beta_1}^2)^{\frac{1}{4}}W^{-\frac{1}{2}}}{2K_{\frac{1}{2}}\left(\sqrt{c\widehat{\beta_1}^2}\right)}\exp\left(-\frac{\widehat{\beta_1}^2}{2W}-\frac{cW}{2}\right)dW \\ =& \frac{2r\exp(\frac{cV^2+\widehat{\beta_1}^2}{2V})\sqrt{V}K_{\frac{1}{2}}\left(\sqrt{c\widehat{\beta_1}^2}\right)}{(c/\widehat{\beta_1}^2)^{\frac{1}{4}}}\int_{W=a+V}^{W=b+V}\frac{(c/\widehat{\beta_1}^2)^{\frac{1}{4}}W^{-\frac{1}{2}}}{2K_{\frac{1}{2}}\left(\sqrt{c\widehat{\beta_1}^2}\right)}\exp\left(-\frac{cW}{2}-\frac{\widehat{\beta_1}^2}{2W}\right)dW. \end{aligned}$$

where $K_{\frac{1}{2}}(.)$ is a modified Bessel function of the second kind. There is no closed form expression for the distribution function of a generalized inverse Gaussian density, but the integrand, $W \sim GIG(c, \widehat{\beta_1}^2, \frac{1}{2})$, can be calculated in **R** using the `pgig` command.

Deriving the Bayes factor when $f(W) = s(V+W)^k \exp\left(-\frac{d}{2(V+W)}\right)$ for $d > -\widehat{\beta_1}^2$, $k < -1$

The hybrid prior takes the form of an inverse gamma density, and this can be used to find the normalising constant,

$$s = \left(\frac{d}{2}\right)^{-k-1} \left[\Gamma\left(-k-1, \frac{d}{2(b+V)}\right) - \Gamma\left(-k-1, \frac{d}{2(a+V)}\right)\right]^{-1}.$$

However, this intrinsically applies the limit $k < -1$.

Following the steps which result in Equation (3.25) for the PPBF, the hybrid prior Bayes factor (HPBF) takes the form

$$\text{HPBF} = \frac{s}{Q\sqrt{2\pi}} \int_a^b (V+W)^{k-\frac{1}{2}} \exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) \exp\left(-\frac{d}{2(V+W)}\right) dW.$$

As with the PPBF, the integrand above takes the form of an inverse gamma density for the values of k we have specified. In this case the shape and scale parameters are $-(k+\frac{1}{2})$ and $(\widehat{\beta_1}^2 + d)/2$, respectively.

$$\begin{aligned}
\text{HPBF} =& \frac{s}{Q\sqrt{2\pi}} \frac{\Gamma(-k-\frac{1}{2})}{((\widehat{\beta_1}^2+d)/2)^{-k-\frac{1}{2}}} \\
&\times \int_{W=a}^{W=b} \frac{((\widehat{\beta_1}^2+d)/2)^{-k-\frac{1}{2}}}{\Gamma(-k-\frac{1}{2})} (V+W)^{k-\frac{1}{2}} \exp\left(-\frac{(\widehat{\beta_1}^2+d)}{2(V+W)}\right) dW \\
=& \frac{s}{Q\sqrt{2\pi}} \frac{\Gamma(-k-\frac{1}{2})}{((\widehat{\beta_1}^2+d)/2)^{-k-\frac{1}{2}}} \int_{W=a+V}^{W=b+V} \frac{((\widehat{\beta_1}^2+d)/2)^{-k-\frac{1}{2}}}{\Gamma(-k-\frac{1}{2})} W^{k-\frac{1}{2}} \exp\left(-\frac{(\widehat{\beta_1}^2+d)}{2W}\right) dW \\
=& \frac{s}{Q\sqrt{2\pi}} \frac{\Gamma(-k-\frac{1}{2})}{((\widehat{\beta_1}^2+d)/2)^{-k-\frac{1}{2}}} \left[\frac{\Gamma(-k-\frac{1}{2}, \frac{(\widehat{\beta_1}^2+d)}{2W})}{\Gamma(-k-\frac{1}{2})}\right]_{W=a+V}^{W=b+V} \\
=& \frac{\sqrt{2V}\exp\left(\frac{\widehat{\beta_1}^2}{2V}\right)\left[\Gamma\left(-k-\frac{1}{2}, \frac{(\widehat{\beta_1}^2+d)}{2(b+V)}\right) - \Gamma\left(-k-\frac{1}{2}, \frac{(\widehat{\beta_1}^2+d)}{2(a+V)}\right)\right]}{(\widehat{\beta_1}^2+d)^{-k-\frac{1}{2}} d^{k+1}\left[\Gamma\left(-k-1, \frac{d}{2(b+V)}\right) - \Gamma\left(-k-1, \frac{d}{2(a+V)}\right)\right]}.
\end{aligned}$$

Deriving the Bayes factor when $f(W) = \frac{t}{(V+W)} \exp\left(-\frac{(V+W)}{2}\right)$

The reciprocal prior approximate Bayes factor (RPBF) is a specific prior distribution with normalising constant $t = [\Gamma(0, (a+V)/2) - \Gamma(0, (b+V)/2)]^{-1}$. The form of the upper incomplete gamma function $\Gamma(0, z)$ is a special case and is calculated using the relationship $\Gamma(0, z) = -\mathrm{Ei}(-z) = -\gamma - \ln(z) - \sum_{n=0}^{\infty}(-1)^n \frac{z^n}{nn!}$, where Ei is the exponential integral and γ is the Euler-Mascheroni constant [22].

The RPBF can be written

$$\begin{aligned}\mathrm{RPBF} = \frac{t\exp(-|\widehat{\beta_1}|)}{Q} \int_a^b \frac{1}{\sqrt{2\pi}(V+W)^{\frac{3}{2}}} \exp\left(-\frac{\widehat{\beta_1}^2}{2(V+W)}\right) \\ \times \exp\left(-\frac{(V+W) - 2|\widehat{\beta_1}|}{2}\right) dW,\end{aligned}$$

which can be further simplified by writing the integrand as the density of an inverse Gaussian distribution with mean and scale parameters of $|\widehat{\beta_1}|$ and $\widehat{\beta_1}^2$ respectively.

$$\begin{aligned}\mathrm{RPBF} &= \frac{t\exp(-|\widehat{\beta_1}|)}{Q} \int_{W=a}^{W=b} \frac{1}{\sqrt{2\pi}(V+W)^{\frac{3}{2}}} \exp\left(-\frac{\left((V+W) - |\widehat{\beta_1}|\right)^2}{2(V+W)}\right) dW \\ &= \frac{t\exp(-|\widehat{\beta_1}|)}{Q|\widehat{\beta_1}|} \int_{W=a+V}^{W=b+V} \left[\frac{\widehat{\beta_1}^2}{2\pi W^3}\right]^{\frac{1}{2}} \exp\left(-\frac{\widehat{\beta_1}^2 (W - |\widehat{\beta_1}|)^2}{2\widehat{\beta_1}^2 W}\right) dW \\ &= \frac{t\exp(-|\widehat{\beta_1}|)}{Q|\widehat{\beta_1}|} \left[\Phi_+(W) + \exp(2|\widehat{\beta_1}|)\Phi_-(W)\right]_{W=a}^{W=b},\end{aligned}$$

where

$$\Phi_+(y) = \Phi\left(\sqrt{y+V} - \frac{|\widehat{\beta_1}|}{\sqrt{y+V}}\right), \quad \Phi_-(y) = \Phi\left(-\sqrt{y+V} - \frac{|\widehat{\beta_1}|}{\sqrt{y+V}}\right)$$

and $\Phi(.)$ is the distribution function of the standard normal distribution. So

$$\mathrm{RPBF} = \frac{\sqrt{2\pi V}\exp\left(\frac{\widehat{\beta_1}^2}{2V} - |\widehat{\beta_1}|\right)\left[\Phi_+(b) - \Phi_+(a) + (\Phi_-(b) - \Phi_-(a))\exp(2|\widehat{\beta_1}|)\right]}{|\widehat{\beta_1}|\left[\ln\left(\frac{V+b}{V+a}\right) + \sum_{n=1}^{\infty}\frac{(-1)^n}{nn!}\left(\left(\frac{b+V}{2}\right)^n - \left(\frac{a+V}{2}\right)^n\right)\right]}.$$

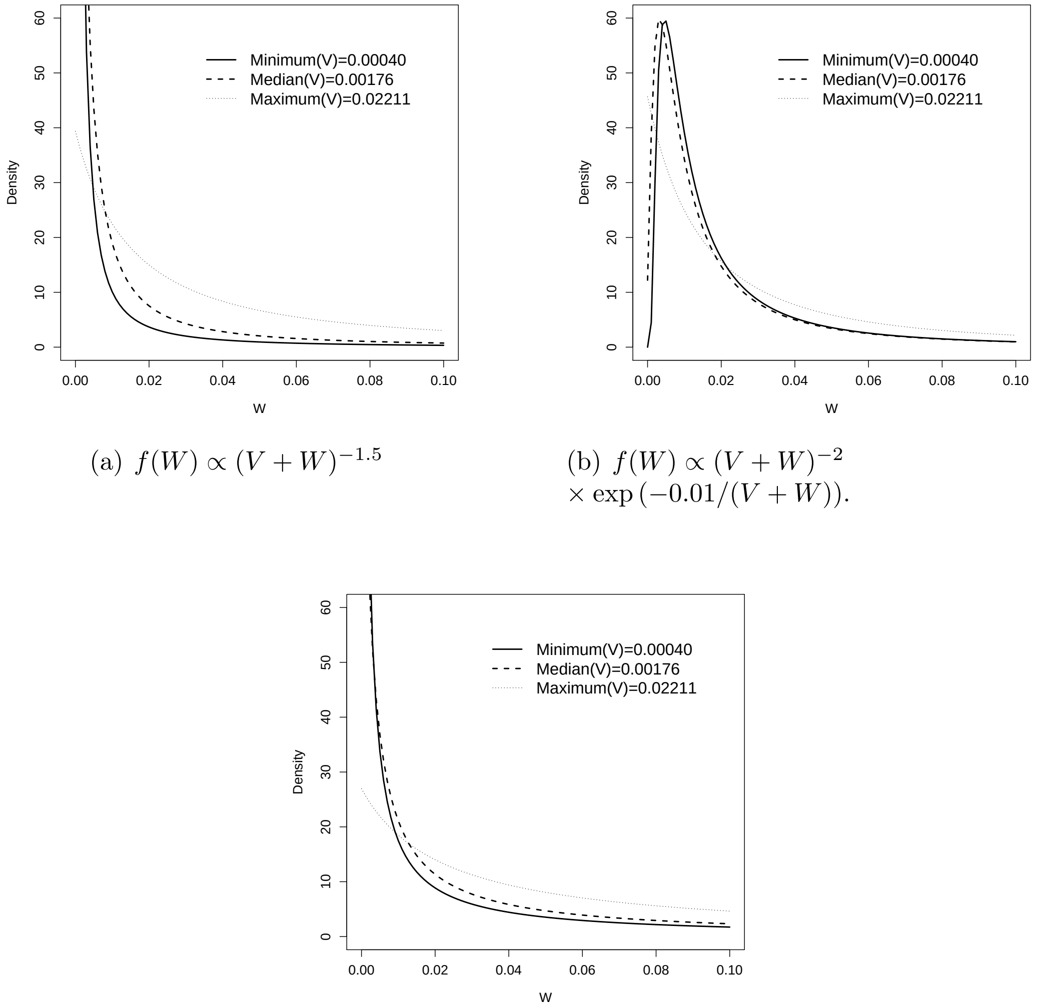
3.3.2 Properties of the priors on W defined for the novel Bayes factors

The dependence of the prior densities of W upon the genotype data

As with the W_{EB} prior variance described in §3.2.3, three of the prior forms, the power, hybrid and reciprocal priors, contain information from the genotype data. In this case, they depend on V, the asymptotic variance of the estimate of the logOR, and therefore may not be considered true priors. However, in order to carry out the integration in the BF calculations, such forms were necessary and we are able to show that, for the values likely to be encountered in large association studies, V has quite a small effect on the prior density of W.

In §3.2.3, we discussed the dependence of V on MAF and sample size. Once again, it is necessary to choose appropriate values of V for our investigations. The sample sizes we are considering should have the power to detect associations with most SNPs, except those with very small MAFs. We therefore we consider the Vs corresponding to SNPs with MAF ≥ 0.005 in one of our simulated datasets of size 20,000. These V values were distributed with minimum, median and maximum of 0.00040, 0.00176 and 0.02211 respectively. Using fixed values of the hyperparameters, we plotted the prior densities for W for the power, hybrid and reciprocal priors for these 3 values of V and these are given in Figure 3.5. These plots show the extent to which the prior for W can be expected to vary dependent on V in a sample size of 20,000. In particular, there appears to be little difference between the densities using the minimum and medium values of V for each prior, and larger V values up to the maximum are also reasonably similar.

We next consider SNPs with extreme values of V as these may lead to extreme priors and potentially large BFs. Because V is bounded below by 0, we only need to consider extreme large values of V. To consider the largest values of V likely to occur, we refer back to the largest value for a SNP with MAF ≥ 0.005 in the simulated sample of 20,000, which was 0.02211. If the number of cases and controls are equal and denoted by n, Slager and Schaid [44] showed that $V \propto 1/n$ approximately. We assume no fine-mapping studies with total sample size less than 2000 would be successful, and we can infer that studies of this size will yield most values of $V \leq 0.2$ for SNPs with MAF greater than 0.005. At smaller sample sizes such as this, we can expect to observe occasional rare SNPs (with MAFs less than 0.005) which will result in unusually large values of V. Although this will have a significant impact on the prior, these SNPs will have broad likelihoods and as such are unlikely to have high BFs. Therefore, although a small proportion of SNPs may have priors



(a) $f(W) \propto (V+W)^{-1.5}$

(b) $f(W) \propto (V+W)^{-2}$
$\times \exp\left(-0.01/(V+W)\right)$.

(c) $f(W) \propto (V+W)^{-1} \exp\left(-(V+W)/2\right)$

Figure 3.5: Prior densities of W given for minimum, median and maximum values of V for SNPs with MAF>0.005 in a sample size of 20,000.

that are different from the rest due to unusual V values, they are likely to be removed in the filtering process.

The dependence of the prior densities of W upon hyperparameters

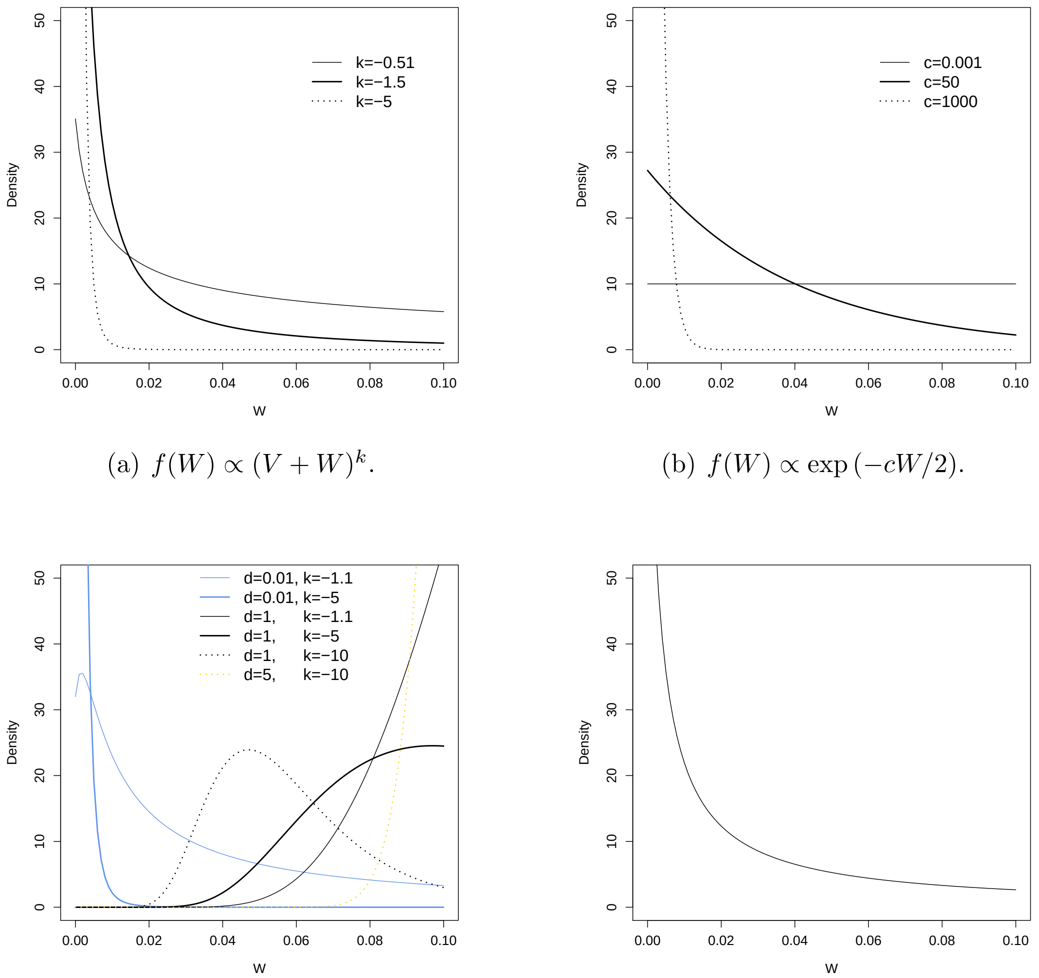
All four forms of the prior must be defined over an interval $0 < a \leq W \leq b$ and each of the power, exponential and hybrid priors are in fact families of priors dependent on further hyperparameters. We suggest choosing the values of a, b and the hyperparameters c, d and k via expert elicitation, which we go on to describe in §3.3.3.

Using $V = 0.003$ for those priors which depend on V, we give the densities of some possible priors in Figure 3.6, demonstrating the range of prior beliefs they are able to capture. The single reciprocal prior only allows for the support to be varied and therefore is very limited. Like this prior, the majority of other distributions also place most of the prior weight of W close to the value of a, the lower limit of the support. However, choosing a value of c close to 0 with the exponential prior results in an almost uniform distribution over W. This, along with its independence of the genotype data, is one of the main advantages for using this family of priors. The hybrid prior also has a unique advantage, as it is the only form which is not necessarily monotonically decreasing with W and can have a stationary point. Its hyperparameters can be chosen such that the mode is located anywhere in the support, specifically it is found at $W = -(V + d/2k)$. This allows for much more flexibility, and can model many more types of uncertainy.

Are these priors consistent with rare alleles having larger effects?

In §3.2.2 we attempted to define a formula for $W(\text{MAF})$ to take into account the suggestion that the effect size of causal SNPs may increase with decreasing MAF [61]. We now investigate whether the three forms of prior which depend on the data through V implicitly have this property. To assess this we examine how $\mathbb{E}(W)$ changes with V, over a support relevant to studies with sample sizes of 2000 or more. Since SNPs with lower MAFs have larger V [44], an appropriate prior would possess the property that $\mathbb{E}(W)$ is a non-decreasing function of V. Then as the MAF decreases, V increases and rarer SNPs have a priori larger effects on average.

Either using integration by parts with respect to (w.r.t.) W, or integrating w.r.t. $(V + W)$ and then using the property $\mathbb{E}(W) = \mathbb{E}(V + W) - V$, we were able to find the expected value of W for each of the prior forms. These are given in Table 3.2 and their dependence on V (where relevant) is plotted in Figure



(a) $f(W) \propto (V+W)^k$.

(b) $f(W) \propto \exp\left(-cW/2\right)$.

(c) $f(W) \propto (V+W)^k \exp\left(-\frac{d}{2(V+W)}\right)$.

(d) $f(W) \propto (V+W)^{-1}$
$\times \exp\left(-(V+W-2|\widehat{\beta_1}|)/2\right)$.

Figure 3.6: Densities of three families of tractable priors and one specific prior for $f(W)$ $(0 < W \leq 0.1)$ where β_1 is log odds ratio with $\beta_1 \sim N(0,W)$. A value of $V = 0.003$ is used in plots (a), (c) and (d).

3.7 for values of V likely to occur. For those forms that are dependent on V, we were unable to show algebraically that the expectation of W is a non-decreasing function of V, although Figure 3.7(c) shows that the reciprocal prior possesses this property. We can see that this is also the case for the power priors plotted in Figure 3.7(a), which use quite different hyperparametric values, so we assume that this applies to all values likely to be used. However, $d\mathbb{E}(W)/dV < 0$ in places for some realisations of the hybrid prior, as shown in Figure 3.7(b). Based on our observations, we suggest that researchers who have prior beliefs that there is a negative relationship between effect size and MAF take care if they wish to use a hybrid prior. In particular, we recommend restricting the value of d to close to zero. Investigators can check whether they believe the dependence on V is appropriate using $\mathbb{E}(W)$ as given in Table 3.2.

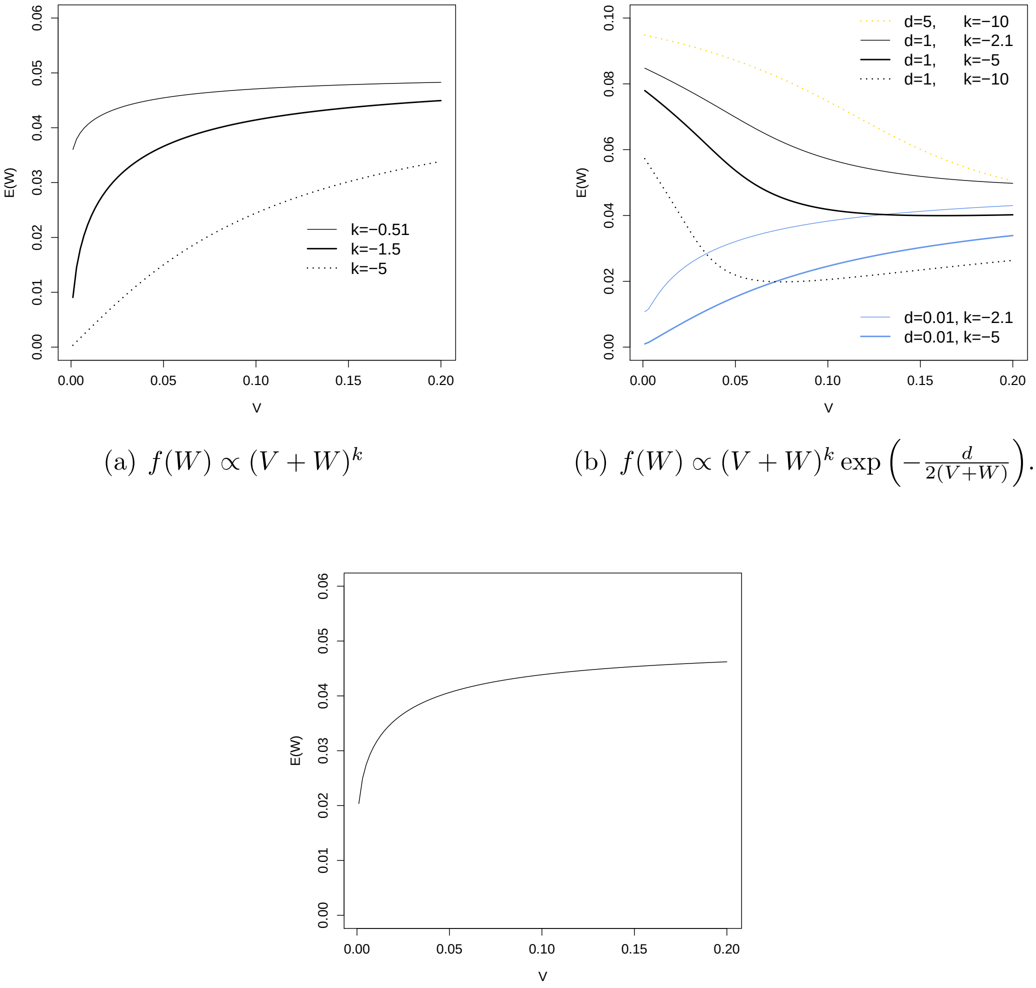
If it is not possible to find a form for the prior on W which fits an expert's overall beliefs closely and also fits their beliefs about the relationship between MAF and effect size, it may be better not to employ our novel BFs, but instead use a generalization of the Savage-Dickey density ratio [56]. We have limited the forms of prior available, as our BF calculations require integration, but the generalization of the Savage-Dickey density ratio approximates the BF without the need to do this. Therefore, many more priors are available and a prior of the form $W^k \exp(-d/2W)$, for example, could be used to calculate a BF using this method. The BF calculation is instead based on posterior sampling, using a method such as MCMC.

Type of prior	$\mathbb{E}(W)$	Limitations
Power	$\frac{[b(k+1)-V](V+b)^{k+1}-[a(k+1)-V](V+a)^{k+1}}{(k+2)[(V+b)^{k+1}-(V+a)^{k+1}]}$	$k < -\frac{1}{2}$
Exponential	$\frac{(a+\frac{2}{k})\exp\left(-\frac{ka}{2}\right)-(b+\frac{2}{k})\exp\left(-\frac{kb}{2}\right)}{\exp\left(-\frac{ka}{2}\right)-\exp\left(-\frac{kb}{2}\right)}$	$c > 0$
Hybrid	$\frac{d[\Gamma\left(-k-2,\frac{d}{2(V+b)}\right)-\Gamma\left(-k-2,\frac{d}{2(V+a)}\right)]}{2[\Gamma\left(-k-1,\frac{d}{2(V+b)}\right)-\Gamma\left(-k-1,\frac{d}{2(V+a)}\right)]}-V$	$d > -\widehat{\beta}_1, k < -2$
Reciprocal	$\frac{2\left(\exp\left(-\frac{V+a}{2}\right)-\exp\left(-\frac{V+b}{2}\right)\right)}{\ln\left(\frac{b+V}{a+V}\right)+\sum_{n=1}^{\infty}\frac{(-1)^n}{nn!}\left(\left(\frac{b+V}{2}\right)^n-\left(\frac{a+V}{2}\right)^n\right)}-V$	

Table 3.2: Expected value of W for each of the four prior forms.

3.3.3 Eliciting hyperparameters of the priors for W

The motivation for putting a prior distribution on W was inconsistency or uncertainty by experts when elicitation was employed to determine a fixed value



(a) $f(W) \propto (V+W)^k$

(b) $f(W) \propto (V+W)^k \exp\left(-\frac{d}{2(V+W)}\right)$.

(c) $f(W) \propto (V+W)^{-1} \exp\left(-(V+W)/2\right)$

Figure 3.7: $\mathbb{E}(W)$ as a function of V for empirical forms of the the prior ($0 < W \leq 0.1$). $\mathbb{E}(W)$ is given over a range of V likely to been seen in sample sizes of 2000 or greater with different values of the hyperparameters, where relevant.

of W. However, we have now specified a whole range of possible prior distributions and are still faced with the problem of how to find the one that fits an expert's beliefs most closely. For all forms, the support of Wat least must be specified, but it is likely that one of the forms with further hyperparameters will be required and therefore these will also need to be defined.

The appropriate distribution can be elicited from an expert using a similar method to when W was fixed. We give an example using the power prior form. To begin with the cumulative distribution is needed:

$$F(W) = \frac{(V+W)^{k+1} - (V+a)^{k+1}}{(V+b)^{k+1} - (V+a)^{k+1}} \quad a \leq W \leq b. \tag{3.26}$$

If it is possible to elicit a single value, w_1, at the p_1^{th} percentile of the distribution of W, Equation (3.26) can be equated to p_1 and solved to find k by replacing W with w_1. However, a better fit will be found if more percentiles can be elicited, for example, h percentiles $(p_1, p_2, ..., p_h)$ of W $(w_1, w_2, ..., w_h)$. These values can then be used to solve

$$\hat{k} = \text{argmin}_k \sum_{i=1}^{h} \left(\frac{(V+w_i)^{k+1} - (V+a)^{k+1}}{(V+b)^{k+1} - (V+a)^{k+1}} - p_i \right)^2 . \tag{3.27}$$

Similar methods can be employed with the exponential and hybrid priors, using their distribution functions, as given in Table 3.3, although the hybrid prior requires a search over a two dimensional space for the pair of hyperparameters which optimise the prior.

Type of prior	$F(W)$	Limitations
Power	$\dfrac{(V+W)^{k+1} - (V+a)^{k+1}}{(V+b)^{k+1} - (V+a)^{k+1}}$	$k < -\frac{1}{2}, k \neq -1$
	$\dfrac{\ln\left(\frac{V+W}{V+a}\right)}{\ln\left(\frac{V+b}{V+a}\right)}$	$k = -1$
Exponential	$\dfrac{\exp\left(-\frac{cW}{2}\right) - \exp\left(-\frac{ca}{2}\right)}{\exp\left(-\frac{cb}{2}\right) - \exp\left(-\frac{ca}{2}\right)}$	$c > 0$
Hybrid	$\dfrac{\Gamma\left(-k-1, \frac{d}{2(V+W)}\right) - \Gamma\left(-k-1, \frac{d}{2(V+a)}\right)}{\Gamma\left(-k-1, \frac{d}{2(V+b)}\right) - \Gamma\left(-k-1, \frac{d}{2(V+a)}\right)}$	$d > -\widehat{\beta}_1, k < -1$
Reciprocal	$\dfrac{\ln\left(\frac{W+V}{a+V}\right) + \sum_{n=1}^{\infty} \frac{(-1)^n}{nn!}\left(\left(\frac{W+V}{2}\right)^n - \left(\frac{a+V}{2}\right)^n\right)}{\ln\left(\frac{b+V}{a+V}\right) + \sum_{n=1}^{\infty} \frac{(-1)^n}{nn!}\left(\left(\frac{b+V}{2}\right)^n - \left(\frac{a+V}{2}\right)^n\right)}$	

Table 3.3: Distribution functions for each of the four prior forms.

Practically, it can be difficult for experts to choose percentiles from a distribution as described. One way which can make it slightly more user-friendly is to instead ask experts to think about the distribution of ORs and to envisage central probability intervals (PIs) for this distribution. We suggest encouraging the subject to choose the particular $z_i\%$ probability intervals themselves. These can then be used to find $p_i = 1-(1-0.01z_i)/2$. The expert should be asked to give their estimate of $\text{PI}_{u,i}$, the upper limit of the $z_i\%$ centralised probability interval for the OR, from which we are able to calculate $w_i = (\ln(\text{PI}_{u,i})/\Phi^{-1}(p_i))^2$.

When carrying out elicitation for either the power or hybrid priors, the optimisation applies to a single value of V, but V is SNP-specific and we ideally want to find the best hyperparameters for all SNPs in the dataset. We suggest fitting the univariate logistic regression models to the data and finding the set of Vs (the squares of the standard errors of the parameter estimates) that apply to those SNPs. The median V can be used in the elicitation, as a representative value for all SNPs. To calculate $\hat{k}$, we also need to specify the limits of the support of W, a and b. For this purpose, it may be possible to elicit a range of plausible values of W at a single percentile, p, once again using the centralised probability interval (PI) method and denoting the elicited minimum and maximum as $\text{PI}_{u,min}$ and $\text{PI}_{u,max}$. The lower limit of the support, a, would then be calculated using $a = (\ln(\text{PI}_{u,min})/\Phi^{-1}(p))^2$, and similarly the upper limit, b, would be found replacing $\text{PI}_{u,min}$ with $\text{PI}_{u,max}$.

To demonstrate this elicitation method, we have carried it out with our expert on breast cancer genetics to determine the prior on W that best fits their beliefs about the causal SNP in the *CASP8* region of chromosome 2. They believe that there is a causal SNP with a small effect size in this region, and we initially asked them to give a range of possible values relating to a single percentile of the OR. They were confident that the 80^{th} percentile of the OR would be in the range of 1.05 to 1.3, which gives us $a = 0.003$ and $b = 0.1$. They were able to give more specific estimates for the 95%, 75% and 50% centralised PIs of the ORs. They provided upper limits of 1.43, 1.21 and 1.14 for these three PIs respectively, yielding $(w_1,\ w_2,\ w_3) = (0.0333,\ 0.0275,\ 0.0377)$. Previously we found the median V in a simulated dataset of size 20,000 to be 0.00176. Using this as the V in the calculations, we were able to find the hyperparameters that resulted in the power, exponential and hybrid priors that fitted the experts beliefs most closely. For example, carrying out a search over $-10 \leq k < -0.05$ at intervals of 0.01 for the PPBF, we found that the minimum sum of squared differences occured at $k = -1.83$. Similarly, we were able find the best fitting priors to be using $c = 145$ for the EPBF and $d = 0.001,\ k = -1.86$ for the HPBF. However, the sum of squared differences for both of these priors was slightly larger than that for the best fitting power prior, so this is the prior that we would choose to use.

We have written **R** code to carry out a search over the hyperparameters for $\hat{k}$ which can be used to find the most appropriate forms of the power, exponential or hybrid priors. We used the values of the hyperparameters in Figure 3.6 to inform the space over which to search, and the code is given in Appendix C. The required form of the prior must be specified, but the output includes the minimum sum of squares from Equation (3.27) (or the equivalents for the EPBF and HPBF), so that it can be run for all forms and that which has the best fit (smallest minimum sum of squares) can be determined. Alternatively, the hyperparameters may be determined by empirical Bayes methods, in which they are signified by (Λ) and $\text{argmax}_{\Lambda}(p(data|\Lambda))$ must be solved. In the case of our priors, this would involve the maximisation of the BF over Λ, which cannot be done analytically.

3.3.4 Fine-mapping using novel BFs on simulated data

We carried out filtering using likelihood percentiles (LP), WBFs and our new BFs on simulated data for several scenarios. We present results for 1000 simulated datasets with a single causal SNP with a MAF of 0.08, an OR of 1.1 and a sample size of 20,000, as well as for data simulated with the same causal SNP but a sample size of only 4000 and several different ORs. Table 3.4 shows the area under the curve (AUC) for ROC curves using the true and mean false positive rates calculated from these analyses. We have highlighted the methods which have resulted in the three highest AUCs for each scenario. Using this measure, LP, which we found to be the most efficacious method that doesn't require prior specification, is constantly ranked among the top three of the filters we have considered, although the most successful BF methods have similar AUCs. When the small sample size of 4000 is used with BF filtering, the prior has much more weight, so the specification is more important. None of the fixed W values we used happened to produce particularly good results for these scenarios, so they do not rank in the top AUCs, but it was possible to achieve better AUCs with some realisations of the novel BFs. It also appears that there is a little more variation in the AUCs when W is varied in the WBF compared to varying the hyperparameters for the PPBF or the HPBF. These new priors average over the WBFs as W varies and are therefore a powerful tool for dealing with uncertainty in W.

Table 3.5 shows the median rank (and other quartiles) of the causal SNP across the 1000 datasets for each scenario using the different filtering methods. In this table, the methods which produce the 3 smallest median ranks for

Analysis method	Parameter values	Sample size and odds ratio			
		SS=20,000 OR=1.1	SS=4000 OR=1.1	SS=4000 OR=1.14	SS=4000 OR=1.18
LP		**90**	**62**	**72**	**80**
WBF	$W = 0.003$	**89**	60	70	79
	$W = 0.03$	85	58	68	76
	$W = 0.1$	82	55	65	73
PPBF	$k = -0.51$	85	58	68	76
	$k = -1.5$	87	60	70	78
	$k = -5$	**89**	**63**	**71**	**79**
HPBF	$d = 0.01,\ k = -1.1$	86	59	69	76
	$d = 0.01,\ k = -5$	88	**62**	**71**	**79**
	$d = 0.05,\ k = -1.1$	85	57	66	74
	$d = 1,\ k = -1.1$	83	58	67	74
	$d = 1,\ k = -5$	83	59	68	76
	$d = 1,\ k = -10$	83	58	67	74
	$d = 5,\ k = -10$	83	59	69	76
RPBF		86	60	68	77

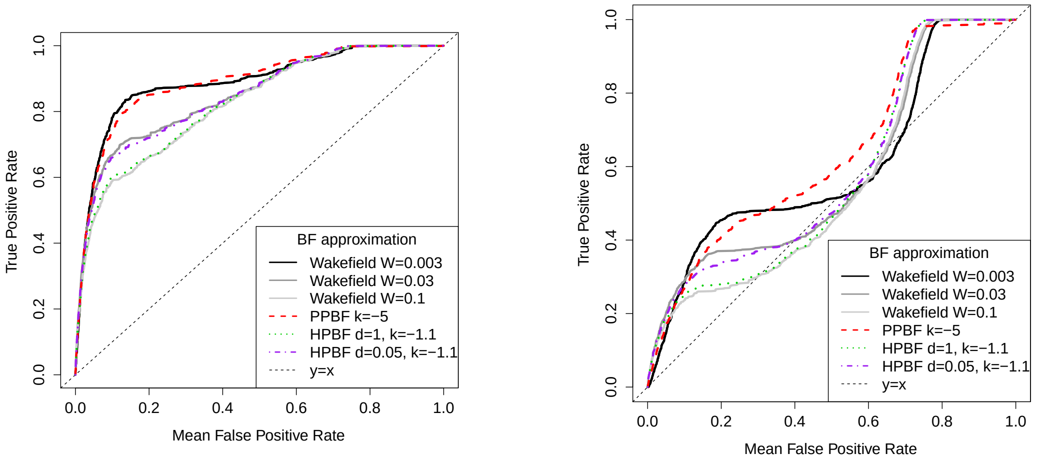
Table 3.4: Results of different methods of filtering on several simulated scenarios in terms of the area under the curve (AUC) of receiver operating characteristic curves, given as a percentage. For the power, hybrid and reciprocal prior Bayes factors (PPBF, HPBF and RPBF), we use $0.003 \leq W \leq 0.1$, and the Wakefield Bayes factor (WBF) and likelihood percentile (LP) methods are also included. For each scenario, 1000 datasets were simulated using the LD structure of the *CASP8* region for a scenario with a single causal SNP that has a MAF of 0.08.

each scenario are highlighted. However, care should be taken when using this to measure the efficacy of filters, as the upper quartile is sometimes high for these same methods. It is interesting to see that using this measure, LP is not always ranked among the top three methods. In particular, for the scenario with sample size 4000 and OR 1.18, it only has the 8th smallest median rank for the causal SNP. This suggests that when using a very small sample size, including information through a prior variance for the logOR can be important. The median rank of the causal SNP is highly variable dependent on the method of filtering, especially for scenarios with very small sample sizes and ORs.

Figure 3.8(a) shows several of the ROC curves resulting from the analysis with 20,000 samples. It includes the results of filtering using the WBF approximation with several values for W and also the results of filtering using one power prior and two forms of hybrid prior for comparison. There is clearly a lot of variation in the effectiveness of the Wakefield BF filter as W changes. Putting too much prior weight on large effect sizes clearly leads to poor performance of the WBF when the actual causal effect size is small, as in this case. The power prior shown puts much of the weight close to $a = 0.003$ but also

Analysis method	Parameter values	Sample size and odds ratio			
		SS=20,000 OR=1.1	SS=4000 OR=1.1	SS=4000 OR=1.14	SS=4000 OR=1.18
LP		126	**896**	504	344
		(39, 348)	**(335,1778)**	(164,1309)	(118,782)
WBF	$W = 0.003$	**91**	1270	**356**	**278**
		(32, 241)	(250,1986)	**(148,1792)**	**(126,624)**
	$W = 0.03$	**122**	1516	**488**	**244**
		(30, 688)	(232,1922)	**(100,1760)**	**(79,1367)**
	$W = 0.1$	188	1528	1088	484
		(42, 839)	(416,1909)	(137,1756)	(92,1422)
PPBF	$k = -0.51$	146	1502	740	333
		(56, 615)	(324,1940)	(157,1764)	(134,1375)
	$k = -1.5$	133	1380	533	331
		(55, 385)	(320,1877)	(169,1672)	(142,1209)
	$k = -5$	**119**	**1052**	**472**	365
		(56, 307)	**(323,1832)**	**(188,1582)**	(188,836)
HPBF	$d = 0.01$,	143	1472	611	332
	$k = -1.1$	(54, 519)	(316,1947)	(160,1763)	(136,1325)
	$d = 0.01$,	131	**1134**	519	367
	$k = -5$	(68, 317)	**(353,1966)**	(237,1707)	(190,862)
	$d = 0.05$,	151	1538	1110	418
	$k = -1.1$	(56, 739)	(394,1906)	(171,1756)	(135,1430)
	$d = 1$,	197	1542	1120	397
	$k = -1.1$	(64, 847)	(345,1887)	(211,1760)	(164,1445)
	$d = 1$,	198	1562	972	**327**
	$k = -5$	(73, 862)	(326,1879)	(197,1768)	**(145,1462)**
	$d = 1$,	171	1513	1118	464
	$k = -10$	(69, 886)	(438,1879)	(226,1757)	(154,1396)
	$d = 5$,	211	1516	736	330
	$k = -10$	(77, 844)	(321,1912)	(155,1761)	(133,1371)
RPBF		150	1474	766	376
		(67, 475)	(388,1952)	(313,1776)	(190,1320)

Table 3.5: Results of different methods of filtering on several simulated scenarios in terms of the median rank (and other quartiles) for the true causal SNP (among 2871 SNPs in total). For the power, hybrid and reciprocal prior Bayes factors (PPBF, HPBF and RPBF), we use $0.003 \leq W \leq 0.1$, and the Wakefield Bayes factor (WBF) and likelihood percentile (LP) methods are also included. For each scenario, 1000 datasets were simulated using the LD structure of the *CASP8* region for a scenario with a single causal SNP that has a MAF of 0.08.



(a) ROC curves for BF filtering results using a total sample size of 20,000.

(b) ROC curves for BF filtering results using a total sample size of 4000.

Figure 3.8: ROC curves showing the results of WBF filtering using 3 different values of W compared to the results for the PPBF with a prior distribution on W ($k = -5$) with most of the weight close to $a = 0.003$ and the HPBF with a prior ($d = 1,\ k = -1.1$) with most of the weight close to $b = 0.1$. Also shown is HPBF with a prior ($d = 0.05,\ k = -1.1$) that is close to uniform. The filtering was carried out on 1000 datasets simulated using the LD structure of the *CASP8* region for two scenarios with a single causal SNP that has an OR of 1.1 and a MAF of 0.08, but different sample sizes.

puts some weight at higher values of W and can be seen to provide an effective way of dealing with the uncertainty in W. The new BFs can be thought of as a weighted average of the BFs over the support of W and so shouldn't suffer to the same degree as using WBF with a value of W that doesn't provide much support for the information in the likelihood. The hybrid prior with $d = 1$ and $k = -1.1$ puts most of the weight close to $b = 0.1$ and, as expected, produces a ROC curve similar to WBF with $W = 0.1$. The other hybrid prior which has been used here ($d = 0.05$ and $k = -1.1$) was specifically chosen because it gives an approximate uniform prior over the support of W representing the case where an expert believes that all the values of W in the support are approximately equally likely. For this prior, the ROC curve has an AUC approximately halfway between those of the two ROC curves obtained when W is at the ends of the support. The eqivalent ROC curves for the same scenario but using a smaller sample size of 4000 are given in Figure 3.8(b). It can be seen that, although filtering is generally less effective with such a small sample size, the same relative efficacies of the methods apply.

Restrictions on the uses of different priors

We have concentrated on the results for the PPBF and HPBF methods, as practically these appear to be the most appropriate. Unlike the families of hyperparametric priors, the reciprocal prior cannot be varied, other than to adjust the support of W. Therefore, it is unlikely to provide the best fit to an expert's prior beliefs and it is unsurprising that it does not provide the ROC curve with the largest AUC, or the highest median rank for the causal SNP for the particular scenarios we have considered.

We have also not provided any results for filtering using EPBF. This is because we discovered that there are computational difficulties. In particular, when $\widehat{\beta_1}$ of a SNP is small, we are not able to produce a value for the EPBF. This is due to the part of the expression which is a generalised inverse Gaussian density. One of the parameters for this distribution is $\widehat{\beta_1}^2$ and the density cannot be computed in **R** if this parameter is close to 0. In fact, it typically cannot be computed if $\widehat{\beta_1} < 0.01$, which is likely to apply to a large proportion of the SNPs in a fine-mapping study, possibly including the causal SNP.

There are also some computational limitations when the HPBF method is used, although these are far less likely to cause a significant problem. We illustrate the combinations of hyperparameters and $\widehat{\beta_1}$ and V values for which there may be problems in Table 3.6, in which 'A' indicates no problems calculating HPBF and 'D' indicates that HPBF cannot be calculated at all. Where, in the table, a 'B' appears under a combination of hyperparameters, HPBF can be calculated unless both $\widehat{\beta_1}$ is large (approximately ≥ 1) and V is small (approximately ≤ 0.1). Where 'C' appears, HPBF cannot be calculated using these hyperparameters if V is small (approximately ≤ 0.1). This problem is caused by both the numerator and the denominator of the HPBF being so close to 0 that **R** processes them as if they were both 0.

3.4 Incorporating external functional data

We chose to use Bayes Factors as an analysis method due to the fact that functional information can be incorporated into the analysis through the priors. So far we have concentrated on how to specify the prior distribution for the effect size, and the options considered have not included functional data. The most intuitive way to include functional information is into the prior probability of association, δ. If we can find a suitable way of doing this, functional information will be incorporated along with the information from the genotype data into the key filtering statistic, the posterior probability of association (Δ), through

	$d =$						
$k =$	≤ 3	4	5	8	10	15	20
-1.1	A	B	B	C	C	D	D
-2	A	B	B	C	C	C	D
-5	A	A	A	B	C	C	D
-10	A	A	A	A	B	C	C
-20	A	A	A	A	A	B	C

Table 3.6: Limitations of the HPBF, as indicated for different combinations of the hyperparameters. 'A' indicates no limitations; 'B' indicates that HPBF can only be calculated for SNPs which do not have both a large $\widehat{\beta_1}$ and a small V; 'C' indicates that HPBF can only be calculated for SNPs which do not have small V; 'D' indicates that HPBF cannot be calculated at all using this combination of hyperparameters.

the formula $\Delta/(1-\Delta) = \delta/(1-\delta) \times \text{BF}$. We would hope that if this could be done effectively, filtering results should be improved.

3.4.1 Using elicitation in conjunction with available functional information

The prior probability of association can be seen to be distinct from the prior on the logOR. δ_i is a measure of the belief that SNP i has a causal association with the disease, but does not quantify the size of the association, whereas β_{1i} is a measure of the size of any association, whether it is causal or not (through LD and/or sampling variation). One way of assigning δ values to SNPs is by using functional genetic data. The choice of the δ values is completely subjective and should be based on the types of functional information that the investigator believes are relevant to the disease they are studying.

We have previously mentioned using elicitation with our expert to determine a distribution for β_{1i}. Now we aim to elicit values for δ_i for each SNP, i. Elicitation is the process of working with an expert (on the subject of interest) to formulate a numeric representation of their beliefs about a certain quantity. There are different methods for carrying out elicitation and Garthwaite *et al.* give a thorough review on the topic [21]. They discuss some issues which are important in the context of our work. First of all, we talk about an expert, and in the example that we give there is a single geneticist whose expertise we seek. Alternatively, multiple experts may also be used, but this adds to the complexity by necessitating the combination of multiple opinions into a single prior. This also slightly changes the problem, as a combination prior is not a subjective prior in the same way that a single expert's opinion is, and therefore the resulting value or distribution does not have an intuitive meaning.

It is important to understand, too, that any prior elicited is subjective and there is no "correct" prior in an objective sense. Section 4 of the Garthwaite *et al.* paper deals with the adequacy of elicitation. The section begins with a discussion on how some researchers have defined an expert's "true" probabilities. The authors suggest that a compromise may be "the result of a method that leads the expert to view the problem from as complete and unbiased a perspective as possible through appropriate use of cognitive tools" [21]. Several such tools are then discussed and examples given.

Here we describe one particular method for assigning prior probabilities based on expert knowledge of breast cancer causal variants and using functional information that is publicly available from the ENCODE database [15]. The method could be adapted to assign prior probabilities relevant to other diseases.

The ENCODE database

The Encyclopaedia of DNA Elements (ENCODE) [15] is a huge online database, available to view using the UCSC Genome Browser, containing much of the known functional information about the human genome. A huge number of variables are recorded at the SNP-level, some of which are likely to be related to whether or not a SNP will have a causal effect on a disease.

Generating prior probabilities of association based on functional and expert prior information

We give an example of a method of assigning δ values that combines functional data from the ENCODE database [15] with expert knowledge of a specific disease. Rather than treating each SNP separately, we assign them to a small number of groups and give all the SNPs in each group the same δ value. This is done using the following steps.

Step 1: Choose a subset of the (many available) ENCODE variables, relevant to the disease of interest.

Step 2: If appropriate, group the ENCODE variables into summary variables indicating broader functionality and choose values of the original variables at which to bifurcate the SNPs into "more likely" and "less likely to be causal" subsets for the relevant broad functionality.

Step 3: By determining the relative importance of the summary variables in terms of probability of causality, use them to divide the SNPs into a small number of prior probability groups, ordered from "very unlikely" to "very likely" to be causal.

Step 4: Determine the probability of no causal SNP in the region of interest and the relative probability of a SNP in each of the prior probability groups being causal. Use this information, and the number of SNPs assigned to each group, to determine δ_g, the prior probability of a SNP being causal given that is is assigned to group g, for each value of g.

Once BF_i has been calculated from the genotype data for SNP i, this can be combined with δ_i to calculate Δ_i, which can then be used for filtering.

Using these steps we were able to assign prior probabilities of association to SNPs in the *CASP8* region based on an expert's knowledge of the genetics of breast cancer. The ENCODE variables [15] that we chose for *step 1* and the summary variables we assigned for *step 2* are given in Table 3.7. There is a lot of missing data in the ENCODE variables, with "gene" being the only one we used that didn't have any values missing. A value of 0 was assigned for all missing values in numeric variables as all given values were positive. Each of the variables that make up *Histone modification* have missing values for between 31% and 39% of SNPs, but only 6% are missing values for all three of these. A larger number of values were missing for the variables that make up *Availability* and *Conservation*, between 71% and 95% for the numeric variables and 98% of "OpenChromSynthGm12878Pk" have no indicator. Figure 3.9 shows diagrammatically how we used these summary variables to group all SNPs into 4 classes (*step 3*), depending on the SNP-specific outcomes and the expert's belief about how much each summary variable influences the probability of the SNPs being causal. This resulted in 1698 SNPs in the *CASP8* region being assigned to the "very unlikely to be causal" group 1, 780 to group 2, 362 to group 3 and 31 to the "very likely to be causal" group 4.

For *step 4*, we were able to elicit the expert's belief that there was no causal SNP in the region as "approximately 0.4". Therefore,

$$\prod_{n=1}^{2871}(1-\delta_n) \approx 0.4 \tag{3.28}$$

$$\Rightarrow (1-\delta_{g=1})^{1698}(1-\delta_{g=2})^{780}(1-\delta_{g=3})^{362}(1-\delta_{g=4})^{31} \approx 0.4. \tag{3.29}$$

Using binomial expansions and considering the fact that all δ_g are expected to be very small relative to the exponent, we can further approximate

$$(1-1698\delta_{g=1})(1-780\delta_{g=2})(1-362\delta_{g=3})(1-31\delta_{g=4}) \approx 0.4 \tag{3.30}$$

$$\Rightarrow 1698\delta_{g=1}+780\delta_{g=2}+362\delta_{g=3}+31\delta_{g=4} \approx 0.6. \tag{3.31}$$

To solve this equation we chose what our expert believed to be an appropriate limitation of $\delta_{g=4}=5\delta_{g=3}=5^2\delta_{g=2}=5^3\delta_{g=1}$. This can now easily be solved to

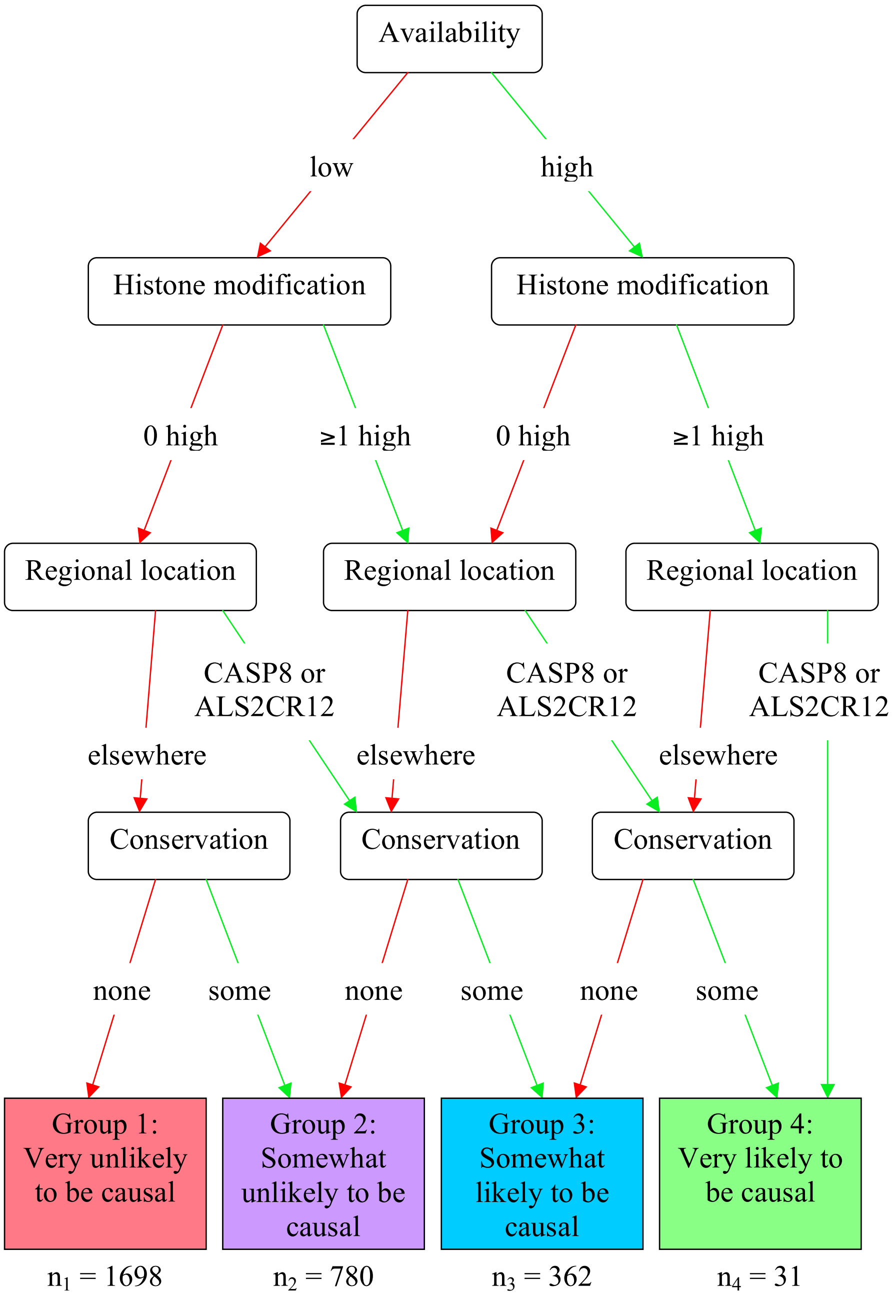


Figure 3.9: Flow diagram showing how SNPs in the *CASP8* region were divided into four groups, depending on four summary variables: *Regional location, Histone modification, Availability* and *Conservation*. The groups represent the subjective belief of a breast cancer geneticist about how likely SNPs are to be causal.

Summary variable	Values of ENCODE variables for which SNPs are likely to be causal	Values of ENCODE variables for which SNPs are unlikely unlikely to be causal
Regional location	gene given as *CASP8* or *ALS2CR12*	elsewhere
Histone modification	layeredGm12878H3k4me1StdSig ≥ 5 or layeredHmecH3k4me3StdSig $\geq e^{1.5}$ or layeredHmecH3k27acStdSig $\geq e^{1.75}$	otherwise
Availability	TranscriptionGm12878 $\geq e^{1.5}$ or TxnFactorChip ≥ 100 or any indicator for OpenChromSynthGm12878Pk	otherwise
Conservation	Conservation score > 0	otherwise

Table 3.7: Four summary variables to describe the SNPs in the 1Mb region surrounding *CASP8*. These are determined based on the following variables downloaded from the ENCODE database: gene, layeredGm12878H3k4me1StdSig, layeredHmecH3k4me3StdSig, layeredHmecH3k27acStdSig, TranscriptionGm12878, TxnFactorChip, OpenChromSynthGm12878Pk and Conservation. Values given in the table were used to determine how likely SNPs with that description/score are to be causal, compared with other SNPs in the region.

give, approximately, $\delta_{g=1} = 3.2 \times 10^{-5}$, from which we can infer $\delta_{g=2} = 1.6 \times 10^{-4}$, $\delta_{g=3} = 8 \times 10^{-4}$ and $\delta_{g=4} = 4 \times 10^{-3}$.

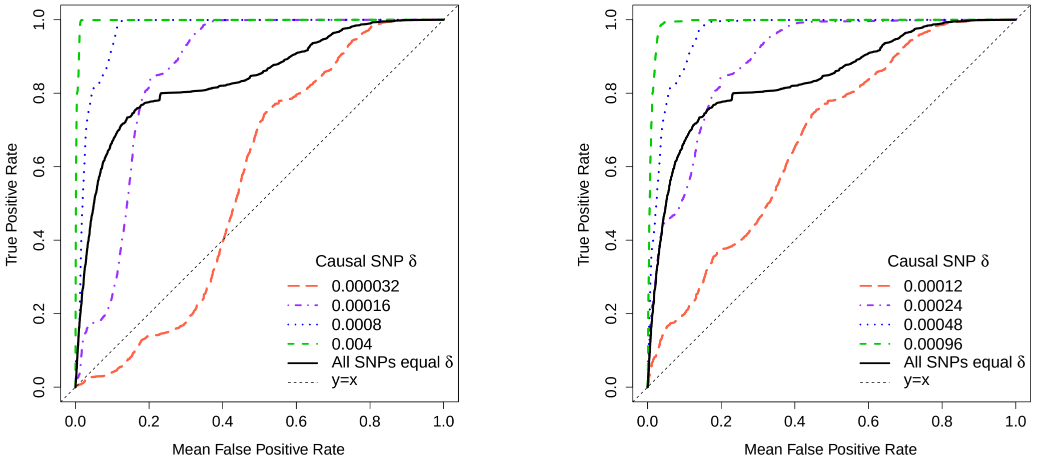
3.4.2 The effect of including prior probabilities of association

Filtering using Δ thresholds was tested on simulated datasets using the δ values assigned to the SNPs as described above. Datasets were simulated to represent 4 scenarios in which the causal SNP had each of the four different prior δ values. All the causal SNPs had a per-allele OR of 1.1 and a similar MAF (in the range 0.037 to 0.049) and were chosen to be in high LD with each other. All pairwise D' values for these four causal SNPs were 1 except for one pair with $D' = 0.916$. However, in this case, the r^2 value was very high ($r^2 = 0.839$). The simulations were analysed using WBF, with a prior on the logOR of $N(0, W_{EB})$ (using the 30% of SNPs with the highest likelihoods to calculate W_{EB}). Using SNPs in such high LD limited the effect of using different causal SNPs so that most of the differences seen would be the result of the different δ values assigned. In fact, we found that the results of filtering using just Bayes Factors to be very similar for each of these four causal SNPs (this is demonstrated in Table 3.8, in the column headed BF filtering).

Figure 3.10(a) shows the results of filtering using posterior probabilities for four scenarios as described. The results of filtering using BF alone for one of the scenarios has also been included for comparison. It can be seen that when the causal SNP was in group 1 and given a very low δ value it was a lot less likely to be retained than when it was assigned any of the other possible δ values. Causal SNPs with the other three δ values also give noticeably different results to each other, but in the case of the scenarios tested here, all resulted in higher TPRs than BF filtering at FPRs ≥ 0.18. If the causal SNP is in group 3 or 4, the TPR > 0.95 at FPRs as small as 0.11. When δ is assigned by group in this way, the filtering results vary dependant on both the relative numbers of SNPs in each group and the precise δ values used.

Some investigators might worry that if their expert has made an incorrect judgement and the causal SNP has in fact been assigned the lowest probability, the possibility of retaining the causal SNP after filtering is reduced significantly. The scenarios for which the results are illustrated in Figure 3.10(a) have causal SNPs with low OR and MAF and a reasonably small sample size for a collaborative fine-mapping study. As these values increase, the information in the data increases and filtering produces better results in general (as demonstrated for LP filtering in §2.3), but this also has the effect of reducing the weight of the priors. If a larger sample size cannot be used and it is suspected that the causal SNP OR and MAF may be very small, or if little is known about the region, investigators may wish to assign δ values to SNPs that are more similar. The prior probabilities could, for example, be assigned according to the limitation $\delta_{g=4} = 2\delta_{g=3} = 2^2\delta_{g=2} = 2^3\delta_{g=1}$, and Figure 3.10(b) shows the results for the same datasets when this is the case. The ROC curves for filtering using Δ are all slightly closer to the BF ROC curve in this figure. When the causal SNP is in group 2, 3 or 4, the ROC curves in Figure 3.10(b) have larger AUCs (89%, 96% and 99%) than the ROC curve for filtering using BF alone (AUC = 84%). However, the AUC is quite a lot smaller when the causal SNP is in group 1 (67%).

We also compared the results of posterior probability filtering to those of BF filtering by examining the numbers of SNPs (both causal and non-causal) retained when the TPR is fixed at 90% for the four scenarios (and two methods of assigning prior probabilities) considered here. These results are given in Table 3.8. This emphasises how much it is possible to reduce the set of candidate causal SNPs (perhaps 28 or fewer) with posterior probability filtering, but only if prior probabilities are assigned appropriately and with confidence. These results also indicate that if SNPs cannot be accurately grouped, for example if



(a) ROC curves for each of the four prior probability scenarios when the values of δ assigned to the SNPs in the four groups were 0.000032, 0.00016, 0.0008 and 0.004. A ROC curve of the results for filtering using BF alone is given for comparison.

(b) ROC curves for each of the four prior probability scenarios when the values of δ assigned to the SNPs in the four groups were 0.00012, 0.00024, 0.00048 and 0.00096. A ROC curve of the results for filtering using BF alone is given for comparison.

Figure 3.10: Effectiveness of posterior probability of association (Δ) as a fine-mapping filter according to the prior probability of association (δ) of the causal SNPs. 1000 datasets were simulated for each of four scenarios using causal SNPs with per-allele OR of 1.1, MAFs close to 0.04 and a total sample size of 20,000 using the LD structure of the *CASP8* region. All SNPs were assigned to one of four prior probability groups and for each scenario a different causal SNP was selected so that it came from each of these groups. A prior on the logOR of $N(0, W_{EB})$ and the Wakefield approximation were used.

Causal SNP group (prior probability)	**BF filtering** mean (s.d.) *threshold*	$\delta_{g=4} = 5\delta_{g=3} = 5^2\delta_{g=2} = 5^3\delta_{g=1}$		$\delta_{g=4} = 2\delta_{g=3} = 2^2\delta_{g=2} = 2^3\delta_{g=1}$	
		Δ filtering mean (s.d.) *threshold*	**Intersection** mean (s.d.)	**Δ filtering** mean (s.d.) *threshold*	**Intersection** mean (s.d.)
Group 1 ($\delta_{g=1}$)	1514 (351) *0.84*	2059 (211) 2.70×10^{-5}	1514 (351)	1926 (284) 1.01×10^{-4}	1514 (351)
Group 2 ($\delta_{g=2}$)	1515 (433) *0.88*	814 (124) 1.41×10^{-4}	638 (182)	823 (156) 2.11×10^{-4}	692 (184)
Group 3 ($\delta_{g=3}$)	1701 (457) *0.83*	264 (61) 6.63×10^{-4}	255 (65)	315 (76) 3.98×10^{-4}	309 (78)
Group 4 ($\delta_{g=4}$)	1678 (441) *0.83*	28 (12) 3.32×10^{-3}	28 (12)	62 (42) 7.97×10^{-3}	62 (42)

Table 3.8: The numbers of SNPs retained out of the total 2871 in the region, such that the true positive rate (TPR) is 0.9. For four scenarios with similar causal SNPs (each in a different prior probability (δ) group), Bayes factor (BF) filtering was carried out and the results are given in the second column. Group-specific δ values were assigned in two different ways, indicated in the top row. Posterior probability (Δ) filtering was carried out and the results are given for this and for the intersection of SNPs retained using the two different methods of filtering. Results are given as the mean and standard deviation (s.d.) of the numbers of SNPs retained and for the two filtering methods, the BF or Δ threshold required to achieve this TPR is given. For each scenario, 1000 datasets, with a causal SNP with a per-allele OR of 1.1, a MAF of 0.037 to 0.049 and a sample size of 20,000 was simulated using the LD structure of the *CASP8* region. To calculate the BFs, a prior on the logOR of $N(0, W_{EB})$ and the Wakefield approximation were used.

it is not know what functional information is important, then BF alone should be used for filtering.

3.5 Summary of filtering using Bayes factors

Several methods of Bayes Factor (BF) analysis were considered in this chapter. There are two filtering statistics that can be used, BF itself and Δ, the posterior probability of association. This probability is calculated using $\Delta/(1-\Delta) = \delta/(1-\delta) \times \text{BF}$, where δ is the prior probability of association. Filtering using BFs alone is equivalent to assigning all SNPs equal δ, and this method was employed by Maller *et al* [32]. Whichever of these filters is being used, to calculate the BF for a SNP, a prior on the logOR (natural logarithm of the per-allele odds ratio) must be specified. We have examined in detail several methods of calculating BF using different priors on the logOR, and considered the scenarios for which they are most appropriate and efficacious.

A simple method of calculating the BF is to use the Wakefield BF (WBF) approximation [59] [60] which requires the prior logOR distribution to be of the form $N(0, W)$. A normal distribution centred around 0 appears to be appropriate, but the choice of the variance, W, is still problematic. If a suitable value of W is used (assigning the same value to all SNPs), then filtering using these BFs produces higher TPRs at particular FPRs than likelihood percentile (LP) filtering, which we found to be the most efficacious method tested in Chapter 2. However, we found the results of filtering to be sensitive to the choice of W.

Collaboration with experts on the genetic region of interest is key when carrying out BF analysis. If a lot is known and it is possible to confidently elicit a single value of W, this can be used. However, if less is known, it is still possible to choose appropriate priors for the logOR based on $N(0, W)$. For example, little may be known about the effect size in the region being investigated, but an expert may be confident that there is a relationship between the causal SNP MAFs and effect sizes in relation to the disease of interest. We have further developed a method published by Wakefield [60] of assigning SNP-specific $W(\text{MAF})$ values based on such a relationship, which will improve the results of filtering if the investigator's beliefs hold for the particular causal SNP in that region. If very little is known, an empirical W value for all SNPs can be assigned with the aim of maximising the BF for the causal SNP, but not for all SNPs. This is done using $W_{EB} = \max(\widehat{\beta_1}^2 - V, 10^{-12})$ which often results in a higher TPR for a particular FPR than most values of W.

Whether a single W value is elicited from an expert or chosen empirically using W_{EB}, or SNP-specific W values are assigned according to MAF, all these priors are applied using the WBF. However, we have also developed several new forms of approximate Bayes factor which can be used as an alternative to WBF. These are applicable if an expert has some ideas about the prior logOR, but these are not consistent with a single W value. Our method allows for the calculation of BFs where $N(0, W)$ is believed to be an appropriate form for the prior, but where there is uncertainty in W, by allowing W to vary according to some prior distribution. The priors on W that we have developed include three parametric families and one fixed form.

If other prior information is available, it can be incorporated through δ and filtering carried out using Δ. In particular, this work was motivated by the large quantities of SNP level functional data now freely available online. Incorporating external data such as these could be a way of countering the problems encountered by fine-mapping studies including high levels of short range LD, enabling smaller samples to produce results with suitable power. A common methodology to determine between the large number of SNPs in a region is to use the results of GWAS and then systematically examine functionality databases to justify the top hits. We have attempted to formalise the incorporation of the functional information within the analysis through using it to specify δ values and combining these with BFs. Our illustration clearly shows that the power of the study can be increased so long as prior probabilities are appropriately assigned to SNPs and we have given an example of how this could be carried out using expert knowledge to select appropriate functional variables from the ENCODE database [15].

Chapter 4

An illustrative example: the iCOGS data

4.1 The association between the *CASP8* region and breast cancer

To demonstrate the use of some of the methods described in Chapters 2 and 3, we have applied the most appropriate to genotype data from the Collaborative Oncological Gene-environment Study (COGS). Specifically, the data analysed in this chapter are the genotypes and imputed genotype doses of the SNPs in the *CASP8* region, which include 501 genotyped SNPs, chosen by the Breast Cancer Association Consortium (BCAC), as well as 1232 imputed SNPs. Although the study recorded the cases/control status of subjects with respect to several cancers, the analyses we have carried out all concern the association of SNPs with breast cancer.

After a borderline association between the D302H variant (rs1045485) in the *CASP8* region and breast cancer was observed in a meta-analysis of 3 studies, Cox *et al.* [16] included it in a candidate variant study with a sample of size 33,532. The observed association in this sample had a p-value of 1.1×10^{-7} and a per-allele odds ratio (OR) and 95% confidence interval (CI) of 1.14 (1.09, 1.19) (with the major allele conveying the increase in risk). A further candidate variant study for this region was carried out in a Korean population by Han *et al.* [25]. The population size was 3135, and the variant rs1861270 (5-UTR C $>$ T) was considered to be associated with breast cancer with a p-value of 0.02. The OR and 95% CI for the one and two risk allele genotypes were 1.13 (0.95, 1.34) and 1.48 (1.04, 2.10), respectively. The same variant in the *CASP8* region that had been included in the Cox *et al.* [16] study was considered in another candidate variant study by Palanca Suela *et al.* [39], this time in relation to modifying the risk of breast cancer in carriers of the known high risk mutations in *BRCA1* and *BRCA2*. This study had a small sample size of 390, and the p-value was 0.01. The investigators considered this to be a significant association in this sample of the sub-population, with an OR of 3.41 and 95% CI (1.33, 8.78) (again, the major allele was the risk allele). Camp *et al.* [14] carried out a fine-mapping study on a smaller scale than COGS (3888 subjects), but on the same region. Haplotype analysis uncovered the most significant association (as measured using p-value), a three-SNP haplotype, with a dominant risk ratio and 95% CI of 1.28 (1.21, 1.35).

Associations have also been found between variants in this region and other types of cancer. A GWAS carried out by Barrett *et al.* [11] also found an association between variant rs13016963 in this region and the risk of melanoma, which had a significant p-value of 8.6×10^{-10}. Another GWAS by Berndt *et al.* [12] investigated association with chronic lymphocytic leukemia. A variant which is classified as either in *CASP8* or *CASP10* (rs3769825) was found to be associated with an OR and 95% CI of 1.19 (1.12, 1.25) and a p-value of 2.50×10^{-9}.

With mounting evidence suggesting there is a breast cancer causal variant in the *CASP8* region, BCAC chose to include this region in their large scale (89,050 subjects) fine-mapping study with the aim of refining this signal. More information about the COGS study is given in §1.3.3, with further details available on the website [2] and in the main study paper, by Michailidou *et al.* [35].

4.1.1 Preliminary analyses

The analyses in this chapter are supplementary to the main analysis carried out on this data by the COGS *CASP8* fine-mapping research group. This analysis is included in a currently pre-publication paper entitled *Identification and fine-mapping of novel associations in the CASP8 region on chromosome 2 with Breast Cancer risk* by Lin *et al.* Details are given in this paper about the study populations, although we only consider the European subjects in this project. There is also information about ethical approval, the selection of SNPs for inclusion on the genotyping chip, the quality control measures taken, and how missing genotypes and further SNPs were imputed.

Logistic regression models were fitted, each including one SNP. COGS was a collaboration between many study groups from around the world, so a group identifier and 7 principal components for ancestry were included as covariates in the models. Information about the principal component analysis is given in Michailidou *et al.* [35]. SNPs were chosen to be taken forward to *in-silico* bioinformatic analysis in Lin *et al.* (pre-publication) using both relative likelihood (RL) filtering and r^2 filtering. The intersection of these two methods was used, such that SNPs would only be retained for further analysis if they had both RL $< 1/100$ and $r^2 > 0.4$ with SNP$_{max}$. In Chapter 2, we found that these two methods were less effective than several other filtering methods, so we tested those methods of filtering we believed to be most appropriate based on the results of previous chapters. However, the logistic regression analysis carried out by Wei-Yu Lin was used as the basis for all further analyses in this described in this chapter.

4.2 Fine-mapping the *CASP8* region using genotype data only

Initially we analysed the genotype data alone. In Chapter 2, the most effective method of this kind was found to be likelihood percentile (LP) filtering. Although this was closely followed by p-value filtering, we would not recommend choosing SNPs using p-value, as we have demonstrated the improved efficacy when LP is used. In the investigation of Bayes factor (BF) filtering methods which was carried out in Chapter 3, we also developed a way to use the Wakefield Bayes factor (WBF) using only information from the data, by using the empirical Bayes prior on the log OR $N(0, W_{EB})$. Using LP filtering may appeal to investigators who are not comfortable with Bayesian or empirical Bayesian methodology, and generally it is a very efficacious method. However, in §3.2.1, we showed that compared to LP filtering, when using WBF, higher true positive rates (TPRs) could be achieved at low false positive rates (FPRs). It is these low FPRs we are interested in, but this advantage only occurs with certain values of W. It is likely that W_{EB} would be such a value, and therefore we have decided to carry out this analysis. Using BFs in this way does have certain other advantages. For example, they can be be easily combined with other information at a later date, as demonstrated by Knight *et al.* [29]. We have also included some of the results obtained by carrying out LP and p-value filtering for comparison.

The empirical Bayesian method for using WBF filtering places a prior on the logOR of the form $N(0, W_{EB})$, and requires the calculation of $W_{EB} = \max(\widehat{\beta_1}^2 - V, 10^{-12})$, using appropriate values of $\widehat{\beta_1}^2$ and V in order to try and maximise the Bayes factor for the causal SNP. The investigation we carried out previously using simulations of this same region suggested that the median of the set of V for all SNPs and the median of the set of $\widehat{\beta_1}^2$ for the top 30% of SNPs ranked by likelihood were appropriate. In the case of this dataset, this resulted in a value of $W_{EB} \approx 0.0440^2 - 0.0002 \approx 0.0018$.

The top 40 SNPs (2.3%) ranked by WBF using W_{EB} are given in Table 4.1. For each of these SNPs we give the SNP number (the 1733 SNPs included in the investigation are numbered by the order of their chromosomal position). We also give the MLE of the per-allele OR from the fitted logistic regression models, along with a 95% confidence interval (CI) and the MAF of the SNP within the whole study population. We have used superscripts to indicate the SNPs for which the major allele conveyed the increase in disease risk and those SNPs that were fully imputed. The SNPs are given in the table in order of their ranks based on the BFs, and these BFs are given. The ranks based on both likelihood and p-value are also included.

It can be seen that the ranks are quite similar for these three methods. In particular, the ranks based on likelihood and p-value are identical for all but one of the SNPs given in this table (SNP number 438, which is ranked 38th by likelihood and 39th by p-value). If we limit the proportion of SNPs to take forward to the next stage of analysis to the top ranked 10% (173 SNPs), the same SNPs would be taken forward whether likelihood or p-value was used as the filter. Using WBF, 170 of these SNPs are also the same as those taken forward based on the other methods. To retain the top 10% of SNPs, a WBF filter threshold 7.7 is applied. It can easily be deduced that if an LP filter were used, the filter which results in 10% retention has a 90% threshold, but if this is translated to a relative likelihood (RL) filtering threshold, it is 1/460, and only 48 SNPs (2.8%) would be retained if an RL threshold of 1/100 was applied.

For this dataset, if p-value filtering were used, a threshold of 0.075 would have to be applied to retain 10% of SNPs. It may also be of interest that the highest ranked SNP had a p-value of 1.08×10^{-5} and that if a p-value threshold of 10^{-4} were applied to this data, 9 SNPs would be retained. If a p-value threshold of 10^{-3} were applied, this number would increase to 41.

4.3 Fine-mapping the *CASP8* region using Bayes factors

Throughout this project we worked with a breast cancer geneticist who had worked on several studies of the *CASP8* region. They had valuable knowledge of the association signal in this region, as well as about breast cancer causal variants across the genome. We were therefore able to include some of their prior knowledge to specify prior distributions, and used these to calculate Bayes factors. We were able to carry out analyses using both WBF with SNP-specific $W(\text{MAF})$ values, and the most appropriate of the novel BF approximations described in §3.3.

We described the methodology for determining parameters for our modified $W(\text{MAF})$ formula

$$W(M) = \alpha_0 + \alpha_1 \exp(\alpha_2 + \alpha_3 \times M). \tag{4.1}$$

in §3.2.2, and showed that with the values elicited from our expert, this gave us

$$W(M) = 0.0123 + 0.172 \exp(-0.451 - 20.3M). \tag{4.2}$$

SNP number	OR (95% CI)	MAF	WBF with W_{EB} = 0.0018	Ranking		
				likelihood	p-value	WBF (W_{EB} = 0.0018)
980 [b]	1.048 (1.027, 1.071)	0.294	1932	1	1	1
1027	1.046 (1.024, 1.068)	0.285	955	2	2	2
992 [b]	1.045 (1.022, 1.067)	0.287	488	3	3	3
909	1.043 (1.021, 1.065)	0.287	352	9	9	4
838	1.041 (1.020, 1.062)	0.338	330	5	5	5
950 [b]	1.043 (1.021, 1.065)	0.286	326	10	10	6
960 [b]	1.043 (1.021, 1.065)	0.285	320	7	7	=7
961 [b]	1.043 (1.021, 1.065)	0.285	320	8	8	=7
985 [b]	1.043 (1.021, 1.066)	0.286	310	4	4	9
837	1.042 (1.021, 1.064)	0.299	306	6	6	10
907	1.042 (1.020, 1.064)	0.287	255	11	11	11
896	1.042 (1.020, 1.064)	0.287	254	13	13	=12
912	1.042 (1.020, 1.064)	0.287	254	15	15	=12
956 [a,b]	1.052 (1.025, 1.078)	0.170	210	16	16	14
1272 [a]	1.075 (1.036, 1.116)	0.071	190	14	14	15
885	1.041 (1.019, 1.063)	0.287	184	19	19	16
878 [a]	1.081 (1.039, 1.125)	0.061	177	12	12	17
1004 [a,b]	1.051 (1.024, 1.078)	0.173	167	18	18	18
999 [a,b]	1.050 (1.023, 1.077)	0.174	152	20	20	19
955 [a,b]	1.050 (1.023, 1.078)	0.173	152	21	21	20
993 [a,b]	1.049 (1.023, 1.076)	0.176	140	22	22	21
681 [a]	1.074 (1.035, 1.116)	0.069	133	17	17	22
994 [a,b]	1.049 (1.022, 1.076)	0.178	124	24	24	23
945 [a,b]	1.049 (1.022, 1.076)	0.176	124	25	25	24
958 [a,b]	1.048 (1.022, 1.075)	0.181	120	26	26	25
924 [a,b]	1.048 (1.021, 1.075)	0.176	109	27	27	26
928 [a,b]	1.048 (1.021, 1.074)	0.176	106	28	28	27
602 [a,b]	1.074 (1.033, 1.116)	0.072	101	23	23	28
1036 [a]	1.049 (1.021, 1.077)	0.143	62	30	30	29
438 [b]	1.045 (1.018, 1.072)	0.176	49	38	39	30
656 [a]	1.059 (1.024, 1.096)	0.087	45	33	33	31
1029 [a,b]	1.057 (1.023, 1.092)	0.100	44	35	35	32
523 [a,b]	1.044 (1.018, 1.072)	0.164	39	44	44	33
971 [a,b]	1.083 (1.034, 1.135)	0.049	37	31	31	34
840 [b]	1.035 (1.014, 1.057)	0.313	35	45	45	35
933 [a,b]	1.083 (1.034, 1.135)	0.049	35	32	32	36
903 [a,b]	1.081 (1.033, 1.132)	0.047	34	34	34	37
844	1.036 (1.014, 1.058)	0.269	34	47	47	38
901 [a]	1.080 (1.032, 1.131)	0.046	33	36	36	39
947 [a,b]	1.082 (1.033, 1.134)	0.049	31	37	37	40

[a] For these SNPs, the major allele is associated with a higher disease risk.

[b] These SNPs were not genotyped but imputed.

Table 4.1: Top ranked SNPs in *CASP8* region based on Wakefield Bayes factor (WBF) approximation with empirical prior logOR $\sim N(0, W_{EB} = 0.0018)$. Rankings using likelihood and p-value are also included, as is the logistic regression estimate and 95% confidence interval (CI) of the odds ratio (OR) for each SNP. The genotype data for *CASP8* region comes from the iCOGS study and has a total sample size of 89,050 and 1733 SNPs.

In §3.3.3, we also showed how to choose the best fitting prior for one of the novel BF methods, such that W is allowed to vary in way that closely fits an expert's beliefs. We also demonstrated this with values elicited from our expert. With this methodology, we use the median of V from the data in the fitting of the distribution. The median V from the iCOGS dataset was approximately 1.74×10^{-4}, and although all the other values were kept the same as previously, this resulted in a very slightly different prior. Again, a power prior form had the closest fit, but with hyperparameter $k = -1.66$.

Although Equation (4.2) describes our expert's beliefs about the relationship between MAF and effect size for breast cancer causal variants in general, they were unsure whether this pattern would apply in this region, so although we have carried out both types of analyses, it is the PPBF results we focus on, and would use for the selection of SNPs to take forward for further analysis. The results for the top 40 SNPs ranked by PPBF are given in Table 4.2, along with ranks based on WBF filtering using both $W(\text{MAF})$ and W_{EB} for comparison. In total there are 23 SNPs with BF > 100 and 113 with BF > 10.

Although the rankings for the methods in Table 4.2 are not so close as those shown in Table 4.1, there is still a lot of similarity, particular between WBF with $W(\text{MAF})$ and PPBF. The most noticeable deviation from similarity is SNP number 1098, which is ranked 27th by both of these methods, but only 76th by WBF with W_{EB}. Applying a filter to retain 10% of SNPs (173) would result in 168 of the same SNPs whichever of these methods were used. The BF thresholds which would need to be applied at this level are 4.85 for PPBF and 3.3 for WBF with $W(\text{MAF})$.

Even though the prior logORs for all these methods have the form $N(0, W)$, it would be expected that the results would differ because of the different choices for W. However, in a sample size as large as this, the BF estimates are more highly weighted by the data than the prior. Perhaps it also should not be surprising that PPBF produces similar ranks to WBF using $W(\text{MAF})$, because we showed in §3.3.2 that the power prior was consistent with rarer alleles generally having larger effects. In fact, this can be seen by plotting the estimated ORs against the sample MAF. Figure 4.1(a) is such a plot created using all 1733 SNPs, and it can be seen that SNPs with MAFs anywhere in the possible range, $0 < \text{MAF} \leq 0.5$, have estimated ORs with a variety of values. Any relationship between MAF and OR for the *CASP8* SNPs is unclear, except that only rare SNPs achieve the highest ORs (for example, the only SNPs with OR > 1.08 have MAF < 0.07). However, as can be seen in Figure 4.1(b), if only the top 10% of SNPs, as ranked by PPBF, are include in the plot, there is a clear relationship between MAF and OR.

SNP number	OR (95% CI)	MAF	PPBF	Ranking		
				WBF with $W =$		PPBF
				W(MAF)	0.0018	
980 [b]	1.048 (1.027, 1.071)	0.294	1387	1	1	1
1027	1.046 (1.024, 1.068)	0.285	664	2	2	2
992 [b]	1.045 (1.022, 1.067)	0.287	334	3	3	3
909	1.043 (1.021, 1.065)	0.287	234	4	4	4
878 [a]	1.081 (1.039, 1.125)	0.061	228	11	17	5
1272 [a]	1.075 (1.036, 1.116)	0.071	217	12	15	6
950 [b]	1.043 (1.021, 1.065)	0.286	217	5	6	7
838	1.041 (1.020, 1.062)	0.338	213	6	5	8
960 [b]	1.043 (1.021, 1.065)	0.285	213	7	=7	=9
961 [b]	1.043 (1.021, 1.065)	0.285	213	8	=7	=9
985 [b]	1.043 (1.021, 1.066)	0.286	206	9	9	11
837	1.042 (1.021, 1.064)	0.299	200	10	10	12
907	1.042 (1.020, 1.064)	0.287	167	13	11	13
896	1.042 (1.020, 1.064)	0.287	166	14	=12	=14
912	1.042 (1.020, 1.064)	0.287	166	15	=12	=14
956 [a,b]	1.052 (1.025, 1.080)	0.170	159	16	14	16
681 [a]	1.074 (1.035, 1.116)	0.069	149	19	22	17
1004 [a,b]	1.051 (1.024, 1.078)	0.173	124	18	18	18
885	1.041 (1.019, 1.063)	0.287	118	17	16	19
955 [a,b]	1.050 (1.023, 1.078)	0.173	112	21	20	20
999 [a,b]	1.050 (1.023, 1.077)	0.174	112	20	19	21
602 [a,b]	1.074 (1.033, 1.116)	0.072	112	24	28	22
993 [a,b]	1.049 (1.023, 1.076)	0.176	102	22	21	23
994 [a,b]	1.049 (1.022, 1.076)	0.178	90	23	23	24
945 [a,b]	1.049 (1.022, 1.076)	0.176	89	25	24	25
958 [a,b]	1.048 (1.022, 1.075)	0.181	85	26	25	26
1098 [a]	1.148 (1.064, 1.238)	0.016	79	27	76	27
924 [a,b]	1.048 (1.021, 1.075)	0.176	77	28	26	28
928 [a,b]	1.048 (1.021, 1.074)	0.176	75	29	27	29
971 [a,b]	1.083 (1.034, 1.135)	0.049	46	31	34	30
1036 [a]	1.049 (1.021, 1.077)	0.143	45	30	29	31
933 [a,b]	1.083 (1.034, 1.135)	0.049	43	33	36	32
903 [a,b]	1.081 (1.033, 1.132)	0.047	42	34	37	33
901 [a]	1.080 (1.032, 1.131)	0.046	40	38	39	34
947 [a,b]	1.082 (1.033, 1.134)	0.049	39	37	40	35
656 [a]	1.059 (1.024, 1.096)	0.087	38	35	31	36
940 [a,b]	1.082 (1.033, 1.133)	0.049	38	39	42	37
931 [a,b]	1.082 (1.032, 1.133)	0.049	37	40	44	38
888 [a,b]	1.081 (1.032, 1.132)	0.047	36	41	43	39
1029 [a,b]	1.057 (1.023, 1.092)	0.100	36	36	32	40

[a] For these SNPs, the major allele is associated with a higher disease risk.

[b] These SNPs were not genotyped but imputed.

Table 4.2: Top ranked SNPs in *CASP8* region based on power prior Bayes factor (PPBF) approximation with hyperparameter $k = -1.66$ and $a = 0.003 \leq W \leq b = 0.1$. Rankings using WBF with W(MAF) and W_{EB} are also included, as is the logistic regression estimate and 95% confidence interval (CI) of the odds ratio (OR) for each SNP. The genotype data for *CASP8* region comes from the iCOGS study and has a total sample size of 89,050 and 1733 SNPs.

Figure 4.1(b) also highlights clusters of SNPs, and if we examine results from Haploview [10], we see that these are SNPs that are in strong LD. This effect can also be seen in Table 4.2. For example, the top hit has OR and 95% CI of 1.048 (1.027, 1.071) and MAF of 0.294. In total, there are 12 other SNPs in the top 10% with $1.04 < \text{OR} < 1.05$ and $0.28 < \text{MAF} < 0.30$, and these are all ranked in the top 19. Eleven of these SNPs have $r^2 > 0.9$ with the top hit, and the twelfth (number 837, the only one with a MAF greater than that of the top hit) has $r^2 > 0.7$. It is likely that, if any, only one of these 13 SNPs is causal, with the others being associated only through LD. Another such cluster which is clear in both the table and figure are those SNPs with $1.045 < \text{OR} < 1.055$ and $0.170 < \text{MAF} < 0.185$. There are 10 such SNPs ranked between 16th and 29th by PPBF. In the case of this group, it is always the major allele which is associated with the increase in risk. These SNPs all have $0.07 < r^2 < 0.10$ with the top hit and $r^2 > 0.8$ with each other.

4.4 Fine-mapping the *CASP8* region using prior probabilities

In §3.4.1, we demonstrated the use of a combination of expert elicitation and information from the ENCODE database to determine prior probabilities of association (δ) for each of the SNPs. We used the same summary variables described in this section to divide the SNPs into four prior probability groups. However, the set of SNPs in this study were different to those used in the simulations, and the total number in each group (from those least likely to those most likely to be causal) were 994, 497, 227 and 15. To find the appropriate values of δ, a solution to the following approximation must be found in order to satisfy the expert's beliefs about the probability of there being no causal SNP in the region:

$$\prod_{n=1}^{1733}(1-\delta_n) \approx 0.4, \tag{4.3}$$

which is approximately equivalent to

$$994\delta_{g=1} + 497\delta_{g=2} + 227\delta_{g=3} + 15\delta_{g=4} \approx 0.6. \tag{4.4}$$

Applying the limitation of $\delta_{g=4} = 5\delta_{g=3} = 5^2\delta_{g=2} = 5^3\delta_{g=1}$ to this approximation gives $\delta_{g=1} = 5.44 \times 10^{-5}$, $\delta_{g=2} = 2.72 \times 10^{-4}$, $\delta_{g=3} = 1.36 \times 10^{-3}$ and $\delta_{g=4} = 6.8 \times 10^{-3}$.

In §4.3, we explained that of the different BF approximations, we are most confident about those calculated using PPBF. We therefore calculated the posterior probability (Δ) for SNP i using

$$\Delta_i/(1-\Delta_i) = \delta_i/(1-\delta_i) \times \mathrm{PPBF}_i. \tag{4.5}$$

The posterior probabilities and the ranks based on them are given for the 40 SNPs with the highest Δ in Table 4.3, and are signified by 'Δ (5)'. In this table we also provide the PPBF values and ranks, as well as the Δ values and ranks obtained by assigning less extreme δ values to each group ($\delta_{g=4} = 2\delta_{g=3} = 2^2\delta_{g=2} = 2^3\delta_{g=1} \Rightarrow \delta_{g=1} = 2 \times 10^{-4}$, $\delta_{g=2} = 4 \times 10^{-4}$, $\delta_{g=3} = 8 \times 10^{-4}$ and $\delta_{g=4} = 1.6 \times 10^{-3}$), signified by '$\Delta$ (2)'. The number of the prior probability group that each SNP was assigned to are also included for these SNPs, where the groups are numbered 1 to 4 in order of increasing prior probability.

It can be seen that, whilst all the methods we have examined previously have generally resulted in very similar ranks for the top 40 SNPs, ranking by Δ is quite different. Posterior probability filtering with the δ values elicited from the expert is the only method for which SNP number 980 is not the top hit, due to the fact that it was assigned a prior suggesting it was 'fairly unlikely to be causal' (group 2). However, using this method, the prior odds and the BF have equal weighting in the calculation of the posterior odds, so because it has such a large PPBF, it is still ranked 4th. It can be seen the top two ranked SNPs were assigned the highest prior probability (group 4), but also have very high BFs. However, there are no SNPs in the top 40 that were assigned the smallest prior probability (group 1). When the δ values limited by a factor of 2 are used, the SNP rankings based on Δ generally lie somewhere between the ranking by PPBF and the ranking by Δ calculated using the more extreme prior probabilities. Using these δs, SNP 980 is still the top hit and there is one SNP in the top ranked 40 (36th) that has $\delta_{g=1}$. This is SNP number 602, which is ranked 22nd by PPBF.

Considering the top 10% of ranked SNPs using Δ calculated with the elicited δs, there are only 99 of the 173 which are also in the top 10% as ranked by PPBF. Of these 173 SNPs, thirteen were assigned the highest prior probability, 91 the next highest and 61 and 8 the two lowest probabilities, respectively. These are equivalent to approximately 87%, 40%, 12% and 1% of the total SNPs assigned to each of the prior probability groups. The highest ranked SNP which had the lowest δ (group 1) is again SNP 602, which is ranked only 73rd in this case. The SNP in the top 10% based on these Δs that is ranked lowest by PPBF is SNP number 786, which only has the 1502nd highest PPBF value, but

SNP			PPBF filtering		Δ filtering				
number	OR (95% CI)	MAF	PPBF	rank	Δ (2)	rank	Δ (5)	rank	δ_g
838	1.041 (1.020, 1.062)	0.338	213	8	0.255	3	0.593	1	4
837	1.042 (1.021, 1.064)	0.299	200	12	0.243	4	0.578	2	4
1027	1.046 (1.024, 1.068)	0.285	664	2	0.347	2	0.475	3	3
980 [b]	1.048 (1.027, 1.071)	0.294	1387	1	0.357	1	0.274	4	2
893 [a,b]	1.080 (1.032, 1.131)	0.047	36	41	0.054	19	0.196	5	4
896	1.042 (1.020, 1.064)	0.287	166	=14	0.118	6	0.185	6	3
885	1.041 (1.019, 1.063)	0.287	119	19	0.087	7	0.139	7	3
839 [b]	1.035 (1.013, 1.057)	0.270	16	=62	0.024	35	0.096	8	4
992 [b]	1.045 (1.022, 1.067)	0.287	334	3	0.118	5	0.083	9	2
909	1.043 (1.021, 1.065)	0.287	234	4	0.086	8	0.060	10	2
971 [a,b]	1.083 (1.034, 1.135)	0.049	46	30	0.036	24	0.059	11	3
878 [a]	1.081 (1.039, 1.125)	0.061	228	5	0.084	9	0.058	12	2
1272 [a]	1.075 (1.036, 1.116)	0.071	217	6	0.080	10	0.056	13	2
950 [b]	1.043 (1.021, 1.065)	0.286	217	7	0.078	11	0.056	14	2
960 [b]	1.043 (1.021, 1.065)	0.285	213	=9	0.079	=12	0.055	=15	2
961 [b]	1.043 (1.021, 1.065)	0.285	213	=9	0.079	=12	0.055	=15	2
903 [a,b]	1.081 (1.033, 1.132)	0.047	42	33	0.033	28	0.054	17	3
985 [b]	1.043 (1.021, 1.066)	0.286	206	11	0.077	14	0.053	18	2
901 [a]	1.080 (1.032, 1.131)	0.046	40	34	0.031	29	0.051	19	3
888 [a,b]	1.081 (1.032, 1.132)	0.047	36	39	0.020	33	0.047	20	3
907	1.042 (1.020, 1.064)	0.287	167	13	0.063	15	0.043	21	2
912	1.042 (1.020, 1.064)	0.287	166	=14	0.062	16	0.043	22	2
956 [a,b]	1.052 (1.025, 1.080)	0.170	159	16	0.060	17	0.042	23	2
883 [a,b]	1.080 (1.031, 1.131)	0.047	31	44	0.025	34	0.041	24	3
681 [a]	1.074 (1.035, 1.116)	0.069	149	17	0.056	18	0.039	25	2
1004 [a,b]	1.051 (1.024, 1.078)	0.173	124	18	0.047	20	0.033	26	2
890	1.028 (1.008, 1.048)	0.386	5	171	0.008	59	0.032	27	4
862 [a,b]	1.076 (1.028, 1.125)	0.048	24	47	0.019	37	0.031	28	3
955 [a,b]	1.050 (1.023, 1.078)	0.173	112	20	0.043	21	0.030	29	2
999 [a,b]	1.050 (1.023, 1.077)	0.174	112	21	0.043	22	0.030	30	2
840 [b]	1.035 (1.014, 1.057)	0.313	21	50	0.016	40	0.028	31	3
844	1.036 (1.014, 1.058)	0.269	20	51	0.016	41	0.027	32	3
993 [a,b]	1.049 (1.023, 1.076)	0.176	102	23	0.039	23	0.027	33	2
994 [a,b]	1.049 (1.022, 1.076)	0.178	90	24	0.035	25	0.024	34	2
945 [a,b]	1.049 (1.022, 1.076)	0.176	89	25	0.034	26	0.024	35	2
958 [a,b]	1.048 (1.022, 1.075)	0.181	85	26	0.033	27	0.023	36	2
1098 [a]	1.148 (1.064, 1.238)	0.016	79	27	0.031	30	0.021	37	2
843	1.035 (1.013, 1.058)	0.269	15	64	0.012	48	0.021	38	3
924 [a,b]	1.048 (1.021, 1.075)	0.176	77	28	0.030	31	0.021	39	2
928 [a,b]	1.048 (1.021, 1.074)	0.176	75	29	0.029	32	0.020	40	2

[a] For these SNPs, the major allele is associated with a higher disease risk.

[b] These SNPs were not genotyped but imputed.

Table 4.3: Top ranked SNPs in *CASP8* region based on posterior probability (Δ (5)) filtering. These values are calculated by combining prior probabilities ($\delta_{g=4} = 5\delta_{g=3} = 5^2\delta_{g=2} = 5^3\delta_{g=1}$) with power prior Bayes factors (PPBF) with hyperparameter $k = -1.66$ and $a = 0.003 \leq W \leq b = 0.1$. Values of PPBF and Δ (2) using more similar δ and the ranks based on these are also included, as are the estimated OR and MAF for each SNP. The genotype data for *CASP8* region comes from the iCOGS study (89,050 subjects and 1733 SNPs).

due to having been assigned the highest prior probability, it is ranked 168th by posterior probability.

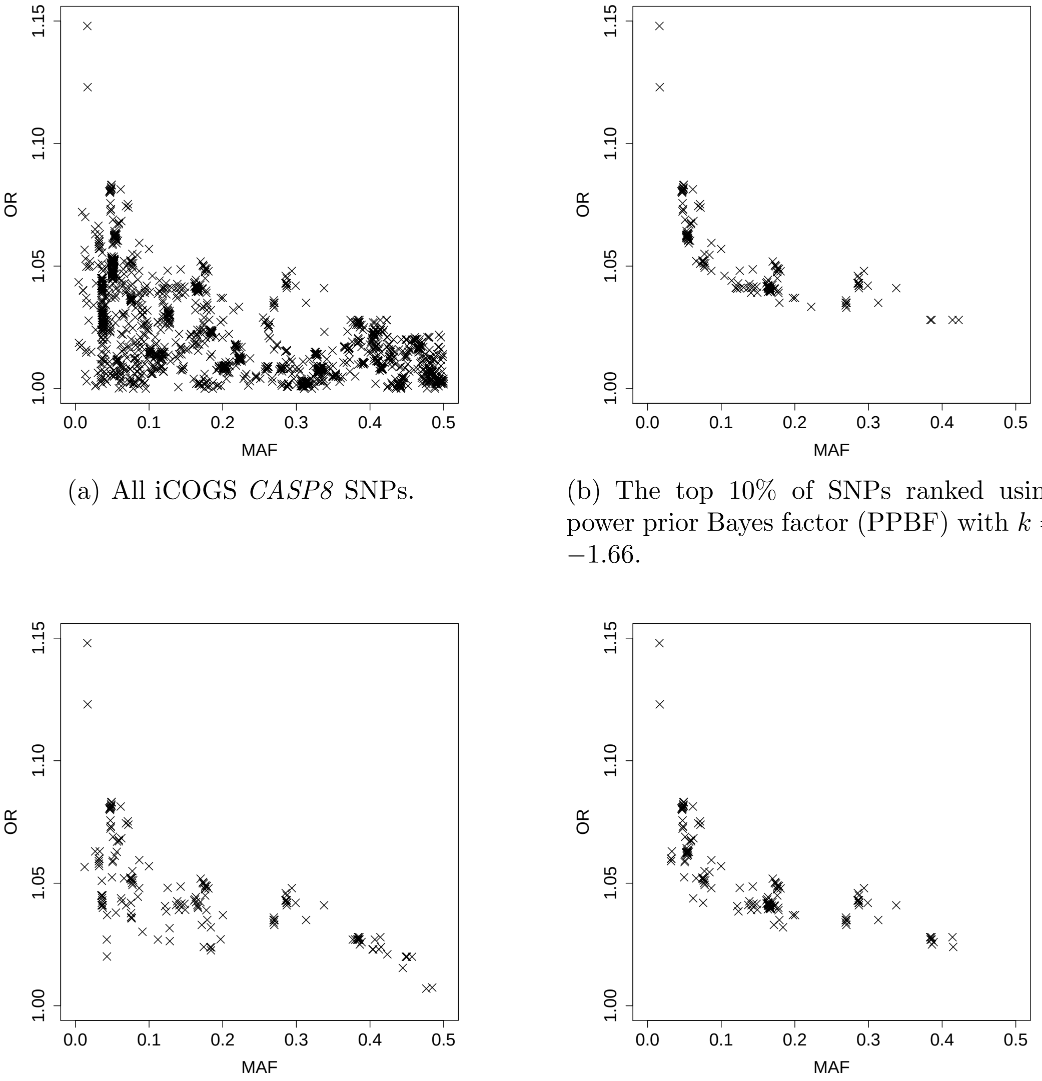
We now compare these to the top 10% when Δ is calculated with the more conservative δ values. Based on these probabilities, there are 155 SNPs in the top 173 as ranked by PPBF. The numbers of these 173 SNPs assigned to each of the prior probability groups (from highest prior probability to lowest, respectively) are 7, 48, 54 and 64, which are approximately 47%, 21%, 11% and 6% of the total SNPs in each of the groups. If the top 10% of SNPs were taken forward using these Δs, the SNP that would be included that has the lowest rank by PPBF is SNP number 798. This is ranked 341st by PPBF value, but 118th by this formulation of Δ.

Figures 4.1(c) and 4.1(d) plot the estimated OR against the sample MAF for the top 10% of SNPs ranked using the two different sets of Δ values. Comparing these to the top 10% ranked by PPBF, as plotted in Figure 4.1(b), we observe that using the more conservative δ values produces a similar plot, because of the high amount of overlap in these two groups, but there are far more SNPs which are not clustered close to other SNPs in Figures 4.1(c). This suggests that including other functional data through prior probabilities of association can help to break up the large groups of SNPs in high LD with each other, particularly when dissimilar prior probabilities are assigned to the SNPs.

4.5 Analysis of a subset of the *CASP8* data

COGS was a collaborative study with a very large number of subjects. In §2.3.3, we demonstrated how the efficacy of likelihood percentile filtering increases with sample size. The same effect can also be shown for the other methods of analysis, as larger studies have more power to distinguish between causal and coincidental associations, for example, those attributed to SNPs in high LD with the causal SNP. With such a large sample size, the Bayesian methods we have used which incorporate external information through prior distributions have been very highly weighted by the likelihood compared to the prior. Therefore, all the methods have produced reasonably similar rankings for most of the SNPs. Many studies will not have such a large sample size as this, so we now give an example of the same analyses used on a stratified random subset of the iCOGS subjects, with 2721 cases and 2517 controls (5238 total).

The results given in Table 4.4 include the top 40 SNPs ranked by PPBF, which due to a different median V across the SNPs, had a different hyperparameter in this analysis, of $k = -1.96$. Also given for these SNPs are the ranks by likelihood, p-value, WBF with $\beta_1 \sim N(0, W(\text{MAF}))$, WBF with



(a) All iCOGS *CASP8* SNPs.

(b) The top 10% of SNPs ranked using power prior Bayes factor (PPBF) with $k = -1.66$.

(c) The top 10% of SNPs ranked using posterior probabilities of association, calculated using PPBF with $k = -1.66$ and $\delta_{g=1} = 5.44 \times 10^{-5}$, $\delta_{g=2} = 2.72 \times 10^{-4}$, $\delta_{g=3} = 1.36 \times 10^{-3}$ and $\delta_{g=4} = 6.8 \times 10^{-3}$.

(d) The top 10% of SNPs ranked using posterior probabilities of association, calculated using PPBF with $k = -1.66$ and $\delta_{g=1} = 2 \times 10^{-4}$, $\delta_{g=2} = 4 \times 10^{-4}$, $\delta_{g=3} = 8 \times 10^{-4}$ and $\delta_{g=4} = 1.6 \times 10^{-3}$.

Figure 4.1: Estimated OR plotted against sample MAF for SNPs in the iCOGS *CASP8* fine-mapping study

$\beta_1 \sim N(0, W_{EB})$, where $W_{EB} = 0.0126$, and Δ, calculated using PPBF and $\delta_{g=1} = 5.44 \times 10^{-5}$, $\delta_{g=2} = 2.72 \times 10^{-4}$, $\delta_{g=3} = 1.36 \times 10^{-3}, \delta_{g=4} = 6.8 \times 10^{-3}$. We see that in general, in a sample this size, these methods do not have the power to detect the small effect sizes that were observed in the full sample size of 89,050. However, there is also less agreement in the ranks of these SNPs than we observed with the top 40 ranked SNPs using the larger dataset. Previously, we observed that likelihood and p-value gave almost identical rankings and these were also similar to the WBF rankings using W_{EB}. Although the ranks based on these methods are still close, there is much more divergence than previously.

It can be seen that, as would be expected, ranks based on BF with different priors are much more variable for this data, as the priors have more influence on the BF values in a smaller sample such as this. Interestingly, the WBF rankings using $W(\text{MAF})$ are similar to the likelihood and p-value ranks, but they vary somewhat from those based on WBF using W_{EB} and PPBF. For example the SNP ranked 1^{st} using the empirical Bayes method is only ranked 10^{th} by PPBF and 22^{nd} using WBF priors based on MAF. Even in the large sample size, we observed that the single method which generally resulted in the largest deviations in ranks from the other methods was Δ filtering. However, all the SNPs ranked 1 to 21 by PPBF were also ranked in the top 30 by Δ. In this sample size, the SNP ranked 19^{th} by PPBF is ranked 196^{th} by Δ and 12 SNPs ranked in the top 40 by PPBF are ranked between 200 and 300 by Δ.

			Ranking					
SNP	OR (95% CI)	MAF	likeli-	p-	WBF with $W =$		PPBF	Δ
number			hood	value	$W(M)$	W_{EB}		
822 [b]	1.514 (1.215, 1.886)	0.037	1	1	1	3	1	1
807 [b]	1.520 (1.216, 1.900)	0.036	2	2	2	6	2	2
820 [b]	1.515 (1.213, 1.893)	0.036	3	3	3	7	3	3
824 [b]	1.514 (1.212, 1.891)	0.036	4	4	4	8	4	4
868 [b]	1.508 (1.209, 1.881)	0.038	5	5	5	9	5	26
378 [b]	1.431 (1.174, 1.745)	0.046	7	7	7	4	6	5
858 [b]	1.495 (1.198, 1.866)	0.036	6	6	6	10	7	30
379 [b]	1.409 (1.162, 1.709)	0.047	8	8	8	5	8	6
854	1.470 (1.181, 1.829)	0.036	9	9	9	11	9	7
346	1.262 (1.099, 1.449)	0.093	26	22	22	1	10	37
845 [b]	1.469 (1.180, 1.829)	0.037	11	11	10	12	11	8
879 [b]	1.480 (1.184, 1.851)	0.037	10	10	11	14	12	39
823 [b]	1.480 (1.183, 1.851)	0.036	12	12	12	15	13	9
339 [b]	1.266 (1.099, 1.459)	0.091	30	28	27	2	14	44
705 [b]	1.430 (1.161, 1.761)	0.043	15	15	15	13	15	50
752 [b]	1.449 (1.168, 1.798)	0.039	14	14	14	16	16	51
900 [b]	1.475 (1.177, 1.849)	0.036	13	13	13	21	17	11
698 [b]	1.454 (1.167, 1.812)	0.036	16	16	16	22	18	57
699 [b]	1.454 (1.167, 1.812)	0.036	17	17	17	23	19	196
700 [b]	1.432 (1.159, 1.771)	0.038	20	20	19	17	20	58
701 [b]	1.453 (1.166, 1.811)	0.036	18	18	18	24	21	12
704 [b]	1.452 (1.165, 1.810)	0.036	19	19	20	26	22	13
694 [b]	1.454 (1.165, 1.814)	0.036	21	21	21	27	23	208
697	1.427 (1.155, 1.763)	0.038	=23	=24	=23	=19	=24	211
702	1.427 (1.155, 1.763)	0.038	=23	=24	=23	=19	=24	14
693 [b]	1.451 (1.163, 1.811)	0.036	22	23	25	31	26	225
690 [b]	1.443 (1.160, 1.796)	0.038	25	26	26	30	27	226
691 [b]	1.425 (1.153, 1.761)	0.039	27	27	28	25	28	232
687 [b]	1.422 (1.151, 1.757)	0.039	32	32	30	28	29	15
692 [b]	1.446 (1.159, 1.804)	0.036	28	29	29	42	30	251
771	1.419 (1.149, 1.753)	0.038	35	35	33	29	31	75
688 [b]	1.439 (1.156, 1.791)	0.037	31	31	31	40	32	16
706 [b]	1.439 (1.156, 1.792)	0.036	33	33	34	43	33	254
756 [b]	1.433 (1.153, 1.780)	0.038	36	36	36	36	34	256
707 [b]	1.433 (1.154, 1.781)	0.037	34	34	35	38	35	257
689 [b]	1.430 (1.152, 1.775)	0.038	37	37	37	35	36	17
746 [b]	1.425 (1.150, 1.765)	0.038	38	38	39	=33	=37	=259
747 [b]	1.425 (1.150, 1.765)	0.038	39	39	38	=33	=37	=259
1005 [b]	1.466 (1.165, 1.845)	0.035	29	30	32	55	39	79
710 [b]	1.414 (1.145, 1.746)	0.038	41	41	40	32	40	264

[b] These SNPs were not genotyped but imputed.

Table 4.4: SNP rankings using different filters on a subset of the genotype data for *CASP8* region from the iCOGS study, which has a total sample size of 5238 and 1733 SNPs. The top 40 SNPs ranked by power prior Bayes factors (PPBF; $k = -1.66$, $a = 0.003 \le W \le b = 0.1$) are given in the table. Estimated ORs, MAF and ranks based on likelihood, p-value, Wakefield Bayes factor (WBF) with W based on both MAF ($W(M)$) and empirical Bayes estimation (W_{EB}) and posterior probability (Δ) using PPBF are also included.

Chapter 5

Discussion

5.1 Scope of the project

This project was an investigation of different statistics that could be used to filter the set of candidate causal SNPs in a known disease-association region, following dense genotyping of the region. Each of these statistics can be calculated individually for every variant and they do not take into account any interactions between variants.

We investigated the relative efficacy of a number of different methods, including some based on likelihoods, p-values, genetic structure and Bayes factors. We chose to use Bayes factors with a prior distribution on the logOR of the form $N(0, W)$, but considered several different methods of choosing W, and when these methods might be appropriate. We also described one method of choosing SNP-specific prior probabilities of association, based on functional genetic data and expert knowledge.

We compared the use of different methods using simulated fine-mapping data and took into account the possibility that much of it may be imputed. To allow for variation, many simulations of each causal SNP scenario were generated, and results of only a selection of the scenarios analysed are given in the thesis. Multiple other scenarios were examined to ensure that overall trends were still observed when other combinations of causal SNP effect, MAF, genomic location and sample size were simulated.

A thorough analysis of real fine-mapped genotype data was also carried out. The data came from the Collaborative Oncological Gene-environment Study and was examining the association between the *CASP8* region on chromosome 2 and breast cancer. We focused on analysing the data using power prior Bayes factors (PPBF), as we believed this to be the most appropriate method, but also compared the results to those from several different analyses.

5.2 Recommendations

5.2.1 Choosing a filtering statistic

We would recommend the use of Bayes factors (BF) in statistical analyses such as these. The prior distribution of the SNP logOR can be chosen to represent the knowledge of the genetic association being fine-mapped, whether that knowledge is specific or vague. They also have the advantage that they can be combined with any other relevant information, through either prior odds of association, or other BFs [29]. Methods which combine the genotype BFs with external information through other BFs allows for the possibility of doing this at a later date, if new information comes to light. If an investigator is not comfortable with Bayesian methodology, we would recommend they use the likelihood percentile method of filtering, as the analyses of simulated data showed that this was consistently the most efficacious method which does not require the specification of any priors.

If it is decided to proceed with a BF analysis, the prior on the logOR (β_1) should be chosen carefully. We advocate collaboration with one or more experts on the genetics of the disease or, if possible, the region of interest. Elicitation can be employed with such experts to determine their prior beliefs (this should be done before they see the data), and prior distributions can be formed based on this elicited information. All the distributions on β_1 we have considered in this project take the general form $N(0, W)$, but software such as Snptest [33] allows for the calculation of BFs with other prior distributions, so these may be used if they appear to fit the beliefs of an expert more closely. Throughout Chapter 3, we made suggestions on what could be elicited to determine either a fixed value of W for all SNPs, SNP-specific values based on their MAF, or a variable W, dependent on a hyperparametric distribution. In the latter case, we also provide code in Appendix C to determine the most appropriate values for the hyperparameters. Initially, though, a judgement should be made on which prior form is the most appropriate. If the expert is very confident about the effect size in the fine-mapped region, a fixed value of W is probably best. If they have some idea, but are not confident, or multiple values elicited from them are not consistent with a single W, a variable W would be better. Most of the priors we suggest for W put most of the weight of W at the lower end of the support but the exponential prior allows for an almost uniform prior which is useful when an expert believes a range of values of W are equally likely a priori. The hybrid prior can be specified so that the mode is anywhere in the given range. This might be useful when an expert has a strong prior belief in a particular value of W but wants to allow for some uncertainty in it.

If the expert knows little about the association in the region of interest, but has prior knowledge about a relationship between the effect sizes and MAFs of causal SNPs for the disease of interest, this can be used to choose $W(\text{MAF})$. If the expert considers the hypothesis that rarer alleles have larger effects to be appropriate and also does have some prior knowledge of the region of interest, the power and reciprocal priors for the novel Bayes factor approximations would be appropriate, because in these cases, the expectation of W appears universally to be an increasing function of V. Depending on the values of the hyperparameters, $\mathbb{E}(W)$ may be a decreasing function of W when the hybrid prior is employed. Such a prior would be inappropriate to use if it was believed that rare causal SNPs do indeed have larger effect sizes. To represent both this belief and a prior on the variance of the logOR that cannot be represented by the power or reciprocal priors, we suggest a Savage-Dickey density ratio approach [56].

In the case where no such expert is available, BF analysis can still be used, and we would recommend using the empirical Bayes method of choosing W, as described in 3.2.3. The choice of W_{EB} is based on the fact that using $W = \widehat{\beta_1}^2 - V$ for any particular SNP maximises the BF for that SNP. However, we only wish to maximise the BF for the causal SNP, so we suggest using the median $\widehat{\beta_1}$ across the top p% of SNPs (ranked by likelihood) and the median V across all SNPs in the region. In the analysis of simulations of the *CASP8* region, we found that $p = 30$ generally gave ROC curves with close to the highest AUCs, but this may not be the case for other regions. The encouraging results of analysis using W_{EB} occur because it is data-driven, but this deviates from purist Bayesian theory as it is not chosen prior to data analysis. Three of the prior forms for the novel Bayes factors also depend upon the genotype data through V, so may not be considered appropriate by purist Bayesians.

Elicitation

Elicitation can be a difficult process, but it is important to obtain the most appropriate values to describe numerically the beliefs of the experts and their certainty in their beliefs. The value of feedback in the elicitation process is worth emphasizing. Take, for example, elicitation to determine the hyperparmeters for a distribution on W, when one of the novel BF approximations is to be employed. Once a distribution for W has been determined based on the quantiles elicited from the expert, it is important to relay back to them what this means about other quantiles not elicited to check that these are acceptable. A web-based tool, MATCH, which may help with this purpose is now available [37]. We have also included **R** code in Appendix C which can be used to determine the most appropriate hyperparameteric values from elicited OR values.

5.2.2 Choosing a filter threshold

In fine-mapping studies, investigators will need to choose an appropriate filter threshold to apply, guided by either the true positive rate (TPR) that they wish to achieve or the false positive rate (FPR) that they are willing to tolerate. We have shown that the relationship between these outcomes will vary dependent on several quantities including the sample size and the MAF and OR of the causal SNP. We suggest estimating the OR by fitting the univariate logistic regression models required for this analysis and then using the median values of the fitted ORs across the top 30% of SNPs, ranked by likelihood, as we showed this worked well in the empirical Bayes method of choosing a prior logOR. Care should be taken to ensure that ORs are adjusted so that they are all greater than 1 (the reciprocal should be taken if the fitted value is less than 1), as it is the magnitude and not the direction of the effect that is of importance in this case. There is also no simple way to estimate the MAF of the causal SNP, but that of the the highest ranked SNP may act as a guide.

For this particular problem, where there is only a single causal SNP among many, the TPR and FPR equate to the probability of retaining the true causal SNP and the approximate proportion of SNPs retained, respectively. These are competing outcomes, which increase together (non-linearly), and it can be difficult to decide what levels of these quantities are acceptable. However, this must be done to justify the choice of a corresponding filter threshold. Bayesian decision theory could be used to help deal with these competing quantities, and a method of this kind has been developed by Wakefield [58]. However, this method still requires the specification of a ratio of the cost of false non-discovery to the cost of false discovery. It is likely that many investigators would find it difficult to confidently quantify such a value.

5.2.3 Choosing a sample size when designing a study

Causal SNP MAF crucially affects the efficacy of LP filtering as we show in Figure 2.5(c), and even with sample sizes as large as 50,000, LP and BF filtering may have a high probability of failing to capture the causal SNP if it has a MAF less than 0.05. Analysing simulated data based on the region of interest and the likely range of causal SNP MAFs and ORs should inform the appropriate sample size required. Such simulation results should also inform appropriate filtering thresholds, taking into account the trade-off that is incurred by using a more lenient threshold: higher TPR but at a cost of higher FPR, as can be seen in Figure 2.5(b).

The appropriate sample sizes to use with RL has been investigated in detail by Udler *et al.* [54], and in §2.3.3 we showed that sample size has a large impact on the efficacy of fine-mapping filters. The Udler paper [54] included the development of an online tool to calculate the required sample size to achieve any given power to ‘discriminate between highly correlated SNPs’. We used this tool to discover that if the causal SNP had a MAF of 0.12 and OR of 1.12 and was in LD at $r^2 = 0.4$ with SNP$_{max}$, a sample size of 46,000 would be required to achieve a power of 0.9 when filtering at a threshold of RL=1/100. However, for a similar scenario, but with the causal SNP in LD at $r^2 = 0.7$ with SNP$_{max}$, a sample size of 92,000 would be required to achieve the same power, due to the increase in difficulty to differentiate between the two SNPs when they are in such high LD.

5.3 Limitations

5.3.1 Computational restrictions of novel Bayes factor approximations

All novel BF approximations can be calculated in **R** (we give code in Appendix B), although the EPBF is computationally intensive and cannot produce results for SNPs that have a small MLE of the logOR, which is likely to be a large proportion of SNPs in a fine-mapping study. Although we also showed in §3.3.4 that there are some combinations of hyperparameters and $\widehat{\beta_1}$ and V for which HPBF also cannot be computed, these are generally extreme combinations, and will rarely occur. The computation for the other forms is simple and efficient. If initial investigation suggests an EPBF should be used, we recommend using PPBF or HPBF instead. In most cases hyperparameters can be found which result in power or hybrid priors very similar to the desired exponential prior. The RPBF is very limited due to there being only a single reciprocal prior form with no hyperparameters and this prior is unlikely to be an appropriate replacement for the desired exponential prior. We would suggest that the EPBF only be used if the investigator particularly does not want to include any information from the data in the prior, and then to proceed with caution. It is the only one of the four prior distribution forms for W which possesses this property.

5.3.2 Incomplete functional data

To carry out Δ filtering, we assigned SNP-specific δ values based on functional SNP-level data found in ENCODE [15]. However, this data is currently limited as information is not complete for all the SNPs across the genome. This means that as well as any uncertainty around the specific values of prior probabilities, there is likely to also be uncertainty about what to do with SNPs for which there is some functional information missing. Of the four summary variables combining the results of several ENCODE variables that we used, *Availability* and *Conservation* in particular had a large amount of missing information. We dealt with missing values in the ENCODE data by replacing them with zeros, but it is unclear how appropriate this is. For some variables, this may not give a good representation of the true missing values. It may be more appropriate to impute these values, using the recorded values for SNPs in close proximity to inform the imputed value for a SNP.

More complete sources of data are expected to come along in the near future and large databases such as ENCODE are regularly updated. We hope that such methods as those described here will become more relevant as these data emerge.

5.4 The project in the context of current genetic analysis

5.4.1 Multiple causal SNP scenarios

This investigation has been restricted to regions of the genome with a single causal SNP. In many cases it will not be clear whether one or more SNPs will be present in the genomic region being investigated. However, several methods aimed at identifying multiple causal variants in a single region have been published, including some which analyse variants simultaneously. These types of methods have been scrutinised in reviews by Ayers and Cordell [9] and Abraham *et al.* [6] and include penalised and non-penalised regression methods and MCMC routines. One popular penalised logistic regression method is HyperLASSO [27], which was demonstrated to be effective in carrying out the analysis of fine-mapped data to uncover the nature of the association in the *HLA* region with Rheumatoid Arthritis [57]. pi-MASS [23] is a piece of software which implements fully Bayesian analyses through MCMC. Many of these methods can be applied through the PUMA (Penalized Unified Multiple-locus Association) framework, which was used to show that carrying out such multiple SNP analyses may result in higher TPRs for a given FPR than single variant p-value analyses, given there are multiple causal SNPs [26]. One problem with using univariate analysis is that if there is some sort of interaction between causal SNPs, then fitting them in single SNP logistic regression models may not result in high enough likelihoods for LP filtering to be effective, particularly if the marginal effect of a SNP is small.

5.4.2 Alternative methods of including functional data

There is a strong tradition within genetics of making data and results publicly available, and such information can be utilised in fine-mapping analysis to increase the power to detect causal effects. This applies whether there is a single or are multiple causal SNPs present in a region. In Chapter 3 we investigated several ways to include both the prior knowledge of experts and SNP-level data from the Encyclopaedia of DNA Elements [15]. However, there are several other published methods which attempt to include external data, such as p-value weighting [42] and a Bayesian latent variable model (BLVM) [20]. Another Bayesian method is stratified false discovery rates [51] [43].

We also stress that in our example of Δ filtering, the given method for assigning δ values to the SNPs in a region is one example of many possible such ways. An alternative method of grouping is to obtain SNP scores from the RegulomeDB database [13]. These categorical scores are assigned based on the regulatory potential of variants and draw information from multiple sources including ENCODE [15]. In this case, a score is between 1 (for most likely to be causal) and 7 (for least likely), although some of these categories have sub-categories such as 1a and 1b. It would be the decision of the analyst of how to assign prior probabilities to these groups. Rather than grouping, a different strategy is to use some sort of continuous score for SNPs. Several such scoring methods have been published recently, based on a SNP's individual likelihood of affecting disease susceptibility, for example the FS score published by Lee and Shatkay [31]. The FS score has the advantage that it integrates a large amount of data from multiple publicly available data sources. It formally combines scores from a number of bioinformatics tools using weighting based on the "reliability" of these tools to give a score between 0 and 1. These scores can be obtained from the F-SNP database. With a continuous scoring method, a function would have to be defined to derive prior probabilities from the scores. Other sources of functional data are also limited, as we explain in relation to the ENCODE database [15] in §5.3.2.

A slightly different way to integrate functional information into this kind of analysis is to use it to form a Bayes Factor rather than a prior probability [29]. This method is effective because "prior knowledge" can be updated any number of times using BFs. Once a posterior odds of association has been calculated, this can be used as a prior odds and multiplied by another BF to get a new posterior odds. Therefore, beginning initially with all SNPs having equal prior probabilities of association, two separate BFs can be used, one containing the association information from the genotyping, as detailed in this study, and the other containing the functional information. Knight *et al.* give some specific values that may be used for these functional BFs [29].

5.5 In conclusion

The problem that we tried to solve was to find the most effective method of filtering SNPs whilst retaining the causal SNP. There is not a simple answer to this problem, as the most appropriate method is likely to depend on what information is available to the investigators and the reliability of that information. Also, we limited our investigations to scenarios where there was a single casual SNP, but this may not be the case. However, we determined that, in general, the best filters to use are likelihood percentile (LP) and Bayes factor (BF). In this chapter, we have aimed to give guidance on when it may be suitable to use LP and the different BF filters, and the sample sizes and thresholds to use with these filters. We have also outlined the limitations of these filters, which may restrict their application in certain situations.

Genetics and genetic epidemiology are still fast growing research areas and we hope that newly available information will aid further applications of the methods discussed here. For example, with more complete reference data, imputation results are likely to become closer to the truth. In particular, though, there is a high potential for including functional information in genotype analysis as the databases that contain this information continue to grow. This incorporation is made simple through the use of Bayes Factors.

Appendix A: Derivation of Wakefield's Bayes factor approximation

We use the Wakefield Bayes factor approximation (WBF) for all calculations where we assign to the logOR (β_1) a prior distribution of $N(0, W)$ with a fixed value of W. We also used it as a basis for our novel BF approximations. The derivation of the WBF is given below, but this is the reciprocal of the BF approximation that Wakefield himself uses in his papers [58] [59] [60].

$$\begin{aligned}
\text{WBF} =& \frac{P(data|H_1)}{P(data|H_0)} && (5.1)\\
=& \frac{\int\int p(\widehat{\boldsymbol{\beta}_F}, \widehat{\beta_1}|\boldsymbol{\beta}_F, \beta_1)\pi(\boldsymbol{\beta}_F, \beta_1)d\boldsymbol{\beta}_F d\beta_1}{\int p(\widehat{\boldsymbol{\beta}_F}, \widehat{\beta_1}|\boldsymbol{\beta}_F, \beta_1 = 0)\pi(\boldsymbol{\beta}_F)d\boldsymbol{\beta}_F} && (5.2)\\
=& \frac{\int p(\widehat{\boldsymbol{\beta}_F}|\boldsymbol{\beta}_F)\pi(\boldsymbol{\beta}_F)d\boldsymbol{\beta}_F \int p(\widehat{\beta_1}|\beta_1)\pi(\beta_1)d\beta_1}{\int p(\widehat{\boldsymbol{\beta}_F}|\boldsymbol{\beta}_F)\pi(\boldsymbol{\beta}_F)d\boldsymbol{\beta}_F \times p(\widehat{\beta_1}|\beta_1 = 0)} && (5.3)\\
=& \frac{\int p(\widehat{\beta_1}|\beta_1)\pi(\beta_1)d\beta_1}{p(\widehat{\beta_1}|\beta_1 = 0)}, && (5.4)
\end{aligned}$$

where H_0 and H_1 are the null and alternative hypotheses; β_1 is the logOR; $\boldsymbol{\beta}_F = \boldsymbol{\beta}_C + \frac{I_{01}}{I_{00}}\beta_1$, where $\boldsymbol{\beta}_C$ is a vector of all other logistic regression coefficients and I is the information matrix; and $\pi(.)$ is the prior over all parameters. This is simplified by considering a prior of $\beta_1 \sim N(0, W)$, and the fact that, asymptotically, as the sample size increases, $\widehat{\beta_1} \sim N(\beta_1, V)$. Consider first the numerator:

$$\int p(\widehat{\beta}_1|\beta_1)\pi(\beta_1)d\beta_1 = \int \frac{1}{\sqrt{2\pi V}} exp\left(-\frac{(\widehat{\beta}_1-\beta_1)^2}{2V}\right)\frac{1}{\sqrt{2\pi W}} exp\left(-\frac{\beta_1^2}{2W}\right)d\beta_1 \tag{5.5}$$

$$= \frac{1}{\sqrt{2\pi(V+W)}} exp\left(-\frac{\widehat{\beta_1}^2}{2V}+\frac{\widehat{\beta_1}^2 W}{2V(V+W)}\right) \times \int \sqrt{\frac{V+W}{2\pi VW}} exp\left(-\frac{(\widehat{\beta}_1 W(V+W)^{-1}-\beta_1)^2}{2VW(V+W)^{-1}}\right)d\beta_1. \tag{5.6}$$

Here, the integrand is the pdf of a normal distribution, and integrates to 1. Hence the WBF can be written as:

$$\text{WBF} = \frac{1}{\sqrt{2\pi(V+W)}} exp\left(-\frac{\widehat{\beta_1}^2}{2V}+\frac{\widehat{\beta_1}^2 W}{2V(V+W)}\right) \div \frac{1}{\sqrt{2\pi V}} exp\left(-\frac{\widehat{\beta_1}^2}{2V}\right) \tag{5.7}$$

$$= \sqrt{\frac{V}{V+W}} exp\left(\frac{\widehat{\beta_1}^2 W}{2V(V+W)}\right). \tag{5.8}$$

Appendix B: R code to calculate the new Bayes factors

The **R** code given below will calculate a vector of approximate Bayes factors for a set of SNPs which have been genotyped and analysed using single SNP logistic regression models. These models should all include the same relevant covariates. The first two inputs are `betas` and `vars`, which are both vectors of length n, where n is the number of SNPs in the genotyping study. Respectively, they should be the fitted logOR estimates ($\widehat{\beta_1}$) and their variances (V) from the logistic regression models. The other inputs are `form`, which can be either `"PPBF"`, `"EPBF"`, `"HPBF"` or `"RPBF"` indicating the form of approximate BF to be used; `hyper`, indicating the values of the hyperparameters, which is a single value if `form="PPBF"` or `"EPBF"`, a vector of length 2 (`c(d,k)`) if `form="HPBF"` and null if `form="RPBF"`; `a` and `b`, the limits of the range ($a < b$) over which W should be defined.

```
BFapprox<-function(betas,vars,form,hyper,a=0.003,b=0.1){
nSNP<-length(betas)
betasq<-betas^ 2
#define Q, the denominator
Q=((2*pi*vars)^ -0.5)*exp(-betasq/(2*vars))
####################################################
switch(form,
#Power prior form
PPBF={
k=hyper
#normalising constant for prior
if(k!=-1){
q<-(k+1)/((vars+b)^ (k+1)-(vars+a)^ (k+1)) }
if(k==-1){
q<-(log(vars+b)-log(vars+a))^ (-1) }
#BF
num<-gamma(-k-0.5)*(pgamma(betasq/(2*(vars+a)), -k-0.5)-pgamma(
betasq/(2*(vars+b)), -k-0.5))
   denom<-sqrt(2*pi)*(betasq/2)^ (-k-0.5)
   PPBF<-(q*num)/(Q*denom)
   return(PPBF)},
   #Exponential prior form
   EPBF={
   c=hyper
   #normalising constant for prior
   r=c/(2*(exp(-c*a/2)-exp(-c*b/2)))
   #integrand
   library("GeneralizedHyperbolic")
   int<-rep(NA,nSNP)
   for(i in 1:nSNP){
   if(abs(betas[i])>=0.01){
   intA<-pgig(a+vars[i], param = c(betasq[i],c,1/2))
   intB<-pgig(b+vars[i], param = c(betasq[i],c,1/2))
   int[i]<-intB-intA }}
   #BF
   num<-2*exp(c*vars/2)*besselK(sqrt(c*betasq), 0.5, expon.scaled =
FALSE)
   denom<-sqrt(2*pi)*(c/betasq)^ 0.25
   EPBF<-(r*num*int)/(Q*denom)
   return(EPBF)},
   #Hybrid prior form
   HPBF={
   d=hyper[1]
   k=hyper[2]
   #normalising constant for prior
   inc.gamma.part<-gamma(-k-1)*(pgamma(d/(2*(vars+a)), -k-1)-pgamma
(d/(2*(vars+b)), -k-1))
   s<-(d/2)^ (-k-1)/inc.gamma.part
   #BF
   num<-gamma(-k-0.5)*(pgamma((betasq+d)/(2*(vars+a)), -k-0.5)-pgamma
((betasq+d)/(2*(vars+b)), -k-0.5))
   denom<-sqrt(2*pi)*((betasq+d)/2)^ (-k-0.5)
   HPBF<-(s*num)/(Q*denom)
   return(HPBF)},
   #Reciprocal prior form
   RPBF={
```

```
#approximation of sum for incomplete gamma with 0 term:
abfun<-function(n){
((-1)^ n)*(((b+vars)/2)^ n-((a+vars)/2)^ n)/(n*factorial(n)) }
absum<-0
for(i in 1:1000){
absum<-absum+abfun(i)}
#normalising constant for prior
t<-1/(log((b+vars)/(a+vars))+absum)
#integrand
ap<-pnorm(sqrt(a+vars)-abs(betas)/sqrt(a+vars))
am<-pnorm(-sqrt(a+vars)-abs(betas)/sqrt(a+vars))
bp<-pnorm(sqrt(b+vars)-abs(betas)/sqrt(b+vars))
bm<-pnorm(-sqrt(b+vars)-abs(betas)/sqrt(b+vars))
int<-bp-ap+(bm-am)*exp(2*abs(betas))
#BF
RPBF<-(t*exp(-abs(betas))*int)/(Q*abs(betas))
return(RPBF)},
)}
```

Appendix C: R code to find $f(W)$ that best fits an expert's beliefs

The **R** code given below will, for the power, exponential or hybrid forms of the prior distrbution on W, find the values of the hyperparameters that best fit an expert's beliefs. The main elicited values which are used as inputs are q and PIu, which should be vectors of equal length and for which $\texttt{PIu}_i$ should be the upper limit of the $\texttt{q}_i$% centralised probability interval for the logOR. We also suggest that a and b, the limits of the range ($a < b$) over which W should be defined are elicited from an expert. For the power and hybrid priors, the data-specific value of the variance of the logOR for which you wish to find the best-fitting hyperparameters should be specified as V, and we suggest using the median of all Vs from the dataset. Finally, form can be either "PPBF", "EPBF" or "HPBF" indicating the form of approximate BF to be used. The output is a vector of length two when form="PPBF" or "EPBF", with the first value being the value of the hyperparameter (k or c) which results in the smallest sum of square distances and in the case where form="HPBF", the output is a vector of length three, with the first values being the two hyperparameters (d followed by k). The final value is always the sum of squared differences that the given hyperparameters produce, and we provide this value so that different forms can be fitted and the best fit discovered as being the one which results in the smallest value.

```
priorfit <- function(q, PIu, a=1e-6, b=0.1, V, form) {
#calculate percentiles, quantiles of normal dist and W values
p=1-(1-0.01*q)/2
quant=qnorm(p)
W=(log(PIu)/quant)ˆ 2
switch(form,
#Power prior
PPBF={
#define space over which to search
ks=seq(-10,-0.5,0.01)
```

```
  fullk=rep(NA,length(ks))
  for(j in 1:length(ks)){
  k<-ks[j]
  sum=0
  for(i in 1:length(W)){
  sum=sum+((((V+W[i])^ (k+1)-(V+a)^ (k+1))/((V+b)^ (k+1)-(V+a)^ (k+
1))-p[i])^ 2
  }
  fullk[j]=sum
  }
  #add k=-1 at end
  sum=0
  for(i in 1:length(W)){
  sum=sum+((log(V+W[i])-log(V+a))/(log(V+b)-log(V+a))-p[i])^ 2
  }
  min1k=sum
  ks=cbind(ks,-1)
  fullk=c(fullk,min1k)
  #find k for which the argument is lowest
  mink=which.min(fullk)
  return(c(ks[mink],summary(fullk)[1]))},
  #Exponential prior
  EPBF={
  #define space over which to search
  cs=seq(0,1000,1)
  fullc=rep(NA,length(cs))
  for(j in 1:length(cs)){
  c<-cs[j]
  sum=0
  for(i in 1:length(W)){
  sum=sum+((exp(-c*W[i]/2)-exp(-c*a/2))/(exp(-c*b/2)-exp(-c*a/2))-p[i])^ 2
  }
  fullc[j]=sum
  }
  #find c for which the argument is lowest
  minc=which.min(fullc)
  return(c(cs[minc],summary(fullc)[1]))},
  #Hybrid prior
  HPBF={
```

```
  #define the upper incomplete gamma function
  inc.gam<-function(gama,gamx){
  pgamma(gamx, gama, lower=FALSE) * gamma(gama)}
  #define space over which to search
  ds=c(seq(0,1,0.001),seq(1,2,0.01),seq(2,5,0.1))
  ks=c(seq(-10,-6,0.5),seq(-6,-4,0.1),seq(-4,-1,0.01))
  fulldk=matrix(NA,length(ds),length(ks))
  for(j in 1:length(ds)){
  d<-ds[j]
  for(h in 1:length(ks)){
  k<-ks[h]
  sum=0
  for(i in 1:length(W)){
  sum=sum+((inc.gam(-k-1,d/(2*(V+W[i])))-inc.gam(-k-1,d/(2*(V+a))))/
(inc.gam(-k-1,d/(2*(V+b)))-inc.gam(-k-1,d/(2*(V+a))))-p[i])^ 2
  }
  fulldk[j,h]=sum
  }
  }
  #find d, k combo for which the argument is lowest
  mindk=which.min(fulldk)
  col=floor(mindk/dim(fulldk)[1])
  mink<-col+1
  mind<-mindk-dim(fulldk)[1]*col
  return(c(ds[mind],ks[mink],fulldk[mind,mink]))}
  )}
```

Bibliography

[1] BCAC website. http://www.srl.cam.ac.uk/consortia/bcac/.

[2] COGS website. http://cogseu.org/.

[3] dbSNP. http://www.ncbi.nlm.nih.gov/projects/SNP/.

[4] Human Genome Project. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.

[5] Iceberg computing. http://www.sheffield.ac.uk/wrgrid/iceberg.

[6] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology*, 37(2):184–195, 2013.

[7] I. Adrianto, S. F. Wang, G. B. Wiley, C. J. Lessard, J. A. Kelly, A. J. Adler, S. B. Glenn, A. H. Williams, J. T. Ziegler, M. E. Comeau, M. C. Marion, B. E. Wakeland, C. Y. Liang, K. M. Kaufman, J. M. Guthridge, M. E. Alarcon-Riquelme, G. S. Alarcon, J. M. Anaya, S. C. Bae, J. H. Kim, Y. B. Joo, S. A. Boackle, E. E. Brown, M. A. Petri, R. Ramsey-Goldman, J. D. Reveille, L. M. Vila, L. A. Criswell, J. C. Edberg, B. I. Freedman, G. S. Gilkeson, C. O. Jacob, J. A. James, D. L. Kamen, R. P. Kimberly, J. Martin, J. T. Merrill, T. B. Niewold, B. A. Pons-Estel, R. H. Scofield, A. M. Stevens, B. P. Tsao, T. J. Vyse, C. D. Langefeld, J. B. Harley, E. K. Wakeland, K. L. Moser, C. G. Montgomery, P. M. Gaffney, Biolupus Network, and Genles Network. Association of two independent functional risk haplotypes in TNIP1 with systemic lupus erythematosus. *Arthritis and Rheumatism*, 64(11):3695–3705, 2012.

[8] D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De la Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. K. Wilson, D. Deiros,

M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. X. Li, M. Jian, G. Li, R. Q. Li, H. Q. Liang, G. Tian, B. Wang, W. Wang, H. M. Yang, X. Q. Zhang, H. S. Zheng, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, D. B. Jaffe, E. Shefler, C. L. Sougnez, N. Gormley, S. Humphray, Z. Kingsbury, P. Koko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte, S. Wang, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[9] K. L. Ayers and H. J. Cordell. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8):879–891, 2010.

[10] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.

[11] J. H. Barrett, M. M. Iles, M. Harland, J. C. Taylor, J. F. Aitken, P. A. Andresen, L. A. Akslen, B. K. Armstrong, M. F. Avril, E. Azizi, B. Bakker, W. Bergman, G. Bianchi-Scarra, B. Bressac-de Paillerets, D. Calista, L. A. Cannon-Albright, E. Corda, A. E. Cust, T. Debniak, D. Duffy, A. M. Dunning, D. F. Easton, E. Friedman, P. Galan, P. Ghiorzo, G. G. Giles, J. Hansson, M. Hocevar, V. Hoiom, J. L. Hopper, C. Ingvar, B. Janssen, M. A. Jenkins, G. Jonsson, R. F. Kefford, G. Landi, M. T. Landi, J. Lang, J. Lubinski, R. Mackie, J. Malvehy, N. G. Martin, A. Molven, G. W. Montgomery, F. A. van Nieuwpoort, S. Novakovic, H. Olsson, L. Pastorino, S. Puig, J. A. Puig-Butille, J. Randerson-Moor, H. Snowden, R. Tuominen, P. VanBelle, N. van der Stoep, D. C. Whiteman, D. Zelenika, J. L. Han, S. Y. Fang, J. E. Lee, Q. Y. Wei, G. M. Lathrop, E. M. Gillanders, K. M. Brown, A. M. Goldstein, P. A. Kanetsky, G. J. Mann, S. MacGregor, D. E. Elder, C. I. Amos, N. K. Hayward, N. A. Gruis, F. Demenais, J. A. N. Bishop, D. T. Bishop, and M. E. L. Consortium Geno. Genome-wide

association study identifies three new melanoma susceptibility loci. *Nature Genetics*, 43(11):1108–1113, 2011.

[12] S. I. Berndt, C. F. Skibola, V. Joseph, N. J. Camp, A. Nieters, Z. M. Wang, W. Cozen, A. Monnereau, S. S. Wang, R. S. Kelly, Q. Lan, L. R. Teras, N. Chatterjee, C. C. Chung, M. Yeager, A. R. Brooks-Wilson, P. Hartge, M. P. Purdue, B. M. Birmann, B. K. Armstrong, P. Cocco, Y. W. Zhang, G. Severi, A. Zeleniuch-Jacquotte, C. Lawrence, L. Burdette, J. Yuenger, A. Hutchinson, K. B. Jacobs, T. G. Call, T. D. Shanafelt, A. J. Novak, N. E. Kay, M. Liebow, A. H. Wang, K. E. Smedby, H. O. Adami, M. Melbye, B. Glimelius, E. T. Chang, M. Glenn, K. Curtin, L. A. Cannon-Albright, B. Jones, W. R. Diver, B. K. Link, G. J. Weiner, L. Conde, P. M. Bracci, J. Riby, E. A. Holly, M. T. Smith, R. D. Jackson, L. F. Tinker, Y. Benavente, N. Becker, P. Boffetta, P. Brennan, L. Foretova, M. Maynadie, J. McKay, A. Staines, K. G. Rabe, S. J. Achenbach, C. M. Vachon, L. R. Goldin, S. S. Strom, M. C. Lanasa, L. G. Spector, J. F. Leis, J. M. Cunningham, J. B. Weinberg, V. A. Morrison, N. E. Caporaso, A. D. Norman, M. S. Linet, A. J. De Roos, L. M. Morton, R. K. Severson, E. Riboli, P. Vineis, R. Kaaks, D. Trichopoulos, G. Masala, E. Weiderpass, M. D. Chirlaque, R. C. H. Vermeulen, R. C. Travis, G. G. Giles, D. Albanes, J. Virtamo, S. Weinstein, J. Clavel, T. Z. Zheng, T. R. Holford, K. Offit, A. Zelenetz, R. J. Klein, J. J. Spinelli, K. A. Bertrand, et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nature Genetics*, 45(8):868–876, 2013.

[13] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, 22(9):1790–1797, 2012.

[14] N. J. Camp, M. Parry, S. Knight, R. Abo, G. Elliott, S. H. Rigas, S. P. Balasubramanian, M. W. R. Reed, H. McBurney, A. Latif, W. G. Newman, L. A. Cannon-Albright, D. G. Evans, and A. Cox. Fine-mapping CASP8 risk variants in breast cancer. *Cancer Epidemiology Biomarkers & Prevention*, 21(1):176–181, 2012.

[15] Encode Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*, 9(4):e1001046, 2011.

[16] A. Cox, A. M. Dunning, M. Garcia-Closas, S. Balasubramanian, M. W. R. Reed, K. A. Pooley, S. Scollen, C. Baynes, B. A. J. Ponder, S. Chanock,

J. Lissowska, L. Brinton, B. Peplonska, M. C. Southey, J. L. Hopper, M. R. E. McCredie, G. G. Giles, O. Fletcher, N. Johnson, I. D. Silva, L. Gibson, S. E. Bojesen, B. G. Nordestgaard, C. K. Axelsson, D. Torres, U. Hamann, C. Justenhoven, H. Brauch, J. Chang-Claude, S. Kropp, A. Risch, S. Wang-Gohrke, P. Schurmann, N. Bogdanova, T. Dork, R. Fagerholm, K. Aaltonen, C. Blomqvist, H. Nevanlinna, S. Seal, A. Renwick, M. R. Stratton, N. Rahman, S. Sangrajrang, D. Hughes, F. Odefrey, P. Brennan, A. B. Spurdle, G. Chenevix-Trench, J. Beesley, A. Mannermaa, J. Hartikainen, V. Kataja, V. M. Kosma, F. J. Couch, J. E. Olson, E. L. Goode, A. Broeks, M. K. Schmidt, F. B. L. Hogervorst, L. J. Van't Veer, D. Kang, K. Y. Yoo, D. Y. Noh, S. H. Ahn, S. Wedren, P. Hall, Y. L. Low, J. J. Liu, R. L. Milne, G. Ribas, A. Gonzalez-Neira, J. Benitez, A. J. Sigurdson, D. L. Stredrick, B. H. Alexander, J. P. Struewing, P. D. P. Pharoah, D. F. Easton, and Consortium Kathleen Cunningham Fdn Consortium; Breast Canc Assoc. A common coding variant in CASP8 is associated with breast cancer risk. *Nature Genetics*, 39(3):352–358, 2007.

[17] D. F. Easton, K. A. Pooley, A. M. Dunning, P. D. P. Pharoah, D. Thompson, D. G. Ballinger, J. P. Struewing, J. Morrison, H. Field, R. Luben, N. Wareham, S. Ahmed, C. S. Healey, R. Bowman, K. B. Meyer, C. A. Haiman, L. K. Kolonel, B. E. Henderson, L. Le Marchand, P. Brennan, S. Sangrajrang, V. Gaborieau, F. Odefrey, C. Y. Shen, P. E. Wu, H. C. Wang, D. Eccles, D. G. Evans, J. Peto, O. Fletcher, N. Johnson, S. Seal, M. R. Stratton, N. Rahman, G. Chenevix-Trench, S. E. Bojesen, B. G. Nordestgaard, C. K. Axelsson, M. Garcia-Closas, L. Brinton, S. Chanock, J. Lissowska, B. Peplonska, H. Nevanlinna, R. Fagerholm, H. Eerola, D. Kang, K. Y. Yoo, D. Y. Noh, S. H. Ahn, D. J. Hunter, S. E. Hankinson, D. G. Cox, P. Hall, S. Wedren, J. J. Liu, Y. L. Low, N. Bogdanova, P. Schurmann, T. Dork, Raem Tollenaar, C. E. Jacobi, P. Devilee, J. G. M. Klijn, A. J. Sigurdson, M. M. Doody, B. H. Alexander, J. H. Zhang, A. Cox, I. W. Brock, G. MacPherson, M. W. R. Reed, F. J. Couch, E. L. Goode, J. E. Olson, H. Meijers-Heijboer, A. van den Ouweland, A. Uitterlinden, F. Rivadeneira, R. L. Milne, G. Ribas, A. Gonzalez-Neira, J. Benitez, J. L. Hopper, M. McCredie, M. Southey, G. G. Giles, C. Schroen, C. Justenhoven, H. Brauch, U. Hamann, Y. D. Ko, A. B. Spurdle, J. Beesley, X. Q. Chen, A. Mannermaa, V. M. Kosma, V. Kataja, J. Hartikainen, N. E. Day, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093, 2007.

[18] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*,

27(8):861–874, 2006. Fawcett, T.

[19] J. D. French, M. Ghoussaini, S. L. Edwards, K. B. Meyer, K. Michailidou, S. Ahmed, S. Khan, M. J. Maranian, M. O'Reilly, K. M. Hillman, J. A. Betts, T. Carroll, P. J. Bailey, E. Dicks, J. Beesley, J. Tyrer, A. T. Maia, A. Beck, N. W. Knoblauch, C. Chen, P. Kraft, D. Barnes, A. Gonzalez-Neira, M. R. Alonso, D. Herrero, D. C. Tessier, D. Vincent, F. Bacot, C. Luccarini, C. Baynes, D. Conroy, J. Dennis, M. K. Bolla, Q. Wang, J. L. Hopper, M. C. Southey, M. K. Schmidt, A. Broeks, S. Verhoef, S. Cornelissen, K. Muir, A. Lophatananon, S. Stewart-Brown, P. Siriwanarangsan, P. A. Fasching, C. R. Loehberg, A. B. Ekici, M. W. Beckmann, J. Peto, I. Dos Santos Silva, N. Johnson, Z. Aitken, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, F. Marme, A. Schneeweiss, C. Sohn, B. Burwinkel, P. Guenel, T. Truong, P. Laurent-Puig, F. Menegaux, S. E. Bojesen, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, R. L. Milne, M. P. Zamora, J. I. Arias Perez, J. Benitez, H. Anton-Culver, H. Brenner, H. Muller, V. Arndt, C. Stegmaier, A. Meindl, P. Lichtner, R. K. Schmutzler, C. Engel, H. Brauch, U. Hamann, C. Justenhoven, K. Aaltonen, P. Heikkila, K. Aittomaki, C. Blomqvist, K. Matsuo, H. Ito, H. Iwata, A. Sueta, N. V. Bogdanova, N. N. Antonenkova, T. Dork, A. Lindblom, S. Margolin, A. Mannermaa, V. Kataja, V. M. Kosma, et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *American Journal of Human Genetics*, 92(4):489–503, 2013.

[20] B. L. Fridley, E. Iversen, Y. Y. Tsai, G. D. Jenkins, E. L. Goode, and T. A. Sellers. A latent model for prioritization of SNPs for functional studies. *Plos One*, 6(6), 2011.

[21] P. H. GARTHWAITE, J. B. KADANE, and A. OHAGAN. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470), 2005.

[22] W. Gautschi and W.F. Cahill. *Handbook of Mathematical Functions: Eponential Integral and Related Functions*, chapter 5, pages 227–252. 1964.

[23] Y. T. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5(3):1780–1815, 2011.

[24] J. B. S. Haldane. The combination of linkage values, and the calculation of distances between linked factors. *Journal of Genetics*, 8:299–309, 1919.

[25] S. H. Han, K. M. Lee, J. Y. Choi, S. K. Park, J. Y. Lee, J. E. Lee, D. Y. Noh, S. H. Ahn, W. S. Han, D. H. Kim, Y. C. Hong, E. Ha, K. Y. Yoo, and D. H. Kang. CASP8 polymorphisms, estrogen and progesterone receptor status, and breast cancer risk. *Breast Cancer Research and Treatment*, 110(2):387–393, 2008.

[26] G. E. Hoffman, B. A. Logsdon, and J. G. Mezey. PUMA: A unified framework for penalized multiple regression analysis of GWAS data. *Plos Computational Biology*, 9(6), 2013.

[27] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *Plos Genetics*, 4(7), 2008.

[28] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[29] J. Knight, M. R. Barnes, G. Breen, and M. E. Weale. Using functional annotation for the empirical determination of Bayes factors for genome-wide association study analysis. *Plos One*, 6(4), 2011.

[30] E. Lander and L. Kruglyak. Genetic dissection of complex traits - guidelines for interpreting and reporting linkage results. *Nature Genetics*, 11(3):241–247, 1995.

[31] P. H. Lee and H. Shatkay. An integrative scoring system for ranking SNPs by their potential deleterious effects. *Bioinformatics*, 25(8):1048–1055, 2009.

[32] J. B. Maller, G. McVean, J. Byrnes, D. Vukcevic, K. Palin, Z. Su, J. M. M. Howson, A. Auton, S. Myers, A. Morris, M. Pirinen, M. A. Brown, P. R. Burton, M. J. Caulfield, A. Compston, M. Farrall, A. S. Hall, A. T. Hattersley, A. V. S. Hill, C. G. Mathew, M. Pembrey, J. Satsangi, M. R. Stratton, J. Worthington, N. Craddock, M. Hurles, W. Ouwehand, M. Parkes, N. Rahman, A. Duncanson, J. A. Todd, D. P. Kwiatkowski, N. J. Samani, S. C. L. Gough, M. I. McCarthy, P. Deloukas, P. Donnelly, and Consor Wellcome Trust Case Control. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, 44(12):1294–1301, 2012.

[33] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

[34] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913, 2007.

[35] K. Michailidou, P. Hall, A. Gonzalez-Neira, M. Ghoussaini, J. Dennis, R. L. Milne, M. K. Schmidt, J. Chang-Claude, S. E. Bojesen, M. K. Bolla, Q. Wang, E. Dicks, A. Lee, C. Turnbull, N. Rahman, O. Fletcher, J. Peto, L. Gibson, I. D. Silva, H. Nevanlinna, T. A. Muranen, K. Aittomaki, C. Blomqvist, K. Czene, A. Irwanto, J. J. Liu, Q. Waisfisz, H. Meijers-Heijboer, M. Adank, R. B. van der Luijt, R. Hein, N. Dahmen, L. Beckman, A. Meindl, R. K. Schmutzler, B. Muller-Myhsok, P. Lichtner, J. L. Hopper, M. C. Southey, E. Makalic, D. F. Schmidt, A. G. Uitterlinden, A. Hofman, D. J. Hunter, S. J. Chanock, D. Vincent, F. Bacot, D. C. Tessier, S. Canisius, L. F. A. Wessels, C. A. Haiman, M. Shah, R. Luben, J. Brown, C. Luccarini, N. Schoof, K. Humphreys, J. M. Li, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, F. J. Couch, X. S. Wang, C. Vachon, K. N. Stevens, D. Lambrechts, M. Moisse, R. Paridaens, M. R. Christiaens, A. Rudolph, S. Nickels, D. Flesch-Janys, N. Johnson, Z. Aitken, K. Aaltonen, T. Heikkinen, A. Broeks, L. J. Van't Veer, C. E. van der Schoot, P. Guenel, T. Truong, P. Laurent-Puig, F. Menegaux, F. Marme, A. Schneeweiss, C. Sohn, B. Burwinke, M. P. Zamora, J. I. A. Perez, G. Pita, M. R. Alonso, A. Cox, I. W. Brock, S. S. Cross, M. W. R. Reed, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, B. E. Henderson, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics*, 45(4):353–361, 2013.

[36] D. Miki, M. Kubo, A. Takahashi, K. A. Yoon, J. Kim, G. K. Lee, J. I. Zo, J. S. Lee, N. Hosono, T. Morizono, T. Tsunoda, N. Kamatani, K. Chayama, T. Takahashi, J. Inazawa, Y. Nakamura, and Y. Daigo. Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. *Nature Genetics*, 42(10):893–896, 2010.

[37] D. E. Morris, J. E. Oakley, and J. A. Crowe. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4, 2014.

[38] N. E. Morton. Sequential tests for the detection of linkage. *American Journal of Human Genetics*, 7(3):277318, 1955.

[39] S. Palanca Suela, E. Esteban Cardenosa, E. Barragan Gonzalez, I. de Juan Jimenez, I. Chirivella Gonzalez, A. Segura Huerta,

C. Guillen Ponce, E. Martinez de Duenas, J. Montalar Salcedo, V. Castel Sanchez, P. Bolufer Gilabert, and Community Group for Assessment of Hereditary Cancer of Valencia. CASP8 D302H polymorphism delays the age of onset of breast cancer in BRCA1 and BRCA2 carriers. *Breast Cancer Research and Treatment*, 119(1):87–93, 2010.

[40] R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

[41] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[42] S. F. Saccone, N. L. Saccone, G. E. Swan, P. A. F. Madden, A. M. Goate, J. P. Rice, and L. J. Bierut. Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics*, 24(16):1805–1811, 2008.

[43] A. J. Schork, W. K. Thompson, P. Pham, A. Torkamani, J. C. Roddey, P. F. Sullivan, J. R. Kelsoe, M. C. O'Donovan, H. Furberg, N. J. Schork, O. A. Andreassen, A. M. Dale, Consortium Tobacco Genetics, Genomics Bipolar Disorder Psychiat, and Co Schizophrenia Psychiat Genomics. All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *Plos Genetics*, 9(4), 2013.

[44] S. L. Slager and D. J. Schaid. Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Human Heredity*, 52(3):149–153, 2001.

[45] A. V. Smith, D. J. Thomas, H. M. Munro, and G. R. Abecasis. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Research*, 15(11):1519–1534, 2005.

[46] A. V. Spencer, A. Cox, and K. Walters. Comparing the efficacy of SNP filtering methods for identifying a single causal SNP in a known association region. *Annals of Human Genetics*, 78(1):50–61, 2014.

[47] C. C. A. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *Plos Genetics*, 5(5), 2009.

[48] M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

[49] T. Strachan and A. P. Read. *Human molecular genetics (4th edition): pages 443-446.* 2010.

[50] Z. Su, J. Marchini, and P. Donnelly. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304–2305, 2011.

[51] L. Sun, R. V. Craiu, A. D. Paterson, and S. B. Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6):519–530, 2006.

[52] M. S. Udler, S. Ahmed, C. S. Healey, K. Meyer, J. Struewing, M. Maranian, E. M. Kwon, J. Zhang, J. Tyrer, E. Karlins, R. Platte, B. Kalmyrzaev, E. Dicks, H. Field, A. T. Maia, R. Prathalingam, A. Teschendorff, S. McArthur, D. R. Doody, R. Luben, C. Caldas, L. Bernstein, L. K. Kolonel, B. E. Henderson, A. H. Wu, L. Le Marchand, G. Ursin, M. F. Press, A. Lindblom, S. Margolin, C. Y. Shen, S. L. Yang, C. N. Hsiung, D. Kang, K. Y. Yoo, D. Y. Noh, S. H. Ahn, K. E. Malone, C. A. Haiman, P. D. Pharoah, B. A. J. Ponder, E. A. Ostrander, D. F. Easton, and A. M. Dunning. Fine scale mapping of the breast cancer 16q12 locus. *Human Molecular Genetics*, 19(12):2507–2515, 2010.

[53] M. S. Udler, K. B. Meyer, K. A. Pooley, E. Karlins, J. P. Struewing, J. Zhang, D. R. Doody, S. MacArthur, J. Tyrer, P. D. Pharoah, R. Luben, L. Bernstein, L. N. Kolonel, B. E. Henderson, L. Le Marchand, G. Ursin, M. F. Press, P. Brennan, S. Sangrajrang, V. Gaborieau, F. Odefrey, C. Y. Shen, P. E. Wu, H. C. Wang, D. Kang, K. Y. Yoo, D. Y. Noh, S. H. Ahn, B. A. J. Ponder, C. A. Haiman, K. E. Malone, A. M. Dunning, E. A. Ostrander, D. F. Easton, and Search Collaborators. FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Human Molecular Genetics*, 18(9):1692–1703, 2009.

[54] M. S. Udler, J. Tyrer, and D. F. Easton. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genetic Epidemiology*, 34(5):463–468, 2010.

[55] W. Valdar, J. Sabourin, A. Nobel, and C. C. Holmes. Reprioritizing genetic associations in hit regions using LASSO-based resample model averaging. *Genetic Epidemiology*, 36(5):451–462, 2012.

[56] I. Verdinelli and L. Wasserman. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618, 1995.

[57] C. M. Vignal, A. T. Bansal, and D. J. Balding. Using penalised logistic regression to fine map HLA variants for rheumatoid arthritis. *Annals of Human Genetics*, 75:655–664, 2011.

[58] J. Wakefield. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics*, 81(2):208–227, 2007.

[59] J. Wakefield. Reporting and interpretation in genome-wide association studies. *International Journal of Epidemiology*, 37(3):641–653, 2008.

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

[61] W. Y. S. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd. Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118, 2005.

[62] Q. Zhu, D. Ge, E. L. Heinzen, S. P. Dickson, T. J. Urban, M. Zhu, J. M. Maia, M. He, Q. Zhao, K. V. Shianna, and D. B. Goldstein. Prioritizing genetic variants for causality on the basis of preferential linkage disequilibrium. *American Journal of Human Genetics*, 91(3):422–434, 2012.