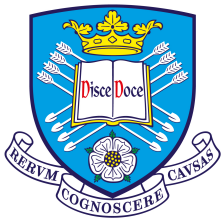


The University Of Sheffield.

Bayesian fine mapping of complex disease genes utilising estimates of the number of yet-to-be-discovered disease-specific SNPs

HANNUUN EADIELA BINTI YAACOB

Submitted for the degree of Doctor of Philosophy

School of Mathematics and Statistics

April 2020

Supervisors: Dr Kevin Walters and Prof Angela Cox

University of Sheffield

ACKNOWLEDGEMENT

All praise to God for giving me the strength and faith throughout the four years of my PhD journey.

First and foremost, I would like to sincerely thanks my supervisor, Dr Kevin Walters for supervising me. I started this PhD journey with a little knowledge about the background of the research, but with Kevin's expertise, knowledge and guidance, I learned and gained a lot throughout the years. Thank you for always helping me with everything. Not to forget, I would like to also thank Professor Angela Cox for your support and inputs towards this thesis. I consider myself fortunate to be able to work with both of you.

This journey could not have been completed if it was not because of my family. Thank you for the love, the understanding, the support, the encouragement, the warmth and the never-ending prayers for me to achieve my dreams. Thank you for believing in me and trusting me to be thousands of miles away from home.

To everyone I met in the UK particularly in Sheffield, thank you for being part of my life here. Everyone has played a big role in cheering me up and making me feels at home. Thank you for always wishing me the best in my PhD. There are a number of people I now consider my own family. Thank you for staying by my side.

Last but not least, thank you to the Ministry of Education Malaysia and University of Malaya (UM) for funding my PhD. Thank you everyone in Department of Applied Statistics, Faculty of Economics and Administration, UM for patiently waiting for me to return and serve the department. My appreciations goes to Raja Zarith Sofiah Foundation for providing me the financial aid for the last couple of months of my PhD.

ABSTRACT

Fine mapping studies aim to prioritize causal variants in complex diseases within genome-wide association studies (GWAS) regions. Bayesian approaches have been widely used in recent fine mapping studies because of the advantage they have in overcoming the limitations of the frequentist approach. The commonly used prior distribution on the causal single-nucleotide polymorphism (SNP) effect size is a Normal distribution with mean zero. Previous studies have shown that the posterior distribution and Bayes factor are both highly sensitive to the Normal prior variance, it is therefore reasonable to assume that posterior summaries are also sensitive to the parametric form of the prior. We show that the Laplace prior for the SNP effect size better reflects both the effect sizes observed in breast cancer GWAS top hits, and the number of yet-to-be-discovered SNPs, than the Normal prior. We estimate the prior parameters from the GWAS top hits and develop single-SNP and multi-SNP approaches for the Laplace prior. We compare our approaches with other existing fine mapping methods using simulated data from HAPGEN and real data from iCOGs. Our analysis shows that the Laplace prior performs better than the current gold standard multi-SNP fine mapping method in terms of causal SNP ranks.

Publications

Walters, K., Cox, A., Yaacob, H. (2019) Using GWAS top hits to inform priors in Bayesian fine-mapping association studies. Genetic Epidemiology. doi: 10.1002/gepi.22212

Yaacob, H., Walters, K. and Cox, A. (2019) Utilizing the information from GWAS data to inform priors in Bayesian fine-mapping association studies. In: Human Heredity. 47th European Mathematical Genetics Meeting (EMGM), 08-09 Apr 2019, Dublin, Ireland. Karger, p. 247. doi: 10.1159/000499459.

Thesis content

Chapter 1 discusses the relevant background of genetics. We begin the chapter by introducing the human genome and explain all genetic terminology relevant to the research. We further discuss protein production which leads to genetic mutation and genetic variations. Moreover, we describe DNA inheritance and cell division. This background gives us an overview of how genetics play a role in causing diseases. At the end of this chapter, we define the minor allele frequency and linkage disequilibrium which are terms often used in this thesis.

Chapter 2 introduces population-based association studies. We briefly explain the role of linkage disequilibrium in association studies. We also introduce the four types of population-based association studies and focus on discussing the genome-wide association studies (GWAS) and also fine mapping studies. This chapter is mainly about the statistical methods that are currently being used in a single-SNP analysis in association studies. Initially, we discuss the standard frequentist approach and highlight the limitations. The limitations in the frequentist approach has been the motivation for us to use Bayesian approach instead. Thus, in the last part of Chapter 2, we review the currently used univariate Bayesian approaches in fine mapping.

Chapter 3 begins with a description of the GWAS breast cancer top hits data. This data is fundamental to our research. Through this data, we inform the effect size prior which is used to develop a Bayesian approach throughout this thesis. The information from the top hits data are used to estimate the hyperparameters. We compare the most common effect size prior used in fine mapping, which is the Gaussian prior, with our proposed prior, the Laplace prior. To estimate the hyperparameters,

we use maximum likelihood estimation. We take into account the number of unidentified SNPs in estimating the hyperparameters. The uncertainty of the estimates is also discussed in this chapter.

Chapter 4 is the derivation of the univariate Bayesian approach to fine mapping. In the beginning of the chapter, we describes the simulated data to be used in evaluating the performance of the Bayesian approaches. ROC curves are used to illustrate the ranking performance of all approaches. A brief discussion of ROC curves is provided in this chapter. We further derive the posterior distribution of the effect size and the Bayes factor using the Laplace prior. From the posterior distribution, we derive the posterior expected value, the posterior median, the posterior credible interval and the posterior highest density interval. In addition to observing the ranking performance of the Laplace Bayes factor, it is also used to identify the noteworthy SNPs. The Laplace Bayes factor is extended to the Laplace Gamma Bayes factor to account for the uncertainty of the parameter estimates which depends on the number of yet-to-be-discovered (YTBD) SNPs.

Chapter 5 reviews current multiple Bayesian approaches to fine mapping. This chapter starts off with a description of the multivariate logistic regression. We then discuss Bayesian variable selection and finally compare and contrast the current Bayesian variable selection methods used in fine mapping.

Chapter 6 contains the derivation of the multiple Bayesian approach. We derive the posterior probability of a model using the Gaussian distribution and the Laplace distribution as the prior for the effect size. We specify the prior probability for the number of causal SNPs to follow a binomial distribution.

Chapter 7 examines the performance of the multiple Bayesian approaches. We first describe the simulated data used in this chapter. We compare the ranking performance (using ROC curves) of the Laplace prior with the Gaussian prior and FINEMAP in 5 different scenarios. The five scenarios are the 5 different maximum number of causal SNPs allowed in the model. The posterior probabilities using the Laplace prior are also provided. We also identify the noteworthy SNPs using the Laplace Bayes factor which can be derived from the posterior probabilities.

Chapter 8 describes the application of the Bayesian approaches to the real data which is available

from the Collaborative Oncological Gene-environment Study (COGS). The data is described at the beginning of the chapter. The same approaches used in Chapter 7 are used to compare the ranks of the SNPs. We also include the p-value in the comparison to see the difference between frequentist and Bayesian approaches. We further use the Laplace prior to identifying the noteworthy SNPs.

Chapter 9 summarises the thesis. We provide the limitations of our method and suggest some possible further work to be considered.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to protein production and DNA inheritance

The discussion in this chapter are mostly referred from Human Molecular Genetics by Tom Strachan and Andrew Read (Strachan, 2011) and Human Molecular Genetics by Peter Sudbery and Ian Sudbery (Sudbery, 2009).

1.1 Chromosome and Inheritance

Genetics is the study of heredity and the variation of inherited characteristics. The inheritance of traits by offspring from each parent is defined by genes. Each human has around 20,000 genes which reside in chromosomes and are made of deoxyribonucleic acid (DNA). Cells in our body contain 46 chromosomes, of which 23 come from our father and the other 23 originate from our mother. There are two types of cells in a human, haploid and diploid. Haploid cells contain a single copy of each chromosome, and diploid cells contain two copies of each chromosome. Most cells in our body are diploid cells, and the sperm and egg are haploid cells. Upon fertilisation, they create a single cell with two sets of the 23 chromosomes.

The chromosomes hold all the genetic information in the form of double stranded DNA. The flow of genetic information within cells follows from a DNA sequence as a template to produce proteins via messengers. The ribonucleic acid (RNA) acts as the messenger in this flow. Proteins carry out most of the body functions and determine traits.

Traits or in other terms, often called phenotypes are observable physical characteristics of a human. Phenotypes are determined by genotypes which are formed by the combination of two alleles at a particular locus or position on the chromosome. Alleles at the same locus explain the same trait. However, there are two alleles for each locus, and the difference in these two alleles will result in different expression of that trait. As an example, for an individual with a facial dimples (trait), there are two forms of allele, dimples (D) and no dimples (d). Thus, the facial dimples can be expressed in either dimples or no dimples depending on the genotype, as follows.

Genotype can either be described as homozygous or heterozygous. A homozygous genotype is when an individual has two identical alleles for a particular trait. In contrast, heterozygous genotype is when an individual has two different alleles at a particular locus. Alleles may be dominant or recessive and it is this feature which contributes to how the trait is expressed. Referring to the previous example, for an individual's facial dimples, having dimples (D) is dominant and not having dimples (d) is recessive. If an individual has a homozygous dominant genotype (DD), that individual has facial dimples. If an individual has the dd genotype, a homozygous recessive, the phenotype for this individual is not having facial dimples. If the individual has the heterozygous genotype Dd, this individual has facial dimples.

Thus for an individual with a heterozygous genotype, the expression of the trait is conditioned on the genetic dominance. In the case of a complete dominance such as eye colour, a heterozygous dominant allele will completely mask the recessive allele and express a dominant phenotype. In incomplete dominance, the phenotype expressed is a mixture of both dominant and recessive alleles. In co-dominance, both alleles completely express their traits, which results in a third phenotype.

1.2 Deoxyribonucleic Acid (DNA)

Deoxyribonucleic acid, commonly known as DNA is the molecular basis of inheritance. DNA stores, replicates and expresses genetic information. It can be found in the nuclei of all cells in the body. The DNA molecule is formed of a double helix. Each strand of the helix is built up by molecules called nucleotides. Each nucleotide consists of a carbon-based sugar called deoxyribose, a phosphate group and a nitrogenous base. The sugar and phosphate are linked together to create the backbone of a DNA. Figure 1.1(a) illustrates the sugar-phosphate backbones on the outside of the double helix. The backbone of each strand has two ends, a $3'$ end and a $5'$ end. One of the backbones run from a $5'$ to $3'$ direction and the other backbone runs the opposite, from $3'$ to $5'$ end. The backbones of these strands are antiparallel.

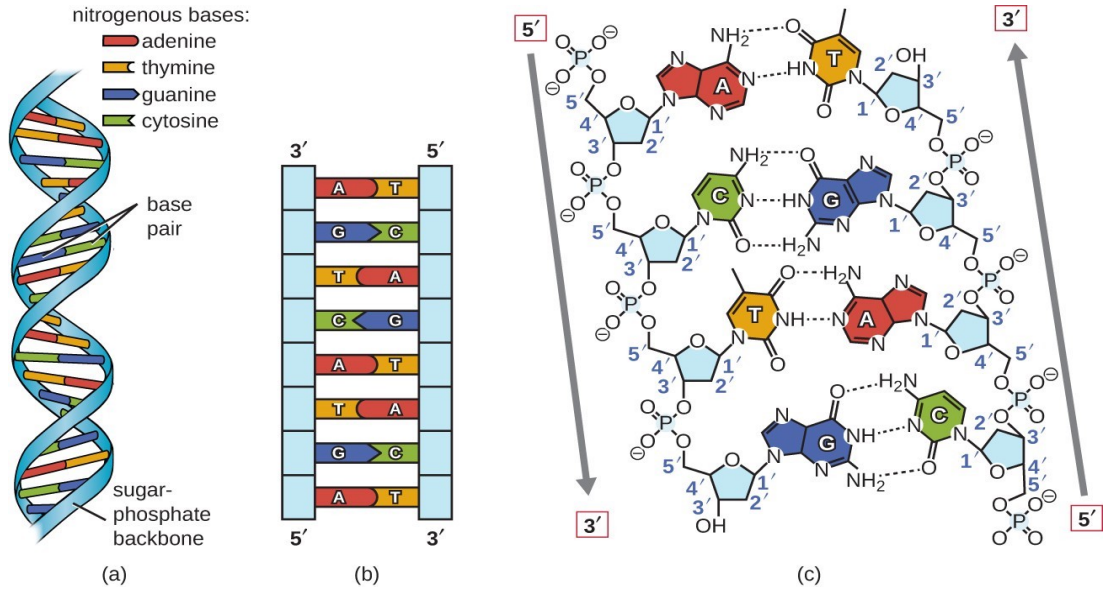


***Figure 1.1*:** An illustration of the deoxyribonucleic acid (DNA). (a) The sugar-phosphate backbones on the outside of the double helix. (b) The complementary base pairs between bases in two antiparallel strands. (c) The molecular structure of DNA (OpenStax, (accessed September 25, 2019)).

There are 4 types of nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). Each base is always attached to the $1'$ carbon of the sugar. T and C are called pyrimidines because these bases have one carbon ring. A and G bases have two rings, they are called purines. These bases make up the rungs of the DNA double helix.

Bases bond with one another across the helix as base pairs. The association between bases are known as complementary base pairs, where A must pair with T while G pairs with C as shown in Figure 1.1(b) . The order of the bases along the length of the helix are sequences that code the genetic information in humans. Due to the pairing in DNA, the composition of A is equal to T, and G is the same as C. The A-T base pair connects with two hydrogen bonds and G-C bonds with three hydrogen bonds.

Figure 1.1(c) illustrates the molecular structure of DNA. The hydrogen bonds stabilize the DNA double helix. The strong covalent bonds between the phosphate and sugar cause the two DNA strands to intertwine. The hydrogen bonds that hold the two strands together through complementary base pairs are non-covalent. These hydrogen bonds are weak which make the helix easily formed and broken for DNA replication prior to cell division. DNA replication allows new cells to receive the full set of genetic instruction. The two strands of DNA act as a template for the synthesis of another two new strands, finally resulting in the production of a set of identical DNA.

1.3 Ribonucleic acid (RNA)

As mentioned above, Ribonucleic acid (RNA) is needed to carry the information from DNA to proteins. Unlike DNA, RNA is commonly single stranded. The nucleotides of RNA are quite similar to the nucleotides of DNA, being built of a phosphate, a nitrogenous base and a sugar. The sugar in RNA is called ribose (rather than deoxyribose in DNA). The other difference between RNA and DNA is in one of the nitrogenous bases. Instead of thymine, RNA contains uracil (U) nucleotide. Enzymes in the cell recognize DNA and RNA through these differences in structure.

The most common RNA types are messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). These three types of RNAs have their individual roles but work together in synthesizing protein. mRNA codes the genetic information from DNA and carries the codes from the nucleus to the cell's cytoplasm. tRNA interprets the codes in mRNA and carries the amino acid to the ribosome to produce protein. The ribosome is a complex structure of rRNA and proteins which translate the mRNA code into protein. A more detailed explanation of the role of RNAs will be discussed in the process of protein synthesis.

1.4 Amino acid, polypeptide and protein

Amino acids are compounds containing carbon, hydrogen, oxygen, nitrogen, and some have sulphur atoms. These atoms form an amino group, a carboxyl group and a side chain. The side chain is the part that differentiates one amino acid from another. There are 20 amino acids and these are linked together in various combinations to become polypeptides and proteins.

The specific side chain of the amino acids define the unique characteristics of each polypeptide. The polypeptide chains fold into a fixed three-dimensional structure to form proteins. There are four levels of protein structural organization. The first is primary structure, amino acids linked in a linear sequence to form polypeptides. A secondary protein structure is where polypeptide chains interact with each other to form beta sheets and alpha helices. The three-dimensional shape of the polypeptide chain is the tertiary structure of protein. Lastly the quaternary structure is where more than one polypeptide chain come together to form a protein.

There are thousands of proteins in each cell. Proteins are responsible for keeping our body functioning. Each protein has its own individual function which includes maintaining the tissues, balancing fluids, storing nutrients and many more. The various molecular structures that form proteins lead to their different functional properties. Our cells continuously produce proteins based on the information coded in DNA through the process of the central dogma of life.

1.5 The central dogma of life

The process of synthesizing protein based on the DNA genetic code is known as the central dogma of life. The process has two parts, transcription and translation. The transcription process is the process of making a single RNA strand from a DNA template. The RNA is then translated into proteins.

1.5.1 Transcription of DNA

In this stage, the genetic information from the DNA is copied to make a single mRNA molecule by the enzyme RNA polymerase. The process begins at a region in the DNA called a promoter. Once the RNA polymerase attaches to the promoter, it starts to "unzip" the DNA helix to expose two DNA strands; template strand and coding strand.

RNA polymerase builds mRNA by stepwise addition of new nucleotides. The synthesizing of new mRNA is in the direction of $5'$ to $3'$. A new nucleotide is added at the $3'$ end of the growing mRNA strand. Each new nucleotide added on the mRNA complements the nucleotide on the DNA template strand. Nucleotide G, C, T and A on DNA template strand will build nucleotide C, G, A and U on mRNA respectively. Thus the mRNA has the same information as the coding strand apart from every T is now U instead.

The mRNA gets longer as RNA polymerase moves along the DNA. The transcription ends after RNA polymerase transcribes a DNA sequence known as terminator. The mRNA is then released and the RNA polymerase will detach from the DNA. In a eukaryotic cell, a RNA transcript is considered as pre-mRNA. It has to go through another stage of processing into a mRNA before being translated into proteins. To modify the pre-mRNA, two components are added to the $5'$ end and $3'$ end of the pre-mRNA strand. $5'$ cap is a modified G nucleotide added to the first nucleotide on the pre-mRNA to protect it and help the ribosome to attach to it. A long chain of A nucleotides which creates a poly-A tail is attached to the $3'$ end. The poly-A tail stabilizes the pre-mRNA and helps to export the mRNA

to cytoplasm.

During transcription, the whole DNA sequence which includes a mix of exons and introns are copied into the pre-mRNA. Exons are the coding sequence use to encode a protein whereas introns are non-coding sequence. Introns are not use for protein translation and they are removed by RNA splicing. A complex molecule known as a spliceosome binds to introns and cuts them out. In addition, the spliceosome pastes the exons together to make a mature mRNA. At this stage, the mature mRNA moves out from the nucleus to go through the process of translation in cytoplasm.

1.5.2 Translation process

In the cytoplasm, the mRNA will engage with the ribosome to initiate the process of translation. The ribosome is a cell structure made of proteins and rRNA, which has two subunits; a large ribosome subunit and a small ribosome subunit. The ribosome travels along the mRNA in the $5'$ to $3'$ direction to read and scan the genetic information. It then uses this information to encode the mRNA into a sequence of amino acids.

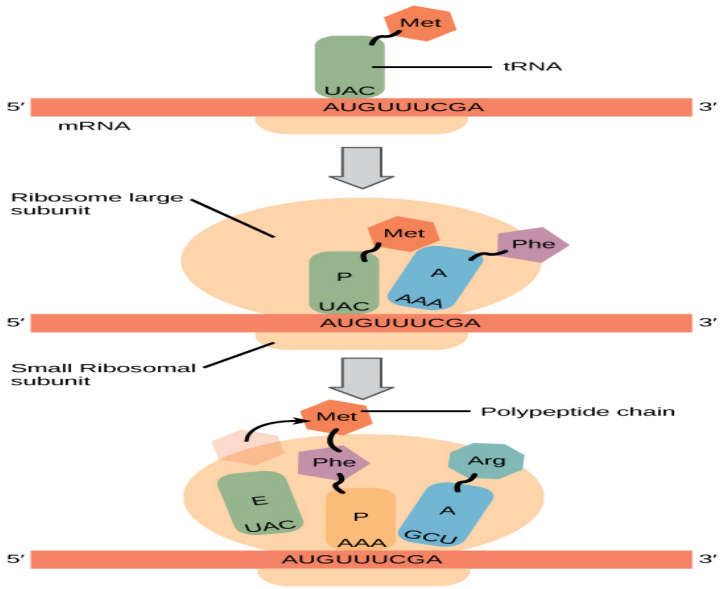


***Figure 1.2*:** An illustration of DNA translation. (Molnar and Gair, (accessed September 25, 2019)).

The information is read in groups of three nucleotides called codons. There are 64 different codons resulting from arranging A, U, G, C nucleotides. 61 of these codons encode for amino acids, while the other three codons are stop codons which terminate the translation. Since there are 61 codons and only 20 amino acids, there is redundancy in the genetic code; more than one codon can code for the same amino acid. Each tRNA carries a specific amino acid and contains the anticodon which is complementary to the codon on mRNA.

An illustration of DNA translation is shown in Figure 1.2. The translation begins with an initiator tRNA carrying the amino acid methionine and attaching to the small ribosome subunit, which binds to the $5'$ end of the mRNA and moves along the mRNA until it arrives at the start codon (AUG). Once the initiator tRNA attaches to the start codon, together with the large ribosome subunit, an initiation complex is formed. At this point, there are now three sites on the ribosome to produce polypeptides. The three sites are known as site P, site A and site E.

The initiation tRNA binds with AUG at site P. Another tRNA carrying an amino acid with an anticodon complementary to the next codon enters site A. The peptide bond forms to connect amino acid on tRNA in P site to amino acid on tRNA in A site. The chain continues to grow in the elongation process when the ribosome moves onto the next codon on the mRNA in $3'$ direction. The tRNA on the P site no longer has an amino acid, this tRNA is shifted to E site and exits the ribosome. At the same time the tRNA at A site shift to P site and a new tRNA with a new amino acid and anticodon lands on site A. This process continues until it terminates when new tRNA at site A binds with the stop codon.

The final stage is where the polypeptide releases to go through further processing to form protein. The small and large subunits separate to allow another translation process to begin on another strand of mRNA.

1.6 Genetic mutation

As outlined above, the synthesis of proteins relies on the DNA sequence. For a protein to function correctly, the DNA must be in the correct sequence to allow the codons to be read correctly and hence form the exact protein. However, changes or in other words mutations in DNA sequence can lead to abnormalities in protein function. Mutation is defined as permanent changes that occur in the DNA nucleotide sequence. Mutation can happened through inheritance or can arise during DNA replication.

There are several types of mutations. One type of mutation is called a frameshift mutation. This occurs by insertion or deletion of one or more nucleotide bases, but not in multiples of three nucleotides. Bases are added or removed, which alters the number of nucleotides in the DNA sequence. This disrupts the triplet codon reading frame and hence affects the amino acid sequence, often leading to the introduction of a premature stop codon and hence a truncated protein which is non-functional.

The second type of mutation is point mutation. Point mutation happens through the substitution of the wrong nucleotide bases, where a single nucleotide base is replaced by another nucleotide base. This changes the structure of a gene. The amino acid coded from a point mutation can result in either a normal protein, a faulty protein or incomplete protein. This is because the point mutation can be a silent, missense or nonsense mutation respectively. In silent mutations, the substitution affects the codon but the new codon encodes the same amino acid. However, if the substitution generates a codon for a different amino acid, this will result in a different type of protein which may be faulty. This is call a missense mutation. Nonsense mutation occurs when the nucleotide being substituted changes the original codon into a stop codon. This stops the translation prematurely, hence producing an incomplete protein. Mutation can cause disease such as cancer. One example is where mutations in the BRCA1 and BRCA2 genes result in dysfunctional BRCA1 or BRCA2 protein, leading to cancer.

1.7 Genetic variation

Genetic variation can be described as the differences in the DNA sequence within and among populations of the same species. This manifests as different forms of alleles in the DNA sequence which may yield different phenotypes, although not all allelic variation results in different phenotype, much of it is silent. Genetic variation plays a role in determining disease susceptibility. It can arise in the population through mutation and through sexual reproduction. Mixing of traits from parents to offspring, and through genetic recombination during gamete formation can lead to different gene combinations and hence increase genetic diversity. An individual's genetic make-up is termed their genome.

Inherited mutations are one form of genetic variation but these tend to be rare. The most common type of genetic variation are Single Nucleotide Polymorphism (SNP). SNPs occur at a particular location where there is a change in a single nucleotide in a DNA sequence. SNPs exist naturally in the genome in more than 1% of a population and hence distinguish one individual from another individual.

DNA sequence in every human are almost identical. To have a better understanding of SNPs, we compare DNA sequences from two individuals in a population. The first individual has a sequence of CGAGGTAAT and the second individual has CGATGTAAT. Notice that both sequences are similar except for the fourth position of the nucleotides. This is an example of a SNP in each individual; the first individual has a G whilst the second individual has a T.

SNPs are stable, inherited and abundantly distributed in the genome. There are approximately 10 million SNPs in one person's genome. SNPs sometimes occur in coding regions of the genome, but more often occur in noncoding regions. SNPs in coding regions may not necessarily change the amino acid sequence of the protein. The majority of diseases are not caused by individual SNPs. However, SNPs can affect disease role and can be informative to predict disease susceptibility. SNPs are used as markers in population studies especially association analysis to identify putative disease genes. These studies rely on the process of Mendelian inheritance of individual SNPs and the patterns of genetic crossing over (recombination) during gamete formation, over many generations.

1.8 Inheritance and cell division

1.8.1 Mendelian inheritance

Inheritance is the process in which genetic material is passed on from parents to their child. Our genome is made of one copy of genome from each of our parents. According to Mendel's experiment using pea plants, the key principles of Mendelian's inheritance are as follows: 1) traits are determined by genes that are passed on from parents to child, 2) for each trait the child has, the child inherits one allele from the mother and one from the father and 3) although the trait may not be visible in a child, he or she can still pass the gene on to their next generation. Mendel's discoveries on the different patterns of inheritance led to Mendel's law of inheritance.

The first law of inheritance is the law of segregation. Every individual has two alleles for each trait. Thus, according to the law of segregation, each gamete will randomly inherit only one of the alleles. The second law is the law of independent assortment. This law states that the sorting of alleles into gametes happens independently, which make all combinations of alleles have the same possibilities to occur. In the third law of inheritance, the law of dominance, because alleles can be dominant or recessive, only one form of trait can appear, thus the dominant allele expresses itself in a trait.

1.8.2 Independent assortment

As mentioned in Section 1.7, sexual reproduction is one of the main sources of genetic variation. Before sexual reproduction occurs, the germ cells go through the process of meiosis, a process of cell division, to produce gametes. In males, the gametes are called sperm cell and in female, they are called egg cells. Gametes are haploid cells. Through fertilization, a diploid cell called the zygote is formed by the union of the sperm and egg cell. The zygote grows through the process of mitosis, producing diploid daughter cells and develops into a new organism. The offspring is unique and is genetically different from both the parents. Each offspring has its own unique combinations of genes from the

gametes produced in meiosis, due to the independent assortment of the chromosome in this process.

The process of meiosis involves two stages of cell division; meiosis I and meiosis II. Prior to meiosis, the chromosomes in the germ cell duplicate. The cell than begins the first stage of meiosis which divides the cell into two haploid cells and further divide into four haploid cells at the end of meiosis II. The four haploid cells are the gametes, each has half the number of chromosomes from the original cell. Each gamete contains a unique assortment of chromosomes, and all gametes are thus genetically distinct from the parents.

Homologous chromosomes look the same and have similar types of gene, however they are non-identical. In a homologous chromosome, one could carry a dominant allele of a gene such as A and the other carry a recessive allele such as a. Independent assortment occurs when homologous chromosomes aligned themselves randomly and independently of one another along the cell's equator during metaphase. This reshuffling of genes results in two different sets of daughter cells. In human, there are 23 chromosomes assorted independently, thus, this will result in having gametes with 2^{23} possible combinations of chromosomes.

Figure 1.3 illustrates the process of independent assortment in meiosis. As an example, we illustrate two pairs of homologous chromosomes. One chromosome carries a recessive allele a and a dominant allele A, and the other carries a recessive allele b and a dominant allele B. In the event of independent assortment, there are two possibilities of alignment. The cell may adopt alignment 1, in which chromosome with allele A and chromosome with allele B end up in the same daughter cell. In an alternative alignment, the alleles in the daughter cell are allele A and allele b. This is because in this alignment, chromosome with allele A aligned together with chromosome having allele b. In the second stage of meiosis, the two daughter cells from each alignment further divides into four gametes with equal number of the genotypes; AB, ab, Ab, aB.

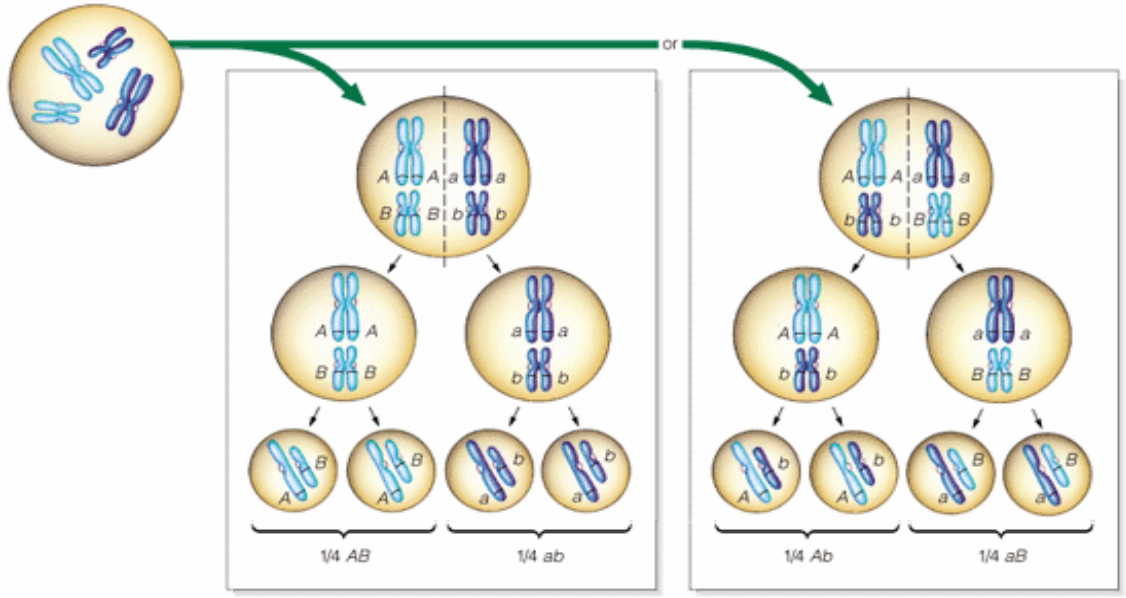


***Figure 1.3*:** An illustration of the process of independent assortment for two pairs of homologous chromosomes. The illustration on the left shows alignment 1. The illustration on the right shows the alternative alignment, alignment 2 (*Meiosis and Formation of Eggs and Sperm*, 2000 (accessed September 25, 2019)).

1.8.3 Genetic recombination

In addition to the independent assortment of homologous chromosomes, the genetic material is further rearranged through crossing over. Crossing over exchanges genetic information between homologous chromosomes. This process takes place during pairing of two homologous chromosomes in the first stage of meiosis I. At this point, paternal chromosome with allele A and B and maternal chromosome with allele a and b are line up in preparation for crossing over. This causes one part of the chromosome to exchange and resulting in four different combination of chromatids (the single strand chromosomes resulting from the second meiotic division, as shown in Figure 1.3); AB, Ab, aB and ab. An illustration of the process of crossing over is shown in Figure 1.4.

If we observe the genotype of the resulting chromatids, there are two gametes having the same genotype as their parents which are AB and ab. These are called non-recombinant genotype. For the other two gametes with genotype Ab and ab, these are the outcomes of recombination and are called recombinant genotype.

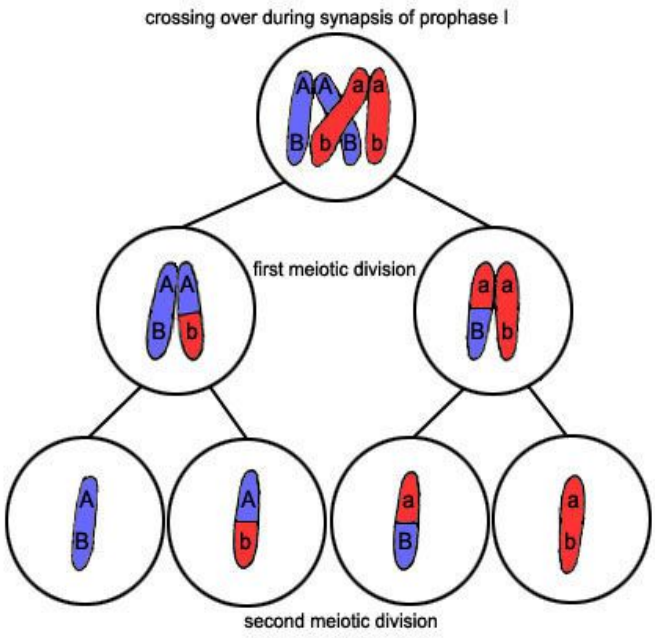


***Figure 1.4*:** An illustration of the crossing over process between two homologous chromosomes during prophase in meiosis (*Mitosis Compared With Meiosis*, 2009 (accessed September 25, 2019)).

The rate of recombination between two loci on the chromosome depends on the distance between them. If two loci on the same chromosome are far apart, the rate of recombination increases. Alleles from distant loci are likely to be in linkage equilibrium, i.e. all combinations of alleles at the two loci are present at their expected frequencies in the population. However, if the loci are close together on the same chromosome, their alleles are likely to be in linkage disequilibrium, i.e. the allele combinations deviate from their expected frequencies. The rate of recombination during inheritance is one of the factors that affects linkage disequilibrium. A more detailed explanation on linkage disequilibrium is described in Section **??**.

1.9 The role of genetics in disease

Geneticists are focused on identifying genes involved in disease. This can explain genes that cause disease or explain how these genes contribute to the cause of disease. Identifying disease genes could

help in measuring people's risk in developing the disease. This could also give further interest in identifying treatment for people at risk. Another purpose of identifying disease genes is to help geneticists understand the mechanism of the disease. As an example, geneticists can investigate the pattern of inheritance to examine disease inheritance in large families.

There are two main categories of genetic diseases; Mendelian and complex diseases. Mendelian diseases, such as cystic fibrosis, sickle-cell disease and Huntingdon's disease, are rare in which they only occur in less than 0.1% of the population. These diseases are caused by single gene mutations. The pattern of inheritance of the single gene mutation is predicted to cause expression of the disease. Thus, individuals carrying this mutation are at high risk of developing the disease.

In contrast, complex diseases such as asthma, diabetes and cancer are more common in the population. Complex diseases are multifactorial diseases, which are influenced by genetic variation, environment and lifestyle. Although an individual inherits genes that affect their susceptibility to a certain disease, it does not mean that he or she will certainly develop the disease. The development of disease is also influenced by the environment and lifestyle of an individual. Having a healthy environment and lifestyle can prevent or change the progression of diseases. Genetic factors can affect the individual's risks associated with disease. A distinction between Mendelian disease and complex disease are summarised in Table 1.1.

***Table 1.1*:** The distinction between two genetic diseases; Mendelian disease and complex disease.

	Mendelian Disease	**Complex Disease**
Examples of disease	Cystic fibrosis Sickle-cell disease Huntingdon's disease	Asthma Diabetes Cancer
Inheritance	Mendelian inheritance	No pattern of inheritance
Factors	Single gene mutation	More than one gene Environmental
Percent in population	Rare ($< 0.1\%$)	Common ($> 1\%$)
Effect on phenotype	Predicts the phenotype	Affects the risk of having the phenotype
Analytic tools	Linkage analysis	Association studies

Identifying disease susceptibility genes is a difficult task. Using genotype technologies, it is possi-

ble to identify the location of the disease genes by their associated genetic markers. The most common analytical tools to locate disease genes are genetic linkage analysis and genetic association studies. The objective of linkage analysis is to locate the disease gene by applying an understanding of the pattern of inheritance within families carrying the disease. In this analysis, the evidence for linkage is tested using a likelihood ratio test at different recombination fractions in families of two or three generations. Linkage analysis has been a powerful tool in identifying disease gene in Mendelian diseases.

Genetic association studies have become the most efficient approach to assess genes associated with complex disease. The main aim for this type of study is to identify genetic variants that can be associated with risk of disease. This association can be measured by an odds ratio, the ratio of the odds of disease in a person with a risk allele and the odds of disease in a person who does not have the risk allele. An approach to this study involves thousands of individuals in a genome-wide association study to seek for association between genotype and disease in a population. By using association studies, we can effectively examine the recombination rates over many generations. Moreover, we can consider including the environmental factors into this study. Since complex diseases are polygenic and multifactorial, genetic association study is the best tool to identify the variants associated to the susceptible to these diseases. Population-based association studies are further discussed in Section 2.1.

1.10 Minor allele frequency

Minor allele frequency (MAF) is widely used in population-based association studies. MAF is derived from allele frequency which calculates the frequency of an allele appearing in a population. Another way of describing allele frequency is how common a specific allele is within a population. In general, allele frequency is defined as follows

$$\text{Frequency of allele A} = \frac{\text{Number of copies of allele A in population}}{\text{Total number of allele A and allele a in population}} \tag{1.1}$$

If we have two alleles in the population, allele A and allele C, the sum of allele frequencies of these alleles must be equal to one. Thus, if in a population, there are 13 alleles A and 5 alleles C, the allele frequency of A is 13/ (13+5) = 0.72 and allele frequency for C is 5/(5+13) = 0.28. The MAF of a SNP is the allele frequency of the rare allele. Over generations, the allele frequency in a population changes. The use of MAF in population studies is to differentiate between common and rare variants in the population.

1.11 Linkage disequilibrium

In Section 1.8.3, we mentioned the phenomena of linkage disequilibrium (LD). Recombination is not the only factor affecting LD. Other factors include non-random mating, mutation rate, genetic drift and population structure. LD is defined as a non-random association between alleles at two or more different loci.

1.11.1 Measures of linkage disequilibrium

Suppose we have two loci. The first loci having alleles A and a. The second loci having alleles B and b. The frequencies for each allele are p_A, p_a, p_B and p_b. Thus, there are four possible haplotypes: AB, Ab, aB and ab. The haplotype relative frequencies are p_{AB}, p_{Ab}, p_{aB} and p_{ab}. One measure of LD is defined as the difference between the haplotype relative frequencies and the product of allele relative frequencies. If we calculate the frequency to measure LD of allele A in the first loci and allele B in

the second loci, LD can be defined as

$$D = p_{AB} - p_A\, p_B. \tag{1.2}$$

However, the magnitude of D is difficult to interpret as a measure of LD. Thus, a preferred measure for LD is by normalising D to the maximum value possible. The normalised D, given as D', is therefore,

$$D' = \frac{D}{D_{\text{max}}}. \tag{1.3}$$

D' varies between 0 for no LD to 1 for complete LD. An alternative measure of LD is by deriving r^2, the correlation between two alleles:

$$r^2 = \frac{D^2}{p_A\, p_a\, p_B\, p_b} \tag{1.4}$$

where r^2 only takes positive values between 0 and 1. In association studies, the most common measures of LD used is r^2, partly because it relates to the statistical power at a tag SNP, relative to the power at the causal SNP. This is the LD measure we use in this thesis.

Chapter 2

Current univariate statistical methods used in population-based association studies

2.1 Population-based association studies

One approach to identifying risk alleles is population-based association studies. In such studies, the variants are determined from hundreds or thousands of unrelated individual with or without the disease. The study is designed to look into whether the frequency distribution of the genotype of affected individuals are significantly different to that of unaffected individuals. Under the null hypothesis, there is no association between the risk allele and disease, the frequency of the risk allele should be equal in both cases (with disease) and controls (without disease). In complex diseases, many unaffected individuals could carry the risk allele and some individuals with disease may not carry the risk allele.

The reason unrelated individuals may share mutations is that they are assumed to came from a common ancestor. Thus, individuals with the disease have inherited the same part of a mutation-carrying chromosomal region from their common ancestor. Many non-causal alleles could be associated with the disease because they might be in linkage disequilibrium (LD) with the causative allele. LD plays an important role in association studies since it increases the number of potential causal variants and

makes it difficult to pinpoint the actual causal SNP.

Population-based association studies can be classified into four types. All studies have the aim of identifying association between variants and disease trait. The first type are candidate polymorphism studies in which the studies focus on determining whether a particular functional polymorphism (SNP) is associated with the disease. The second type are candidate gene studies. In these studies, SNPs under investigation are not necessarily functional but are simply within the gene region. The third and fourth types of studies are genome-wide association studies and fine mapping studies which we describe in more detail.

2.1.1 Genome-wide association studies

Genome-wide association studies (GWAS) has been successful and have become a powerful tool in genetic studies in current research (Bush and Moore, 2012). This is because GWAS overcome the limitation in candidate gene studies in which the studies require prior knowledge of the biological information of the gene. GWAS scans markers across the genome of many people to assess possible association with disease in every region. This is made possible with the completion of The Human Genome Project in 2003 and The International HapMap Project in 2005. The information from these projects include databases contain human genome sequence, pattern of DNA sequence variation, variant frequencies and correlation between variants which are to facilitate GWAS in finding the genetic contribution to diseases.

A chip-based microarray technology has been used in GWAS to measure SNP variation in the population. With 10 million SNPs in the genome, it will be costly and time consuming to genotype all SNPs to include in a GWAS. By utilizing this technology, we can avoid genotyping SNPs with redundant information. One or a few SNPs can be genotyped (tag SNPs) to gain information about 10 to 20 other SNPs. This is because many of the SNPs are associated with each other. Thus, this bring us to the concept of linkage disequilibrium (LD).

Linkage disequilibrium (LD) (Section 1.11) gives an advantage in carrying out a GWAS. The presence of LD would either results in SNP having a direct or an indirect association with the disease. A direct association refers to the genotyped SNP statistically found to be having an association with the disease. However, this is not a typical case in GWAS. More often, the tag SNP has an indirect association because it does not actually cause the disease but might be in high LD with other SNPs that could be a causal SNP.

With hundreds of thousands of SNPs are evaluate in GWAS, a very large number of cases and controls are used in GWAS to detect association of the risk alleles. In a traditional statistical test, an association is based on a p-value threshold of 0.05. However, the large number of SNPs lead to increase in false positive rates. Thus, in GWAS, the p-value threshold (5×10^{-8}) has been modified using a Bonferroni correction for multiple testing with million tests. An effective sample size is required in order to have the power to detect a GWAS p-value threshold. In GWAS, the typical statistical power used is 80% (Hong and Park, 2012).

Several genome-wide association studies have successfully identified regions harbouring common variants associated with complex diseases such as type II diabetes and schizophrenia (Schaid et al., 2018). As an example, in breast cancer, as of 2017, there are 148 breast cancer susceptible loci identified through GWAS (Stacey et al., 2007, 2008; Zheng et al., 2009; Ahmed et al., 2009; Thomas et al., 2009; Turnbull et al., 2010; Fletcher et al., 2011; Ghoussaini et al., 2012; Siddiq et al., 2012; Michailidou et al., 2013; Cai et al., 2014; Milne et al., 2014; Michailidou et al., 2015, 2017) . Results from GWAS tell us that these loci are associated with the disease but does not identify the causal SNP. Within the GWAS significant locus, there is atleast one SNP likely to be in LD with the tag SNP and could be the potential causal SNP. With many highly correlated SNPs in the associated region, it is challenging to determine that a particular SNP contributes to increasing the risk of having disease.

2.1.2 Fine mapping genes

GWAS provide insights into plausible causal variants in GWAS disease-associated region. Current research is interested in prioritizing causal SNPs within GWAS-associated regions and identifying the target genes which can provide information about the disease causing mechanism. It is known that there are hundreds or thousands of SNPs in a small region of interest. Filtering the SNPs to pinpoint the causal SNPs is challenging. This is because SNPs in the region are highly correlated with each other and this makes it difficult to distinguish between causal SNPs and those in LD with them. To disentangle the signal in correlated SNPs, various statistical approaches can be used to fine map the genes.

There are different statistical approaches used to prioritize casual SNPs from an associated region. One approach is filtering the SNPs based on p-value or a certain LD threshold (Spencer et al., 2014). In another statistical approach, the Bayesian framework, the strength of association is measured using Bayes factors which compare the evidence under the null with the evidence under the alternative hypothesis. Variants are filtered into a credible set believing that the causal variant will be included in the set.

An advantage of fine mapping is that causal variants filtered from a GWAS region using statistical approaches can be further investigated based on their biological function. Various types of functional data exist for use in fine mapping studies. The functional annotation can be incorporated into Bayesian statistical analyses and hence can improve the performance of fine mapping the causal variants.

2.2 Standard frequentist statistical approach

2.2.1 Logistic regression

Our interest is to fit statistical models to assess association of SNPs with case-control binary outcomes. We want to understand how the predictors can affect the chances of the outcome to occurs. This can best be done using logistic regression. The reason is, the logit function in logistic regression directly relates to the probability of the outcome occurring (usually of having a disease) to not having a disease. Therefore, we let

$$p(y = 1) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}. \tag{2.1}$$

Where y is a binary variable with 1 representing a diseased individual and 0 a control, and x represents the genotype coded in some way and possibly imputed. Now, the right-hand side of Equation (2.1) could only yield the values between 0 to 1. This expression is called logistic function. In the case where an event occurs with probability $p(y = 1)$, the odds in favour is $p(y = 1)/p(y = 0)$. Using Equation (2.1), the odds in favour of an occurring event is,

$$\frac{p(y = 1)}{p(y = 0)} = exp(\alpha + \beta x). \tag{2.2}$$

By taking the logarithm of both sides of Equation (2.7), we now have,

$$log\left[\frac{p(y = 1)}{p(y = 0)}\right] = \alpha + \beta x. \tag{2.3}$$

Thus, we assume a linear relationship between $log\left[p(y = 1)/p(y = 0)\right]$ and x instead of assuming

$p(y = 1)$ has a linear relationship with x. Therefore, in logistic regression, we want to fit the model as in Equation (2.3).

2.2.2 Odds Ratio

We are interested in the association between variants and disease. This is usually quantified in terms of odds ratio. Odds ratio (OR) relates the odds of an occurring event in cases to the odds of an event occurring in controls. In the context of genetics, OR measures the ratio of odds of a disease in an individual with a genotype over an individual with another genotype. The genotype is coded as 0, 1 or 2 which represent the number of copies of the risk alleles at a SNP. A homozygous wildtype is SNP with no risk allele, thus the genotype is coded as 0. A homozygous risk SNP has genotype coded as 2 which indicates two risk alleles. Genotype coded as 1 is for heterozygous SNP, one of the alleles is the risk allele.

In this thesis, we assumed SNP has an additive effect on the disease, from Equation 2.7, the odds of disease for a homozygous wildtype is

$$\frac{p(y = 1|x = 0)}{p(y = 0|x = 0)} = exp(\alpha). \tag{2.4}$$

For a heterozygous SNP, the odd of disease is

$$\frac{p(y = 1|x = 1)}{p(y = 0|x = 1)} = exp(\alpha + \beta) \tag{2.5}$$

and for homozygous risk SNP, the odd of disease is given by

$$\frac{p(y = 1|x = 2)}{p(y = 0|x = 2)} = exp(\alpha + 2\beta). \tag{2.6}$$

Thus, if we let homozygous wildtype be the reference genotype, the odds ratio for an individual with

a heterozygous SNP having a disease is

$$\begin{aligned} \text{OR} &= \frac{exp(\alpha + \beta)}{exp(\alpha)} \\ &= exp(\beta) \end{aligned} \tag{2.7}$$

ORs can be easily related to the parameters of logistic regression model (with a logit link) when we fit the model as in Equation 2.3. The estimated coefficient, β, which is determined by maximum likelihood estimation, is the estimate of the log OR. Throughout this thesis, the log OR is referred as the effect size of a SNP.

2.2.3 Frequentist approach and its limitation

Frequentist approaches had been widely used in assessing evidence for true causal association between genetic variants and disease present in a population. In this approach, the p-value is computed under the null hypothesis of no association. In GWAS, if a SNP has a p-value less than the p-value threshold (5×10^{-8}), this shows some evidence against the null hypothesis of no association with the disease. This SNP is now considered as a candidate causal SNP.

P-value is often used to place the rejection region to make conclusions either to reject or not to reject the null hypothesis. If p-value is less than the rejection region, thus, we reject the null hypothesis. In hypothesis testing, there is some probability of rejecting the null hypothesis known as the power of the test. Thus, p-values should be interpreted with regard to power. However, our question turns to how confident are we in quantifying the evidence that the SNP is truly associated to the disease?

Identical p-values calculated at different SNPs and in different studies lead to different conclusion about the evidence of true association, because the power maybe different. A SNP's minor allele frequency, the effect size and the sample size of the study affect the power of the test. Thus, with a powerful test, we will have smaller p-values with more evidence to reject null hypothesis. However,

if a test has low power, such p-value may be supporting more evidence on the null hypothesis of no association. With large or small power, there are still risks of discarding evidence of detecting SNPs with true association.

2.3 Bayesian approaches

Bayesian approaches are becoming increasing common as an alternative to overcome the limitation of the frequentist approach. A Bayesian approach does not suffer from having to take account of power, it is incorporated into the Bayes factor. In a Bayesian analysis based on Bayes factors, the strength of evidence of an association can be computed among SNPs and throughout the studies without needing to interpret them with respect to allele frequencies and sample sizes (Wakefield, 2008). Furthermore, this approach has an advantage in providing a way to incorporate genomic information in the analysis (Spencer et al., 2016).

2.3.1 Summarising posterior distributions

The essential elements in Bayesian analysis are the prior distribution and likelihood function. The prior distribution, $\pi(\theta)$ specifies knowledge about the parameter, θ before the data is observed. A likelihood function, $f(x|\theta)$ gives the likelihood of the data x given the parameter. Bayes theorem allows the computation of the posterior distribution by incorporating both prior distribution and the likelihood function, which is a probability distribution. Bayes theorem states that, the posterior probability distribution of a parameter given the data can be computed as follows

$$f(\theta\,|x) = \frac{f(x|\theta)\ \pi(\theta)}{\int_\theta f(x|\theta)\ \pi(\theta)\ d\theta}$$

where the integral could be multi-dimensional.

The posterior distribution captures the posterior uncertainty of the parameter. This can be de-

scribed by the summaries of the posterior distribution. In Bayesian analysis, the summaries that are often used are the point estimates and the interval estimates. Posterior point estimates, i.e expected value and median, give a value that best estimates, in some sense, the unknown parameter. On the other hand, interval estimates give an interval indicating the uncertainty in the parameter. Commonly used Bayesian interval estimates are credible interval and highest density interval. All the summary statistics can be derived from the posterior distribution.

2.3.2 Bayes factor

The purpose of hypothesis testing is to evaluate if there is enough statistical evidence to support one hypothesis about a parameter. Early research used univariate frequentist approaches in hypothesis testing which utilise p-values from a likelihood ratio test. However, the p-value has it weaknesses. The p-value indicates the frequency of observing a more extreme test statistics in multiple imaginary experiments given the null hypothesis is actually true, with no consideration of the alternative hypothesis.

Another alternative to hypothesis testing is by using a Bayesian approach. One common Bayesian approach considers the marginal likelihood of the data in both null and alternative hypotheses. The ratio of the values of the marginal likelihood under each hypothesis are compared. This ratio between these two competing hypotheses is called a Bayes factor (BF). Generally, the Bayes factor is given by

$$\text{Bayes factor} = \frac{f(data \mid H_1)}{f(data \mid H_0)}. \tag{2.8}$$

A Bayes factor is equal to 1 shows that the data is equally likely under both hypotheses. If the Bayes factor is greater than 1 it gives more evidence to the alternative, and if it is less than 1 the data is more supported under the null hypothesis.

Since Bayes factors take account of both hypotheses, this gives a major advantage in hypothesis testing compared to p-values. Interpretation of Bayes factor is straightforward as the Bayes factor

specifies how much more likely the data are under one hypothesis compared to the other hypothesis. A drawback of Bayes factor is that a threshold needs to be applied, which is not easy to specify meaningfully.

2.3.3 Posterior odds

To assess the evidence that a causal association exists, we are required to obtain the posterior odds on the alternative hypothesis. This requires the prior probability of the alternative hypothesis to be specified. The prior probability of the alternative hypothesis is denoted by π, so the prior probability on the null hypothesis is given by $1 - \pi$. Using Bayes theorem, the posterior probability of the alternative hypothesis is

$$P(H_1 \mid data) = \frac{f(data \mid H_1) \times \pi}{f(data)}. \tag{2.9}$$

Thus, the posterior odds on the alternative hypothesis is given by

$$\frac{P(H_1 \mid data)}{P(H_0 \mid data)} = \frac{f(data \mid H_1)}{f(data \mid H_0)} \times \frac{\pi}{1-\pi}. \tag{2.10}$$

Equation (2.10) can be written as

$$\text{Posterior odds on } H_1 = \text{Bayes factor} \times \text{prior odds (PO) on } H_1. \tag{2.11}$$

Computing the Bayes factor for each SNP therefore leads to the posterior probability of association (PPA) which is given by

$$\text{PPA} = \frac{\text{Posterior Odds on } H_1}{(1 + \text{Posterior Odds on } H_1)}. \tag{2.12}$$

PPA is basically a probability, regardless of power, sample size and the number of SNPs being tested. PPA can also be thought of as p-value in a Bayesian analysis. Thus, it can be used to make decision on which SNPs to take forward for further analysis. To have strong evidence of true association (large PPA), the Bayes factor has to be large relative to the odds on H_0. This can be proven by using Equation (2.11) and Equation (2.12).

$$\begin{aligned} \text{PPA} &= \frac{\text{BF} \times \frac{\pi}{1-\pi}}{1 + \text{BF} \times \frac{\pi}{1-\pi}} \\ \text{PPA} &= \frac{\text{BF} \times \pi}{(1-\pi) + \text{BF} \times \pi}. \end{aligned} \tag{2.13}$$

Following Equation (2.13), for PPA to be large, it requires

$$\begin{aligned} 1 - \pi &\ll \text{BF} \times \pi \\ 1 &\ll \pi(\text{BF} + 1) \\ \text{BF} &\gg \frac{1-\pi}{\pi} \\ \text{BF} &\gg \text{Prior odds on } H_0. \end{aligned} \tag{2.14}$$

2.3.4 Wakefield Bayes factor in univariate analyses

The currently most commonly used Bayes factor in GWAS and fine mapping was derived by Wakefield (2008) who came up with an asymptotic approximate Bayes factor that overcomes both concerns about computation and specifying the prior on the intercept. His approach is applicable to be used with large sample sizes since it requires the asymptotic Gaussian distribution of the maximum likelihood estimation (MLE) from the logistic regression in Equation (2.3) as follows

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{00} & \boldsymbol{I}_{01} \\ \boldsymbol{I}_{01}^T & \boldsymbol{I}_{11} \end{bmatrix}^{-1} \right)$$

where α is the intercept and $\begin{bmatrix} \boldsymbol{I}_{00} & \boldsymbol{I}_{01} \\ \boldsymbol{I}_{01}^T & \boldsymbol{I}_{11} \end{bmatrix}$ is the Fisher Information matrix. Wakefield derived the Bayes factor from the ratio comparing H_0 and H_1 which is the reciprocal of the Bayes factor we defined in Equation (2.8). Thus, the Bayes factor is

$$\begin{aligned} \text{Bayes Factor} &= \frac{P(data \,|\, H_0)}{P(data \,|\, H_1)} \\ &= \frac{\int \, p(\hat{\alpha}, \hat{\beta} \,|\, \alpha, \beta = 0) \; \pi(\alpha) \; d\alpha}{\int \int \, p(\hat{\alpha}, \hat{\beta} \,|\, \alpha, \beta) \; \pi(\alpha, \beta) \; d\alpha \, d\beta}. \end{aligned} \tag{2.15}$$

Previously, Wakefield (2007) assumed α and β were independent. Later, he relaxed the assumption by reparameterising α to be θ (Wakefield, 2008). The new parameter θ depends on β and α as follows

$$\theta = \alpha + \frac{\boldsymbol{I}_{01}}{\boldsymbol{I}_{00}} \beta.$$

We show in Appendix A that $cov(\hat{\theta}, \hat{\boldsymbol{\beta}}) = 0$ where $\hat{\boldsymbol{\beta}}$ is a vector and that $var(\hat{\boldsymbol{\beta}})$ is the same in both parameterisations. This case is a special of that in Appendix A. Following Equation (2.15), the Bayes factor can be written in the same form by replacing α by θ. Assuming independent priors on θ and β, the joint prior is $\pi(\theta, \beta) = \pi(\theta) \, \pi(\beta)$. Asymptotically, the Bayes factor now becomes

$$\begin{aligned} \text{Bayes Factor} &= \frac{\int \, p(\hat{\theta}, \hat{\beta} \,|\, \theta, \beta = 0) \; \pi(\theta) \; d\theta}{\int \int \, p(\hat{\theta}, \hat{\beta} \,|\, \theta, \beta) \; \pi(\theta, \beta) \; d\theta \, d\beta} \\ &= \frac{\int \, p(\hat{\theta} \,|\, \theta) \; \pi(\theta) \; d\theta \; p(\hat{\beta} \,|\beta = 0)}{\int \, p(\hat{\theta} \,|\, \theta) \; \pi(\theta) \; d\theta \; \int \, p(\hat{\beta} \,|\beta) \; \pi(\beta) \; d\beta} \\ &= \frac{p(\hat{\beta} \,|\beta = 0)}{\int \, p(\hat{\beta} \,|\beta) \; \pi(\beta) \; d\beta}. \end{aligned} \tag{2.16}$$

The main idea is to have a Bayes factor that no longer depends on the prior of the intercept (α). Basically, it is easy to put a prior on the effect size (β) since we have some idea about the effect size. It is not really clear what is the sensible prior for the intercept could be. Thus, assuming independent on

θ and β allow us to derive Bayes factor without the intercept. According to Wakefield (2008), although we assume independent prior on θ and β, this does not mean α and β are independent priors.

Wakefield (2008) used the asymptotic Gaussian distribution of the maximum likelihood estimator (MLE) of β ($\hat{\beta}$), which gives N(β,V). β is the log odds ratio (OR) of the causal SNP which Wakefield (2008) assumed, a priori, to follow a normal distribution with mean 0 and variance W. Thus, the prior specification for this parameter is, $\beta \sim N(0, W)$. By specifying prior a N(0,W) on β and a likelihood $\hat{\beta}|\beta \sim N(\beta, V)$, Equation (2.16) results in an asymptotic Bayes factor derived by Wakefield (2008), which from now on is called WBF. WBF can be calculated as

$$\text{WBF} = \sqrt{\frac{V+W}{W}} exp(-\frac{z^2}{2}\frac{W}{V+W}).$$

To calculate WBF, it requires the Z-score ($z^2 = \hat{\beta}/V$) which is the usual Wald test, the standard error $\sqrt{V}$ and the prior effect size variance, W. WBF uses the summary statistics from a univariate logistic regression to estimate $\hat{\beta}$ and V. However, the prior variance, W also needs specification. We use the Gaussian approximation of the MLE of $\hat{\beta}$ as our likelihood in the rest of the thesis.

The choice of W is crucial because WBF highly depends on the value of W. Wakefield suggested two distinct choices in specifying W: (1) W independent of the minor allele frequency and (2) W dependent on the minor allele frequency. Spencer et al. (2015) suggested that, instead of specifying a fixed value to W, they specify a probability distribution on W and allow uncertainty about W. They considered 4 priors for W which include three priors from parametric families and a fixed form. These priors depend on the variance obtained from the genotype data, which therefore makes it not a true priors in a Bayesian sense. However, these priors allow flexibility in the calculation of Bayes factor and yield tractable integrals and therefore easy to evaluate Bayes factor. Another approach to specifying W is by maximizing the marginal likelihood (Spencer et al., 2016), a so-called Empirical Bayes estimate.

2.3.5 Bayesian decision theory

Bayes factor also can be used in Bayesian decision theory to assess whether the strength of an association is noteworthy i.e is an association worth paying attention to. The posterior probability of the null and cost related to the decision making are required to describe a Bayesian decision theory approach in reporting either the null hypothesis (H_0) or the alternative hypothesis (H_1) (Wakefield, 2007). Reporting an association to be noteworthy is based on minimizing the posterior expected cost. Table 2.1 provides the cost related to making a decision (δ). C_ω is the cost of a false non-discovery and the cost of a false discovery is represented by C_η.

***Table 2.1*:** Cost $C(\delta, H)$ of Decision Making. C_η is the cost of a false discovery and C_ω is the cost of a false non-discovery.

		Decision	
		Non Noteworthy ($\delta = 0$)	**Noteworthy** ($\delta = 1$)
Truth	H_0	0	C_η
	H_1	C_ω	0

We let $\delta = 0$ if the decision is non-noteworthy and $\delta = 1$ if the decision is noteworthy. The posterior expected cost of making decision δ is given by

$$E[C(\delta)] = C(\delta, H_0)\, P(H_0 \mid \hat{\beta}) + C(\delta, H_1)\, P(H_1 \mid \hat{\beta}).$$

where $C(\delta, H)$ is the cost in Table 2.1. Thus, for making both decisions, the posterior expected cost are

$$\begin{aligned} E[C(\delta = 1)] &= C_\eta \times P(H_0 \mid \hat{\beta}) + 0 \times P(H_1 \mid \hat{\beta}) \\ &= C_\eta \times P(H_0 \mid \hat{\beta}) \qquad (2.17) \\ E[C(\delta = 0)] &= 0 \times P(H_0 \mid \hat{\beta}) + C_\omega \times P(H_1 \mid \hat{\beta}) \\ &= C_\omega \times P(H_1 \mid \hat{\beta}). \qquad (2.18) \end{aligned}$$

If we choose to report on a noteworthy association ($\delta = 1$), a decision can be made by minimizing the posterior expected cost, i.e $E[C(\delta = 1)] < E[C(\delta = 0)]$. Therefore, an association is noteworthy is based on

$$\text{Posterior Probability of } H_0 < \frac{C_\omega/C_\eta}{1 + C_\omega/C_\eta}. \tag{2.19}$$

where $R = C_\omega/C_\eta$ is a ratio of costs of making incorrect decisions. In the case $R = 4$, which means the cost of a type II error is 4 times bigger than the cost of a type I error, we can conclude that the association is significant if the posterior odds on the null hypothesis is less than $R = 4$ since the posterior odds on the null hypothesis is, using Equation (2.19), $R = C_\omega/C_\eta$.

We can easily calculate the posterior probability of H_0 from the posterior odds of H_0. Let π_0 be the prior probability of H_0. The posterior odds is given by

$$\text{Posterior Odds of } H_0 = \text{prior odds of } H_0 \text{ / BF}.$$

So the posterior probabilty of H_0 is

$$P(H_0 \mid \hat{\beta}) = \frac{\pi_0}{\pi_0 + \text{BF}(1 - \pi_0)}. \tag{2.20}$$

Though the interpretation of the Bayes factor is straightforward, there are a few concerns to take into consideration about the Bayes factor. The main concern is regarding the computation of Bayes factors. From Equation (2.8), the Bayes factor in a univariate analysis of fine-mapping data is given by

$$\begin{aligned} \text{Bayes factor} &= \frac{f(data \mid H_1)}{f(data \mid H_0)} \\ &= \frac{\int_{\beta \in \mathbb{R}/\{0\}} f(data \mid \beta)\, \pi(\beta)\ d\beta}{f(data \mid \beta = 0)}. \end{aligned}$$

Where β is the effect size given in Equation (2.1). $\pi(\beta)$ is the prior for the parameter under the alternative. From the above equation, we can see that the computation involves (possibly multi-dimensional) integration which, depending on the prior, could require numerical methods, for example Monte Carlo integration.

Chapter 3

Using GWAS top hits data and estimates of the number of yet-to-be-discovered SNPs to inform the effect size prior

3.1 Breast cancer top hits data

In previous years, genome wide association studies have been conducted to identify breast cancer susceptibility loci. Easton et al. (2007) had first identified five new independent SNPs associated with breast cancer. Following this discovery, more SNPs are identified to be significant (Stacey et al., 2007, 2008; Zheng et al., 2009; Ahmed et al., 2009; Thomas et al., 2009; Turnbull et al., 2010; Fletcher et al., 2011; Ghoussaini et al., 2012; Siddiq et al., 2012; Cai et al., 2014; Milne et al., 2014) with 41 SNPs being the largest identification by Michailidou et al. (2013). These SNPs were identified through similar genome wide approaches using up to 70000 cases and 68000 controls from both European and Asian ancestry. All identified SNPs are associated at genome wide significant level of 5×10^{-8} . With large statistical power, these analyses are able to capture SNPs with odds ratio between 1.05 and 1.26 (Fachal and Dunning, 2015).

Michailidou et al. (2013) suggested that there are a large number of SNPs associated with breast cancer risk although in current study these SNPs are not significantly associated at genome wide significant level. Hence, Michailidou et al. (2013) further analyze a set of 10668 SNPs selected from two GWAS. The estimated OR are calculated for each SNP to observe its direction in both GWAS. Assuming if all SNPs are actually non causal, they expect half of the SNPs are in the same direction and the other half are in the opposite direction. However, of 10668 SNPs, 5918 SNPs are in the same direction (both positive or both negative) in the two GWAS meanwhile 4750 SNPs are in the opposite direction. Thus, it is estimated that there are about 1168 additional loci that are associated with higher risk of having breast cancer, most will presumably have very small effect sizes.

In 2015, Michailidou et al. carried out a meta-analysis of 11 GWAS using case-control breast cancer data and also case-control genotyped data from 41 studies restricting only the women of European ancestry. They resulted in identifying 15 more SNPs associated with breast cancer risk at GWAS p-value threshold (Michailidou et al., 2015). In a further GWAS of breast cancer, another 65 new SNPs are reported to be associated at p-value less than 5×10^{-8} with the smallest odd ratio of 1.02 (Michailidou et al., 2017). In this analysis, they used a larger number of case-control data from European ancestry and also included case-control data of East Asian ancestry.

Combining all SNPs identified in various studies above, as of 2017, we have a total of 148 GWAS significant SNPs (top hits) associated with breast cancer risk. Figure 3.1 shows the histogram of the frequency distribution of the log odds ratios from the top hits data used. From the histogram, we can observe that the log odds ratios have an absolute value greater than log 1.02. This shows that the statistical power in previous studies was not sufficient enough to capture SNPs with very small odds ratios.

In this thesis, we will develop a Bayesian statistical approach to fine mapping by suggesting a suitable prior for the effect size. With the 148 Breast Cancer top hits data, we can utilize this data to inform the prior for the effect size. Important information such as the estimated number of unidentified SNPs and the threshold value of the odds ratio are also taken into account. The summaries of the top

hits data and the additional information we have are crucial in estimating the prior parameters. This is because the prior parameters may strongly influence the results of a Bayesian statistical approach to fine mapping.

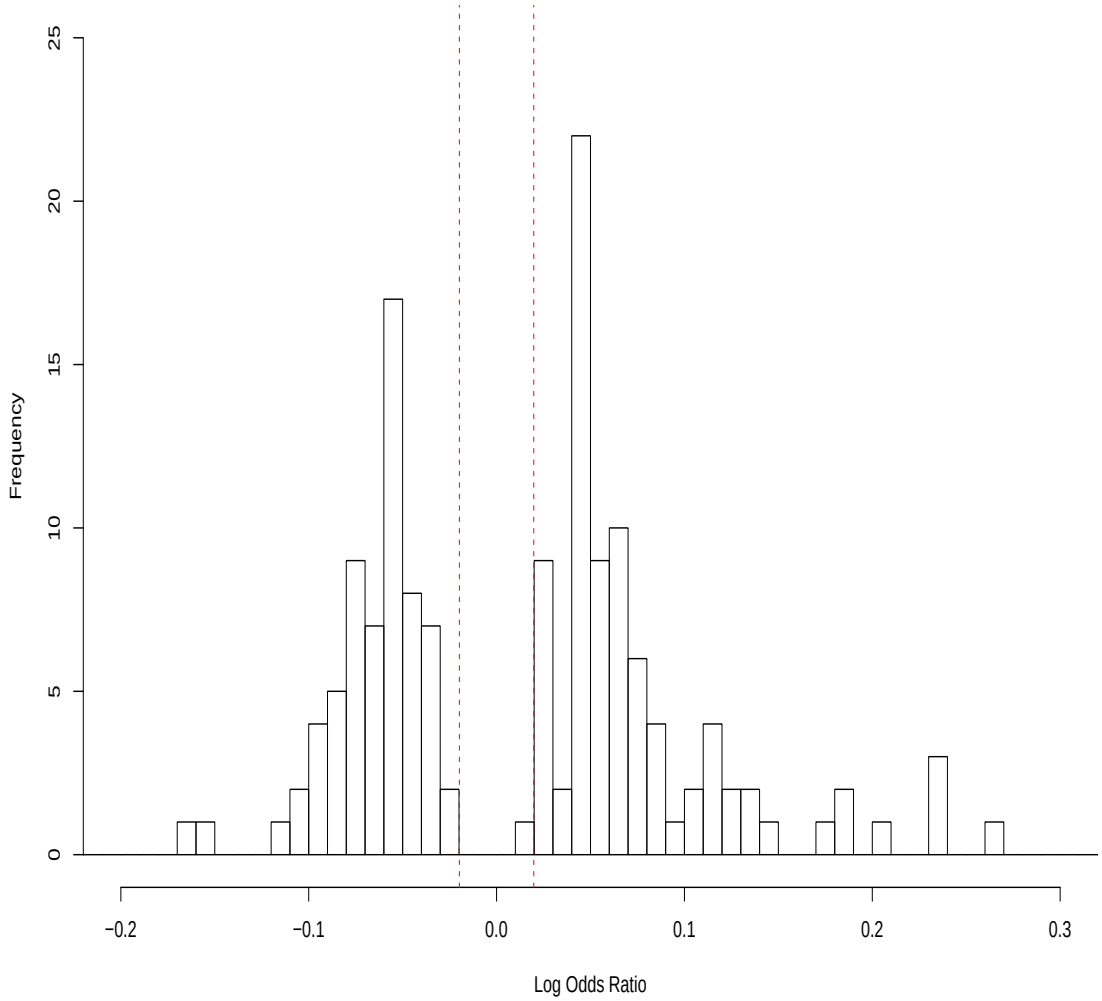


***Figure 3.1*:** Histogram shows the frequency distribution of the log odds ratio from the Breast Cancer top hits data with 148 samples. The red dotted lines represent $|\log(1.02)|$.

3.2 Estimate the hyperparameter used in Wakefield Bayes factor

Wakefield Bayes factor (WBF) is a common used Bayes factor in GWAS which allow the effect size prior to follow a normal distribution. WBF requires a prior on the log OR, β having mean zero and variance W. To be able to calculate WBF, we need to specify W. There are several ways to specify W, one assumes independence of the effect size and the minor allele frequency (MAF), another specifies W so that the WBF ranks match those of the p-values, the so called p-value prior (Wakefield, 2009). However, generally there is uncertainty about the value of W. Instead of specifying a fixed value for W, Spencer et al. (2015) allow for uncertainty about W in the calculation of WBF. Besides having a prior for the log OR, they considered priors for W from three different parametric families (power, exponential and hybrid) and a fixed form (reciprocal).

In this research, we are interested in assessing how well the normal prior fits the top hits data. The specification of W is made through estimation by using the Maximum Likelihood Estimate (MLE) obtained from the known top hits data.

3.2.1 MLE for W without considering the number of yet-to-be-discovered SNPs

Previously we mentioned about the top hits data having values of absolute log ORs ($|\beta|$) greater than a certain value of log OR (β_c). To calculate the MLE for W, we have to take this into consideration. Let β be the random variable for the log OR having a Gaussian distribution with mean=0 and variance=W

$$\beta \sim N(0, W).$$

If we condition on $|\beta|$ to be greater than β_c (where $\beta_c > 0$) we require

$$P(|\beta| > \beta_c \mid \beta \sim N(0, W)) = \int_{-\infty}^{-\beta_c} \frac{1}{\sqrt{2\pi W}} exp\Big(- \frac{\beta^2}{2W}\Big) d\beta + \int_{\beta_c}^{\infty} \frac{1}{\sqrt{2\pi W}} exp\Big(- \frac{\beta^2}{2W}\Big) d\beta$$
$$= 2\left[\Phi\left(\frac{-\beta_c}{\sqrt{W}}\right)\right]$$

where $\Phi(\cdot)$ is the the distribution function of a standard normal random variable. Thus

$$f(\beta \,\big|\, |\beta| > \beta_c, \beta \sim N(0, W)) = \frac{\frac{1}{\sqrt{2\pi W}} exp\Big(- \frac{\beta^2}{2W}\Big)}{2\left[\Phi\left(\frac{-\beta_c}{\sqrt{W}}\right)\right]}$$
$$= \frac{exp\big(- \frac{\beta^2}{2W}\big)}{\sqrt{8\pi}\sqrt{W}\left[\Phi\big(\frac{-\beta_c}{\sqrt{W}}\big)\right]}.$$

If β_i $(1 \leq i \leq n)$ are the observed top hits log odds ratio then the likelihood function is

$$\mathcal{L}(W; \beta_1, \beta_2, ..., \beta_n) = \prod_{i=1}^{n} \frac{exp\big(- \frac{\beta_i^2}{2W}\big)}{\sqrt{8\pi}\sqrt{W}\left[\Phi\big(\frac{-\beta_c}{\sqrt{W}}\big)\right]}$$
$$= \frac{exp\big(- \frac{1}{2W}\sum_{i=1}^{n} \beta_i^2\big)}{\left[\sqrt{8\pi}\sqrt{W}\left(\Phi\big(\frac{-\beta_c}{\sqrt{W}}\big)\right)\right]^n}$$

and the log-likelihood is

$$l(W; \beta_1, \beta_2, ..., \beta_n) = -n \log \sqrt{W} - n \log \left[\Phi\big(\frac{-\beta_c}{\sqrt{W}}\big)\right] - \frac{\sum_{i=1}^{n} \beta_i^2}{2W} + constant \tag{3.1}$$

Using the top hits data, with $n = 148$ and $\beta_c = \log 1.02$ we obtained a MLE for W which is $\hat{W} = 0.0069$ using optim in R. The cumulative distribution function (CDF) for the normal prior when

we condition $|\beta|$ to be greater than a positive β_c is given as follows

$$F(\beta) = \begin{cases} A\,\Phi(\frac{\beta}{\sqrt{W}}) & \text{if } \beta \leq -\beta_c \\ \frac{1}{2} & \text{if } |\beta| < \beta_c \\ \frac{1}{2} + A\left[\Phi(\frac{\beta}{\sqrt{W}}) - \left(1 - \Phi\left(\frac{-\beta_c}{\sqrt{W}}\right)\right)\right] & \text{if } \beta \geq \beta_c \end{cases}$$

where $A = \left(2\Phi\left[\frac{-\beta_c}{\sqrt{W}}\right]\right)^{-1}$.

Figure 3.2 shows the empirical cumulative distribution function (ECDF) for the log odds ratio from the 148 Breast Cancer top hits data and also the CDF for the normal prior (N(0,W)) using $\hat{W} = 0.0069$. The plot shows that the normal prior with $\hat{W} = 0.0069$ does not provide a good fit for the top hits data. The reason behind the poor fit could be that we did not take into consideration of the number of SNPs with very small ORs which have not been picked up in the top hits data when we did the calculation for the MLE.

3.2.2 MLE for W by taking account the number of yet-to-be-discovered SNPs

In Section 3.2.1, the CDF for the estimation of $\hat{W}$ does not fit well with the ECDF for the top hits data as seen in Figure 3.2. As mentioned before, we did not consider the number of unidentified SNPs with very small ORs which were not captured in the top hits data. In order to contributes to a more accurate estimate and have a better MLE that fits the ECDF, we now take account the number of yet-to-be-discovered (YTBD) SNPs in obtaining the MLE for W.

The estimation of W is performed by forming three separate groups in the top hits data (Kulldorff, 1961). The first group consists of $\beta \leq -\beta_c$, the second group consists of $|\beta| < \beta_c$ and the last group consists of $\beta \geq \beta_c$. Let P denote the probability that a Gaussian random variable falls in an interval

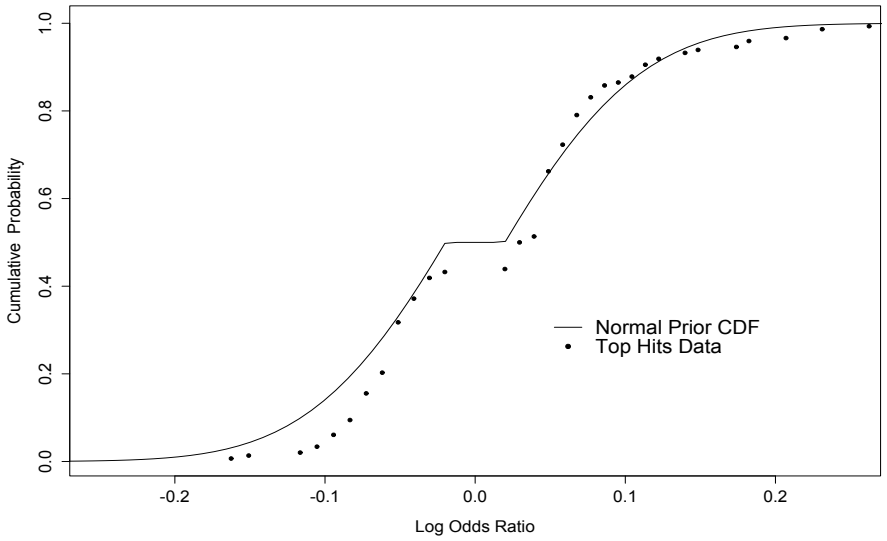


***Figure 3.2*:** Empirical cumulative distribution function (ECDF) for the log odds ratio in 148 Breast Cancer top hits data and the Normal prior (N(0,W)) cumulative distribution function (CDF) using $\hat{W}$ = 0.0069.

$[\beta_0, \beta_1]$ and let F be the distribution function

$$
\begin{aligned}
P &= F(\beta_1) - F(\beta_0) \\
P &= \left[1 - \Phi\left(\frac{-|\beta_1|}{\sqrt{W}}\right)\right] - \Phi\left(\frac{-|\beta_0|}{\sqrt{W}}\right) \\
P &= 1 - 2\,\Phi\left(\frac{-|\beta_c|}{\sqrt{W}}\right); \qquad \text{if} \quad |\beta_1| = |\beta_0| = |\beta_c|
\end{aligned} \tag{3.2}
$$

Following Equation (3.2), the likelihood function of the parameter W can be written as

$$
\begin{aligned}
\mathcal{L}(W; \beta_1, \beta_2, ..., \beta_n) &= P^{n_2} \prod_{j=1}^{n_1+n_3} f(\beta_j) \\
&= \left[1 - 2\,\Phi\left(\frac{-|\beta_c|}{\sqrt{W}}\right)\right]^{n_2} \prod_{j=1}^{n_1+n_3} \frac{1}{\sqrt{2\pi W}} exp\left(-\frac{\beta_j^2}{2W}\right)
\end{aligned}
$$

where,

n_1 = the number of SNPs in the first group ($\beta \leq -\beta_c$)

n_2 = the number of yet-to-be-discovered SNPs in the second group ($|\beta| < \beta_c$)

n_3 = the number of SNPs in the third group ($\beta \geq \beta_c$)

β_j = log odds ratio of the j^{th} observed SNP in the top hits data.

The log-likelihood is

$$l(W;\beta) = n_2 \log\left[1 - 2\,\Phi\left(\frac{-|\beta_c|}{\sqrt{W}}\right)\right] - \left(\frac{n_1+n_3}{2}\right)\log(W) - \frac{1}{2W}\sum_{j=1}^{n_1+n_3}\beta_j^2 + \text{constant.} \qquad (3.3)$$

We compare four different numbers of YTBD SNPs (250,500,750 and 1000) to calculate the values of $\hat{W}$. Using the same 148 Breast Cancer top hits data and the same value for $\beta_c = \log 1.02$, using optim in R, we obtained the values for $\hat{W} = 0.0032, 0.002, 0.0015, 0.0012$ respectively.

Figure 3.3 shows the CDF for every $\hat{W}$ with its respective number of YTBD SNPs and the ECDF for the 148 Breast Cancer top hits data. All the CDF plots with each value of $\hat{W}$ give a slightly better fit to the ECDF of the top hits data compared to the Normal prior CDF when the number of YTBD SNPs was not taken into account.

3.3 The uncertainty in the hyperparameter W

In both cases either we consider the number of YTBD SNPs or not in estimating the MLE, we can see that there are uncertainties in the values of estimated W depending on how many unidentified SNPs are not in the top hits data. Since we estimated W using MLE, the uncertainty of $\hat{W}$ can be translated into a likelihood interval based on the log-likehood function of W.

If the data is a set of n i.i.d observations that depends on W and is distributed according to the likelihood function, l for the true model parameter W, we define

$$\Lambda = -2(l(W^*;\beta) - l(\hat{W}_n;\beta))$$

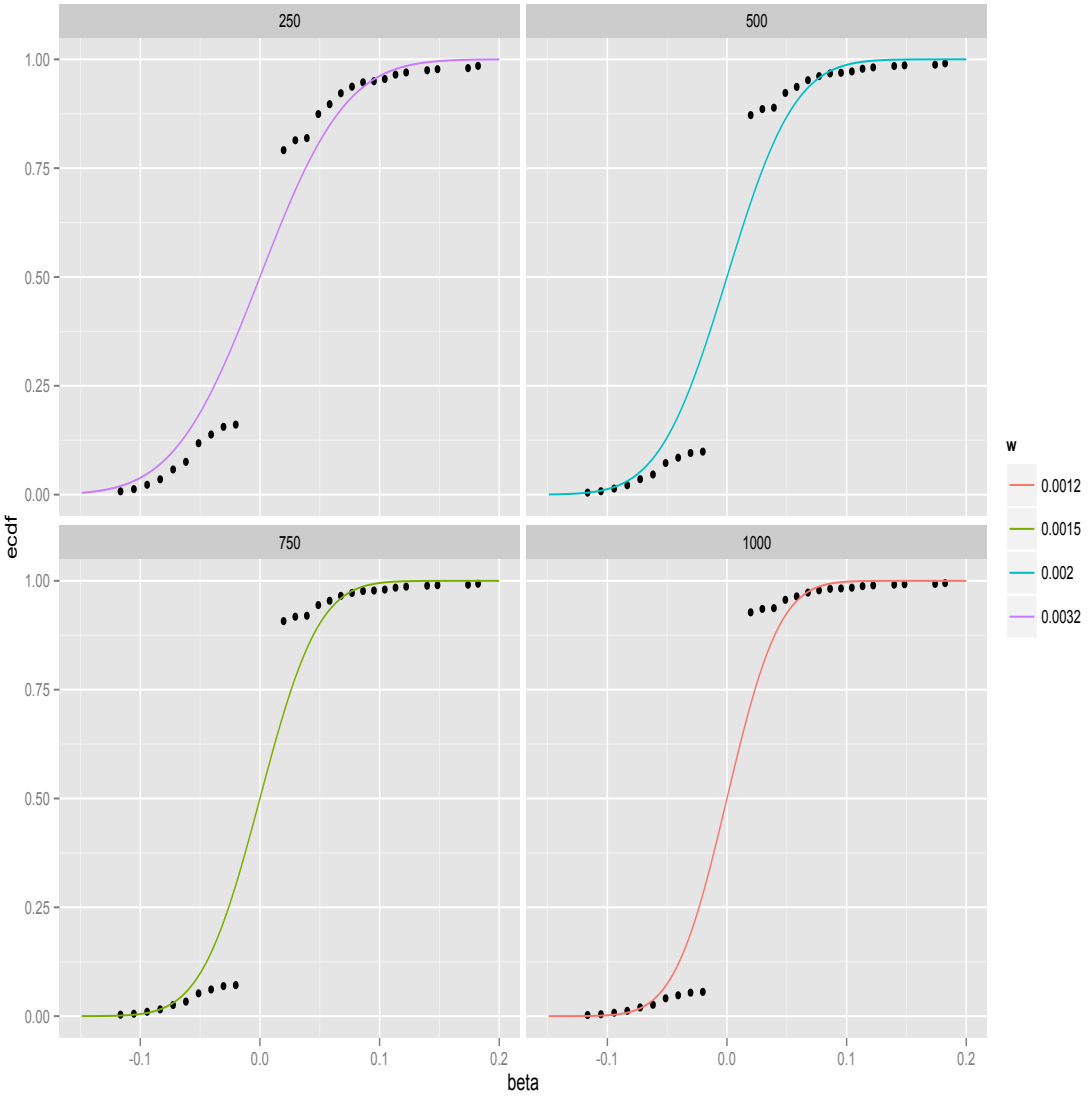


***Figure 3.3*:** Empirical cumulative distribution function of the log odds ratio in 148 Breast Cancer top hits data with different number of yet to be discovered SNPs (250, 500, 750 and 1000) and the cumulative distribution function for Normal prior (N(0,W)) with values of W obtained as the MLE using the respective number of yet-to-be-discovered SNPs.

where W^* is the true value of W and $\hat{W}_n$ is the MLE of W based on a sample of n i.i.d realisation. Wilk's Theorem stated that in a large sample limit, as $n \to \infty$, $\Lambda \sim \chi^2_1$. Therefore we can define a likelihood region as

$$\{\Lambda : l(W;\beta) > l(\hat{W}_n;\beta) - c\}$$

where for a single parameter W, c is defined as $c = \frac{1}{2}\chi^2_{1,0.95} \approx 1.92$ in a 95% confidence region.

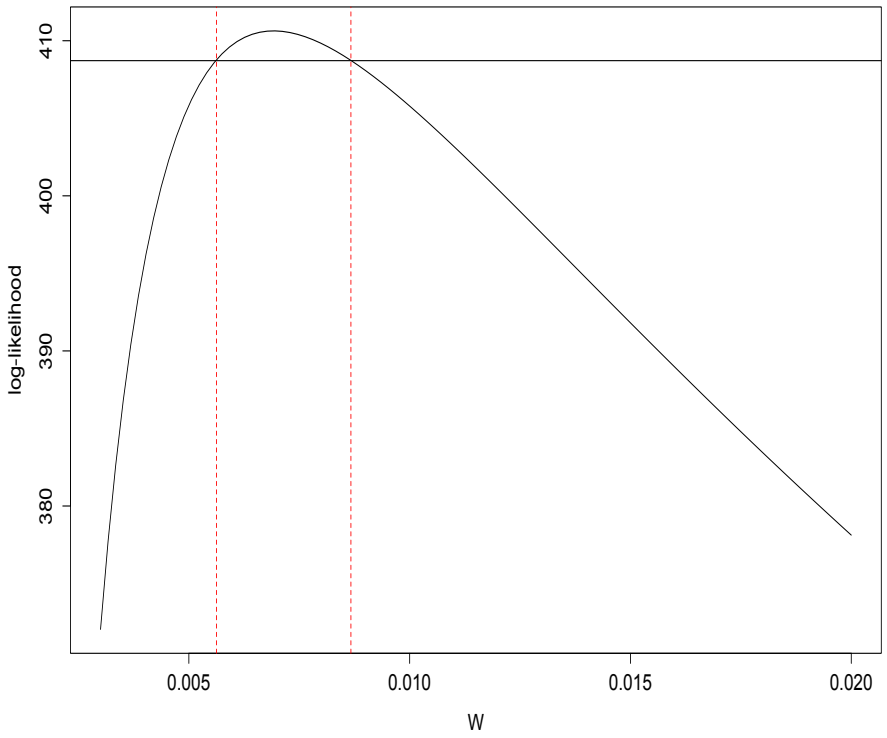
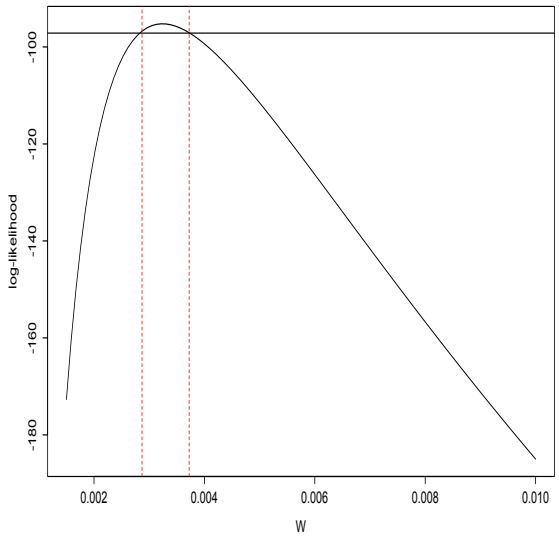
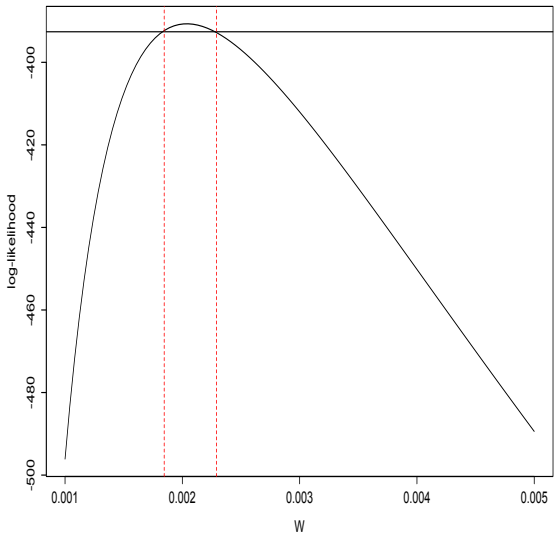


***Figure 3.4*:** Log likelihood interval (in red) limits for W without considering yet-to-be-discovered SNPs based on Wilk's Theorem.

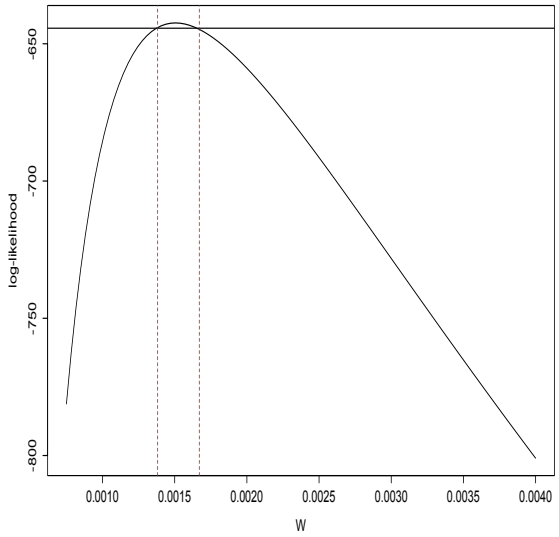
Figures 3.4 and 3.5 show the log likelihood region for the estimated W in both cases either considering or not the number of YTBD SNPs. In Figure 3.4, we did not take into account the number of YTBD SNPs when estimating W where as in Figure 3.5(a), 3.5(b), 3.5(c) and 3.5(d) we did account for the number of YTBD (250, 500, 750, 1000) when estimating W. The likelihood region is between



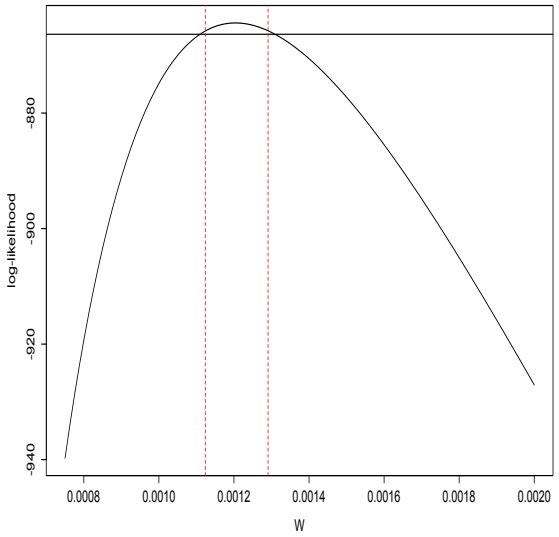
(a) Interval limits for W by taking into account **250** yet-to-be-discovered SNPs



(b) Interval limits for W by taking into account **500** yet-to-be-discovered SNPs



(c) Interval limits for W by taking into account **750** yet-to-be-discovered SNPs



(d) Interval limits for W by taking into account **1000** yet-to-be-discovered SNPs

***Figure 3.5*:** Log likelihood interval (in red) limits for W based on Wilk's Theorem. (a) shows the log likelihood interval (in red) for W by taking account of 250 yet-to-be-discovered (YTBD) SNPs. (b), (c) and (d) shows the log likelihood interval for W by taking account of 500, 750 and 1000 yet-to-be-discovered SNPs respectively.

the limits are in red lines. The lower and upper limits for every estimated W are shown in Table 3.1. The values of W ranges from 0.0011 up to 0.0087.

The uncertainty of W should be accounted for in the calculation when obtaining the Bayes factor using the normal prior. Estimation of W either considering or not the number of YTBD led to a CDF that has a poor fit to the ECDF of the top hits data. The density function for each normal prior with different values of estimated W were plotted on the histogram showing the frequency density of the 148 Breast Cancer top hits data as shown in Figure 3.5. Since the Normal prior does not reflect the distribution of the top hits data, we need a better distribution that has heavier tails and more mass for $|\beta| < \beta_c$. We also need to consider a prior that gives a tractable Bayes factor.

***Table 3.1*:** Maximum likelihood estimation (MLE) for W and its 95% likelihood interval using various number of yet-to-be-discovered (YTBD) SNPs estimated using the 148 top hits data with a critical value of log odd ratio, β_c=log 1.02.

Top Hits data	no. YTBD SNPs	$\hat{W}$	Lower Limit	Upper Limit
148 Top Hits Data	not considered	0.0069	0.0056	0.0087
	1000	0.0012	0.0011	0.0013
	750	0.0015	0.0014	0.0017
	500	0.0020	0.0018	0.0023
	250	0.0032	0.0029	0.0037

3.4 Laplace prior

Previous studies have been using the Gaussian distribution as the prior for the log ORs. Specifying the hyperparameter for the Gaussian prior by using the MLE from analysing the top hits data does not reflect the top hits data since the top hits data has a heavier tail as shown in Figure 3.6. A sensible choice would be a distribution that has more mass close to zero with heavier tails and could have a tractable integral when computing the Bayes factor or posterior distribution.

The Laplace distribution meets the criteria we wanted in reflecting the log ORs from the 148 top hits data. A random variable X have a Laplace distribution La(μ,λ) if its probability density function

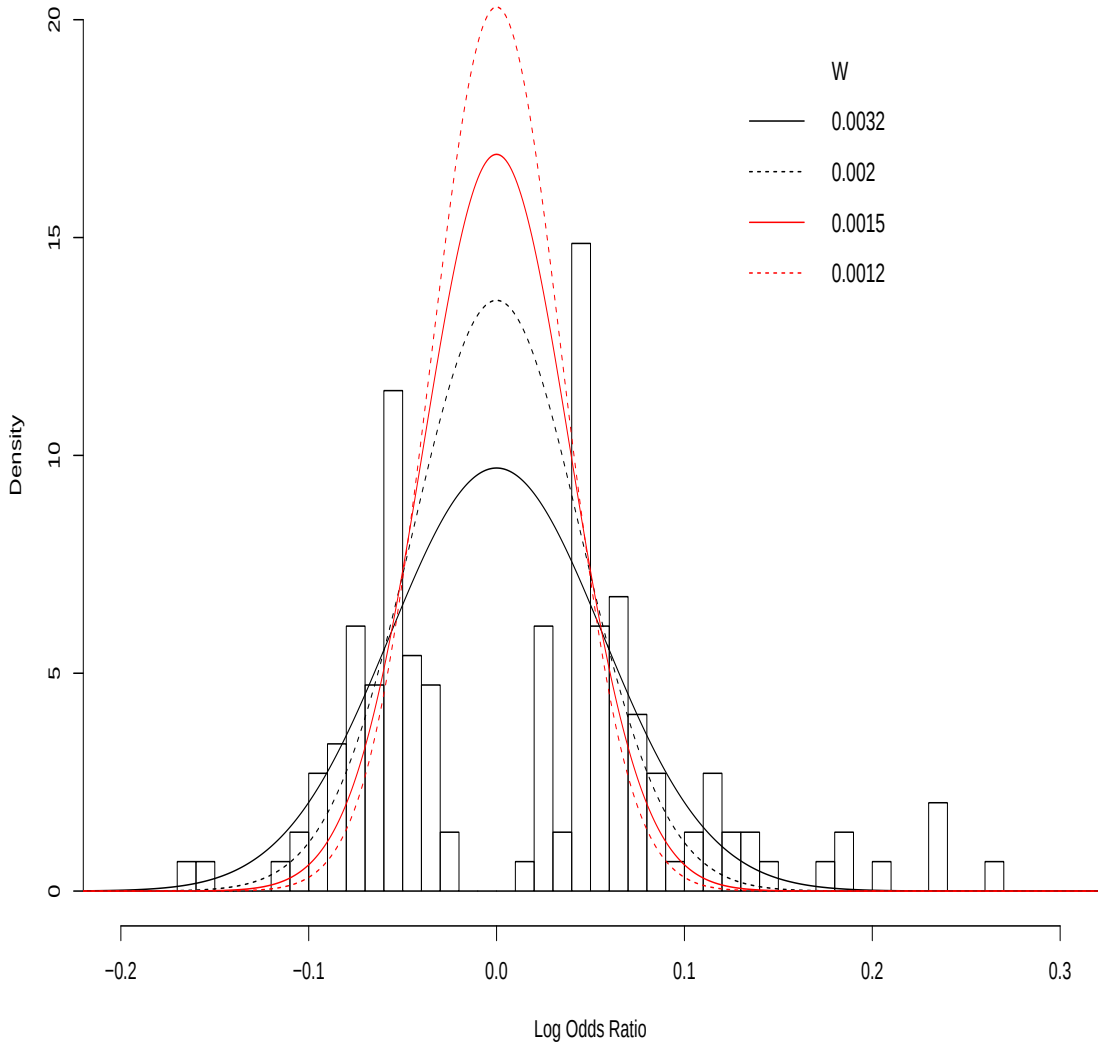


***Figure 3.6*:** Histogram shows the frequency density of the log odds ratio in 148 Breast Cancer top hits data with probability density function (PDF) for Normal prior (N(0,W)) with different values of W (0.0032,0.002,0.0015,0.0012).

is defined as follows

$$f_X(x \mid \mu, \lambda) = \frac{\lambda}{2} exp(-\lambda|x-\mu|). \tag{3.4}$$

The cumulative distribution function is given as

$$F(x) = \begin{cases} \frac{1}{2}\, exp(\lambda(x-\mu)) & \text{if } x < \mu \\ \\ 1 - \frac{1}{2}\, exp(-\lambda(x-\mu)) & \text{if } x \geq \mu \end{cases}$$

The log ORs estimates are from a large sample and so the true logistic likelihood can be replaced with the asymptotic distribution of the MLE which is a Normal distribution with mean (β) and variance (V). Figure 3.7 shows the likelihood function and the probability density for the Normal prior with variance, W = 0.0069. A Laplace probability density function is plotted on top of the likelihood and normal prior (N(0,W)) with the parameter, λ obtained by equating the Laplace variance and the Gaussian variance ($\frac{2}{\lambda^2} = W$).

In this research, a Laplace prior $\beta \sim La(\lambda)$, assuming $\mu = 0$ is chosen for the log odds ratio for the reason that we only consider probability density functions symmetric at 0 to become our prior because the rare alleles are equally likely to be protective as to increase disease risk.

3.5 Estimate MLE for λ from top hits data

The hyperparamter λ requires specification to be used in further calculation in any Bayesian method. The same MLE approach we used in estimating W for the Gaussian prior is used in estimating λ. We will look at both including the number of YTBD SNPs and not including them in estimating the MLE.

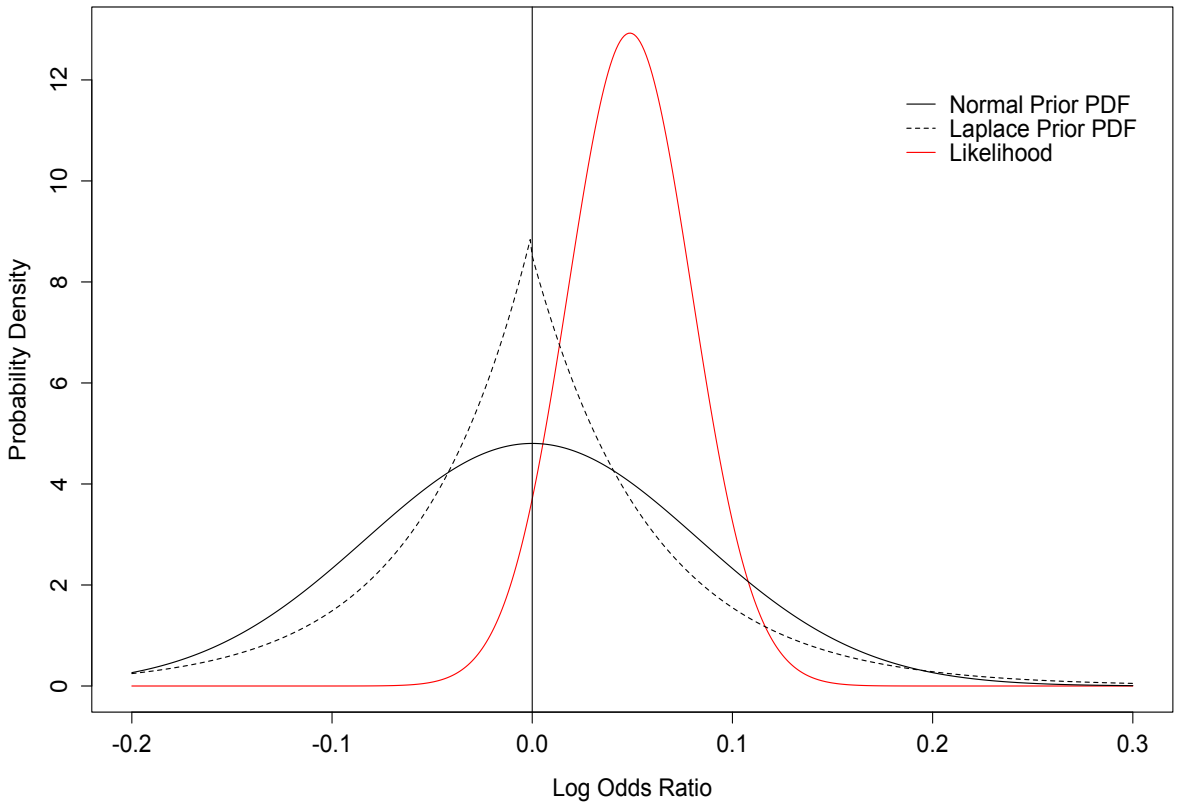


***Figure 3.7*:** Probability densities for priors and likelihood. The likelihood for log odds ratio follows a Normal distribution with mean = 0.05 and variance = 0.03. The Laplace prior uses $\lambda = 17.02$ giving the same variance as the Normal prior ($W = 0.0069$).

3.5.1 MLE for λ without considering the number of yet-to-be-discovered SNPs

In this section, we estimate λ using the MLE from the log ORs in the 148 Breast Cancer top hits data by conditioning the absolute value of the odd ratios to be more than log (β_c). The same reason was given in Section 3.2.1.

Let β be the random variable for the log ORs following a Laplace distribution of λ and $\mu = 0$. The Laplace prior can be written as follows,

$$\beta \sim La(\lambda).$$

If we condition on $\mid \beta \mid$ to have values greater than β_c and β_c is positive, we require

$$\begin{aligned} P(|\beta| > \beta_c \mid \beta \sim La(\lambda)) &= P(\beta > \beta_c \mid \beta \sim Ex(\lambda) \\ &= exp(-\lambda\beta_c) \end{aligned}$$

which leads to a density function

$$f(\beta \,|\, |\beta| > \beta_c, \beta \sim La(\lambda)) = \frac{\frac{\lambda}{2} exp(-\lambda|\beta|)}{exp(-\lambda\beta_c)}$$

and the cumulative distribution function is given as follows

$$F(\beta) = \begin{cases} \frac{1}{2}\ exp(-\lambda(\beta + \beta_c)) & \text{if } \beta \leq -\beta_c \\ \\ \frac{1}{2} & \text{if } -\beta_c < \beta < \beta_c \\ \\ 1 - \frac{1}{2}\ exp(-\lambda(\beta - \beta_c)) & \text{if } \beta \geq \beta_c. \end{cases}$$

If β_i $(1 \leq i \leq n)$ are the observed top hit log odds ratio then the likelihood function is

$$\begin{aligned}\mathcal{L}(\lambda; \beta_1, \beta_2, ..., \beta_n) &= \prod_{i=1}^{n} \frac{\frac{\lambda}{2} exp(-\lambda|\beta_i|)}{exp(-\lambda\beta_c)} \\ &= \prod_{i=1}^{n} \frac{\lambda}{2} exp\bigg(-\lambda[|\beta_i| - \beta_c]\bigg) \\ &= \bigg(\frac{\lambda}{2}\bigg)^n exp\bigg(-\lambda \sum_{i=1}^{n} [|\beta_i| - \beta_c]\bigg)\end{aligned}$$

and the log-likelihood is

$$l(\lambda; \beta_1, \beta_2, ..., \beta_n) = n \log\lambda - \lambda \sum_{i=1}^{n} (|\beta_i| - \beta_c) + constant. \tag{3.5}$$

To estimate the MLE ($\hat{\lambda}$), we set $\frac{dl}{d\lambda} = 0$

$$\begin{aligned}\frac{n}{\lambda} - \sum_{i=1}^{n} (|\beta_i| - \beta_c) &= 0 \\ \hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} (|\beta_i| - \beta_c)}.\end{aligned} \tag{3.6}$$

Estimating λ requires the log ORs from the top hits data, the number of samples (n) in the top hits data and specifying β_c. Using the 148 Breast cancer top hits data with $n = 148$ and β_c = log 1.02 the $\hat{\lambda}$ obtained is equal to 18.3116. The uncertainty in $\hat{\lambda}$ will be considered in Section 3.6.

The Laplace prior CDF with $\lambda = 18.3116$ was plotted together with the Normal prior CDF from Section 3.2.1 on top of the ECDF of the 148 top hits data as shown in Figure 3.8. From the plots in Figure 3.8, the CDF for the Laplace prior shows a better fit to the ECDF for the top hits compared to the Normal prior CDF. Figure 3.9 shows another illustration of the comparison by using the probability density function (PDF) of both priors. The PDFs were plotted on the histogram showing the probability density of the top hits data. In this section, we did not take into consideration the number of YTBD SNPs in estimating lambda. From Figure 3.9, it is predicted that the Laplace prior would have a better

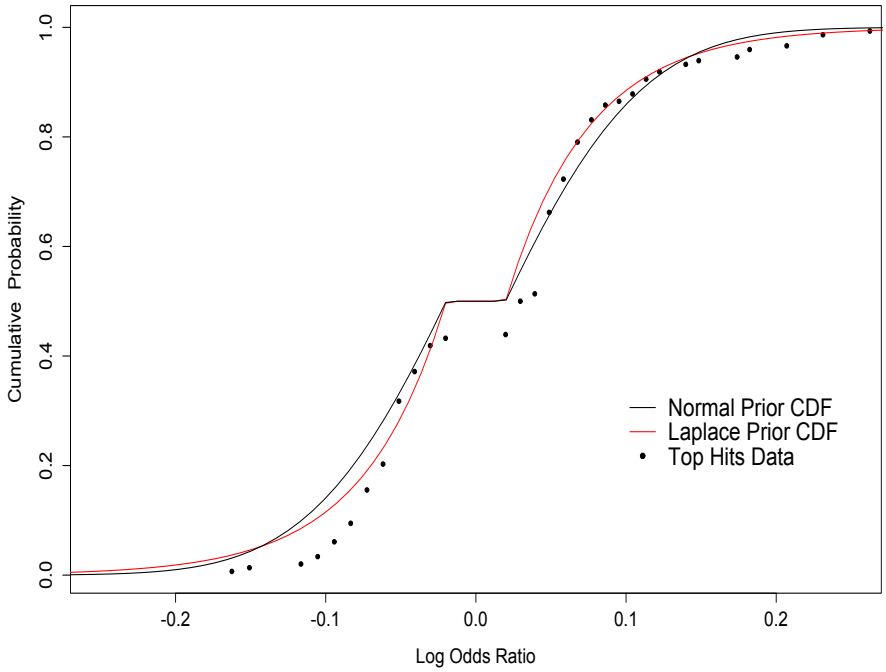


***Figure 3.8*:** Empirical cumulative distribution function (ECDF) for the log odds ratio in 148 Breast Cancer top hits data and the cumulative distribution function (CDF) for Normal prior with $\hat{W} = 0.0069$ and Laplace prior with $\hat{\lambda} = 18.3116$ in cases where we conditioned $|\beta| > \beta_c$.

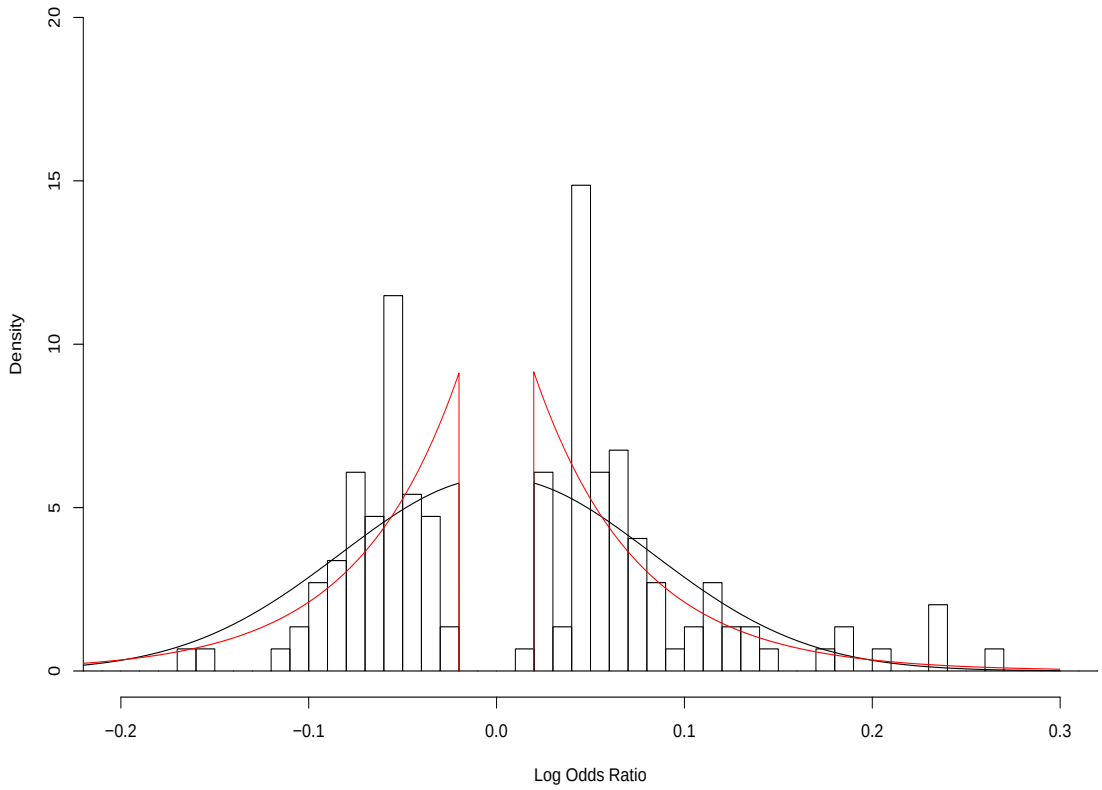


***Figure 3.9*:** The truncated Normal probability density function (PDF) in black and the truncated Laplace PDF in red with the symmetric truncation for $|\beta| < \log 1.02$ in addition to the histogram of the log odds ratio from the 148 Breast Cancer top hits data.

fit to the top hits data if we included the number of YTBD SNPs into the calculation of MLE.

3.5.2 MLE for λ taking account the number of yet-to-be-discovered SNPs

As mentioned before, the Laplace prior may have a better fit to the top hits data if we take into account the number of YTBD SNPs in the calculation of the likelihood of λ. The same concept was applied in Section 3.2.2 where we conditioned the absolute value of the log ORs (β) to have values greater than the critical value of log ORs (β_c).

First, we let P_1 denote the probability of a random variable with a La(λ) distribution falling in an interval ($|\beta| < \beta_c$),

$$\begin{aligned} P_1 &= F(\beta_1) - F(\beta_0) \\ P_1 &= \left[1 - \frac{1}{2}exp(-\lambda|\beta_1)|\right] - \frac{1}{2}exp(-\lambda|\beta_0|) \\ P_1 &= 1 - exp(-\lambda|\beta_c|); \qquad \text{if} \quad |\beta_1| = |\beta_0| = |\beta_c|. \end{aligned} \tag{3.7}$$

The likelihood function of the parameter λ can be written as

$$\begin{aligned} \mathcal{L}(\lambda; \beta_1, \beta_2, ..., \beta_n) &= P_1^{n_2} \prod_{j=1}^{n} f(\beta_j) \\ &= \left[1 - exp(-\lambda|\beta_c|)\right]^{n_2} \prod_{j=1}^{n} \frac{\lambda}{2} exp(-\lambda|\beta_j|) \\ &= \left[1 - exp(-\lambda|\beta_c|)\right]^{n_2} \left(\frac{\lambda}{2}\right)^n exp\left(- \lambda \sum_{j=1}^{n} |\beta_j| \right) \end{aligned}$$

where

n = the number of SNPs in the top hits data

n_2 = the number of yet-to-be-discovered SNPs

β_j= log odds ratio of the j^{th} observed SNP in the top hits data.

The log-likelihood,

$$l(\lambda;\beta) = n_2 \log\left[1 - exp(-\lambda|\beta_c|)\right] + n\log\lambda - \lambda\sum_{j=1}^{n}|\beta_j| + constant \tag{3.8}$$

The calculation for the MLE is undertaken using optimize in R. To estimate the MLE for λ, we set the first derivative of the log-likelihood to become zero and considered the same number of YTBD SNPs as the ones in Section 3.2.2. Using the same top hits data with β_j =log ORs, β_c =log 1.02, $n = 148$ and $n_2 = (1000, 750, 500, 250)$, the values for $\hat{\lambda}$ with different numbers of YTBD SNPs are shown in Table 3.2. In making sure that the values are a maximum value, the second derivatives were checked to be less than zero.

***Table 3.2*:** The Maximum Likelihood Estimation (MLE) for λ estimated using different number of yet-to-be-discovered (YTBD) SNPs.

number of yet-to-be-discovered SNPs	$\hat{\lambda}$
1000	60.47
750	52.27
500	42.41
250	30.05

The CDF for each value of $\hat{\lambda}$ with respective number of YTBD SNPs are shown in Figure 3.10. We can observe that by considering the number of YTBD SNPs into the calculation of the MLE, the Laplace prior has a better fit to the top hits data.

Walters et al. (2019) show how to formally compare the fit of the two priors. Using the hyper-parameter as the MLE, the CDF for both priors were plotted on the ECDF of the top hits data with its respective number of YTBD SNPs (Figure 3.11). The Laplace prior CDF shows a better fit to the ECDF of the top hits data in all scenarios of the different number of YTBD SNPs compared to the CDF for the Normal prior. Therefore, the Laplace prior is the better choice of prior for the log ORs to be used in a Bayesian approach of fine-mapping SNPs. Before we go any further in developing a Bayesian approach using Laplace distribution as the prior, we have to look into the uncertainty of the

hyperparameter.

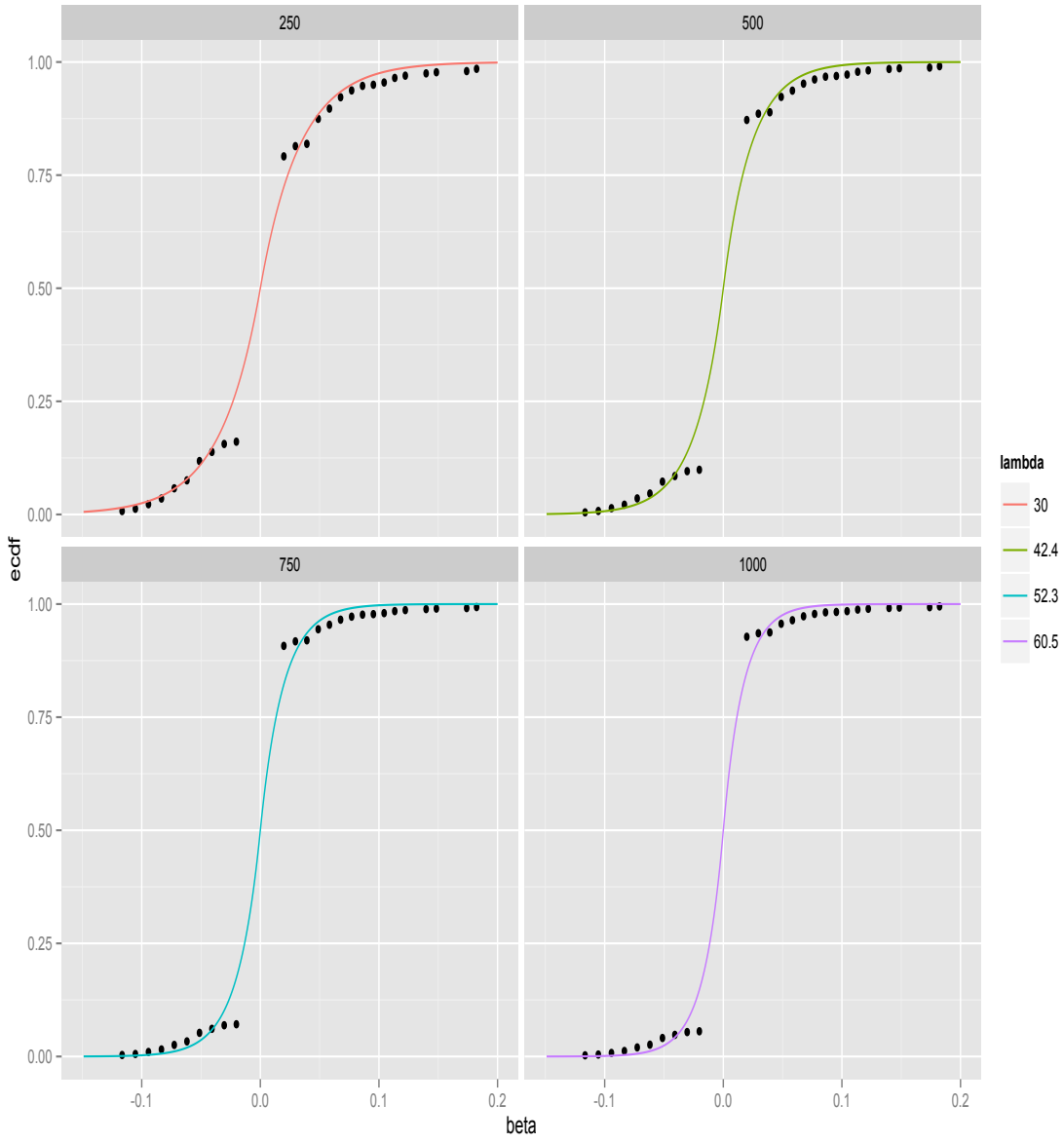


***Figure 3.10*:** Empirical cumulative distribution function (ECDF) showing the log odds ratio in 148 Breast Cancer top hits data with different number of yet to be discovered SNPs (250, 500, 750 and 1000) and the cumulative distribution function (CDF) for Laplace prior with values of λ obtained from the number of yet to be discovered SNPs respectively.

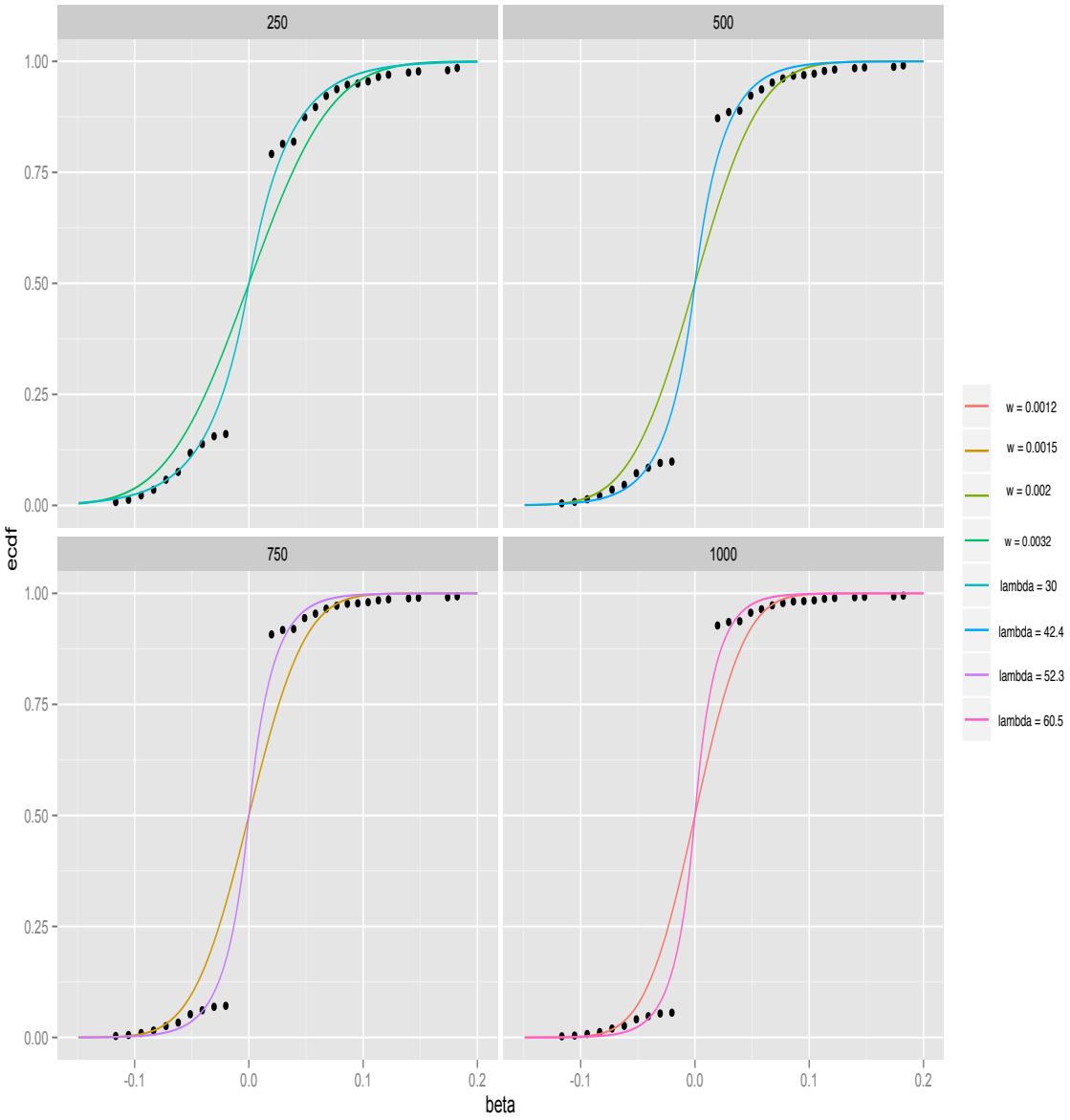


***Figure 3.11*:** Comparing the resemblance of the cumulative distribution function (CDF) for Normal and Laplace prior using their respective estimated hyperparameter to the empirical cumulative distribution function (ECDF) for the log odds ratio in the 148 Breast Cancer top hits data with different number of yet-to-be-discovered (YTBD) SNPs (250, 500, 750 and 1000).

3.6 The uncertainty in $\hat{\lambda}$

Table 3.2 shows how the values of $\hat{\lambda}$ vary according to the choices we made on the number of YTBD SNPs in the top hits data. The uncertainty in $\hat{\lambda}$ has to be accounted for, in order to provide information on plausible values of the parameter. In this section, we will look at the uncertainty of the parameter using the standard error of the parameter rather than the likelihood interval used in Section 3.6.

3.6.1 Standard Error of $\hat{\lambda}$ in two different Breast cancer top hits data

Typically, the MLE has good properties when the sample size is large. To measure the uncertainty of the estimate, we calculate the standard error of the MLE by using the observed information. The standard error (se) for $\hat{\lambda}$ can be obtained by first calculating the variance. Asymptotically we have

$$\hat{\lambda}\ N(\lambda, I^{-1})$$

where

$$I = -\frac{d^2 l}{d\lambda^2}\bigg|_{\lambda=\hat{\lambda}}.$$

In the first case where we did not take into account the number of YTBD SNPs as in Section 3.5.1, following the log-likelihood function in equation 3.5, we obtain the first and second derivatives of the log-likelihood function as follows

$$\frac{dl}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n}(|\beta_i| - \beta_c)$$
$$\frac{d^2 l}{d\lambda^2} = -\frac{n}{\lambda^2}.$$

The variance for $\hat{\lambda}$ when we did not consider the number of YTBD SNPs is therefore

$$Var(\hat{\lambda}) \approx \frac{\lambda^2}{n}. \tag{3.9}$$

Given the log-likelihood in Equation 3.8 in Section 3.5.2 in the case when the number of YTBD SNPs are considered, the derivatives are as follows

$$\frac{dl}{d\lambda} = \frac{n_2|\beta_c|exp(-\lambda|\beta_c|)}{1-exp(-\lambda|\beta_c|)} + \frac{n}{\lambda} - \sum_{j=1}^{n}|\beta_j|$$
$$\frac{d^2l}{d\lambda^2} = -n_2\frac{|\beta_c|^2exp(-\lambda|\beta_c|)}{[1-exp(-\lambda|\beta_c|)]^2} - \frac{n}{\lambda^2}$$

therefore, the variance for $\hat{\lambda}$ in this case is

$$Var(\hat{\lambda}) \approx \left[n_2\frac{|\beta_c|^2exp(-\lambda|\beta_c|)}{[1-exp(-\lambda|\beta_c|)]^2} + \frac{n}{\lambda^2}\right]^{-1}. \tag{3.10}$$

The standard error of $\hat{\lambda}$ were calculated in two different Breast Cancer top hits datasets. Table 3.3 shows the variation of $\hat{\lambda}$ in both datasets. We first look at the top hits data with 148 samples that we currently used, with the critical value of the log odd ratio, $\beta_c = \log 1.02$, $n = 148$ causal SNPs and varying numbers of YTBD SNPs. The values of $\hat{\lambda}$ when we take 95% confidence intervals of the MLE have a range from 15.30 up to 64.15. As we increase the numbers of YTBD SNPs, the uncertainty also increases. Thus, this shows that the estimates of λ are sensitive to the choice of number of YTBD SNPs.

Another top hits dataset was used to estimate λ and to look at its uncertainty using the same method. This dataset was based on the previous availability of Breast Cancer top hits data at an earlier time point having a smaller (68) number of top hits and a bigger critical value of log odds ratio, β_c=log 1.05 (Fachal and Dunning, 2015). Using the same number of YTBD SNPs as before, it is shown in Table 3.3 that the value of $\hat{\lambda}$ varies according to the number of YTBD SNPs with interval extremes

ranging from 16.52 to 52.14. We noticed that, the estimated lambda were not only sensitive on the choices of number of YTBD SNPs, but are also sensitive towards the critical value and the number of top hits.

***Table 3.3*:** Maximum likelihood estimates of λ and its standard error (se) using various number of yet-to-be-discovered (YTBD) SNPs for two different Breast Cancer top hits data. The 148 top hits data have a critical value of log odd ratio, β_c = log 1.02and the 68 top hits data have β_c =log 1.05. The confidence interval (CI) for the MLE is based on 95% confidence.

Top Hits data	no. YTBD SNPs	$\hat{\lambda}$	var($\hat{\lambda}$)	se($\hat{\lambda}$)	Lower CI	Upper CI
148 Top Hits	not considered	18.31	2.27	1.51	15.30	21.32
	1000	60.47	3.53	1.88	56.79	64.15
	750	52.27	3.28	1.81	48.65	55.89
	500	42.41	2.90	1.70	39.01	45.82
	250	30.05	2.30	1.52	27.02	33.09
	50	17.25	1.51	1.23	14.84	19.65
68 Top Hits	not considered	21.81	6.99	2.64	16.52	27.10
	1000	48.56	3.44	1.83	44.98	52.14
	750	43.84	3.26	1.81	40.23	47.45
	500	37.63	3.15	1.78	34.08	41.18
	250	28.48	2.86	1.69	25.09	31.86
	50	15.73	2.14	1.46	12.86	18.60
34 Top Hits	not considered	23.27	8.46	2.91	17.57	28.97
	1000	37.98	3.06	1.75	44.98	52.14
	750	35.03	3.05	1.75	40.30	47.39
	500	31.01	3.03	1.74	34.14	41.11
	250	24.65	2.94	1.71	25.15	31.81
	50	13.84	2.47	1.57	10.76	16.92

3.6.2 Estimate $\hat{\lambda}$ by halving the top hits data

The estimates for λ are sensitive to the number of YTBD SNPs, the number of top hits in the data and the presence of the critical value of log odds ratio. In this section, we would like to look at the sensitivity of the value of $\hat{\lambda}$ when we reduce the number of top hits data to half.

We consider halving the number of top hits data based on quartiles. The first and fourth quartiles were merged together to form another dataset of top hits. The new critical value (β_c) will then be

identified by observing the smallest absolute value of the log ORs. The values of λ will be estimated using the same approach in Section 3.5. The standard error of the estimated λ was calculated to observe the uncertainty of the estimates. The variation of estimated λ with its respective standard error when we halve the 68 top hits data are shown in Table 3.3. The top hits data now has 34 sample size with a new critical value $\beta_c = \log 1.09$ with $\hat{\lambda}$ ranging from 17.57 to 52.14.

As we reduced the number of top hits to half, we could see the estimated λ for all numbers of YTBD SNPs considered decrease. However, the uncertainty of the $\hat{\lambda}$ increases. Furthermore, β_c changes by having a bigger value compared to when we have larger number of top hits. This is because when we have smaller number of top hits, we have less information and this leads to having more uncertainty on the estimated λ. In most cases, as studies gets bigger over time, sample size gets bigger and this will lead to increasing in power which allows them to find causal SNPs with smaller effect size.

Given the results in Section 3.5.2 and by reducing the number of top hits, the value of $\hat{\lambda}$ is sensitive to the number of top hits data, the choices we made on the number of YTBD SNPs and the critical value of the log OR in the top hits data. Therefore, we must take into account the uncertainty in $\hat{\lambda}$ when estimating the hyperparameter for the Laplace prior.

Chapter 4

Univariate Bayesian approaches to fine mapping using the Laplace prior

4.1 A description of the simulated data used

Before we go any further in developing the Bayesian approach with a Laplace prior, we require simulated genotype data with a single causal SNP to test our approach later in this chapter.

The performance of the Bayesian approach will be analysed based on scenarios with different causal SNPs, ORs and sample sizes. The aim is to observe whether the known "true" causal SNP in the simulated data is picked or not as a causal SNP among the candidates ones. The datasets were simulated from the HAPGEN2 software (Su et al., 2011) which produces haplotype sequences based on LD structure in a reference dataset. The reference data used in this case is the European haplotypes of the August 2010 release of the 1000 genome data. These generated sequences depend on the OR specified by the user for the causal SNP. In particular, we simulated data on Chromosome 2 around the CASP8 region between base pair 201666128 and 201866128.

We are interested in testing the Bayesian approach on simulated data with OR varying from 1.08 to 1.15 in two scenarios with rare causal SNPs (MAF=0.09) and common causal SNP (MAF=0.3). 20

datasets were simulated with sample sizes chosen for every scenario to achieve 80% and 60% power where power is given by

$$\begin{aligned}\text{Power} &= P(\hat{\beta} > c_\alpha |\, \hat{\beta} \sim N(\beta, V)) \\ &= P\left(\frac{\hat{\beta}-\beta}{\sqrt{V}} > \frac{c_\alpha - \beta}{\sqrt{V}}\right) \\ &= P\left(Z > \frac{c_\alpha - \beta}{\sqrt{V}}\right) \end{aligned} \tag{4.1}$$

where Z is the standard normal distribution, $c_\alpha =$ is the critical value at level α, $\beta =$ is the log OR, and $V = (n \times \text{MAF} \times (1 - \text{MAF}))^{-1}$ is the variance (Wakefield, 2008).

The sample sizes refer to the total number of cases and controls where we assume that the number of cases is always equal to the number of controls. Table 4.1 shows the values used in the data simulation from HAPGEN2 in scenarios with causal SNPs with different MAF, ORs and sample sizes.

***Table 4.1*:** Simulated data scenarios used in HAPGEN2 with SNPs having different MAF, odds ratio and sample sizes

Power	MAF	Odds Ratio	Sample Size	Number of SNPs
80%	0.3	1.15	10000	193
		1.12	15000	225
		1.08	32000	229
	0.09	1.15	24500	232
		1.12	38000	226
		1.08	81000	236
60%	0.3	1.15	7900	191
		1.12	12100	225
		1.08	26000	229
	0.09	1.15	20300	230
		1.12	31000	225
		1.08	67000	235

As mentioned before, the simulated data from HAPGEN2 were haplotype sequences and we require genotype data to test the Bayesian approach. R codes were created to create the genotype data from the haplotype sequence and to calculate the MAF for all SNPs in the dataset to check if there are extremely rare or monomorphic SNPs. These SNPs need to be removed because the statistical power

is very low for rare and monomorphic SNPs to detect an association with a phenotype. Monomorphic SNPs are not SNPs in our dataset. The simulated data have 412 SNPs in each dataset for every scenario. However, after removing the rare and monomorphic SNPs, each scenario have different number of SNPs as shown in Table 4.1.

4.2 Deriving the posterior distribution and the posterior summaries

Previously in Chaper 2 Section 2.3.4, we mentioned about Wakefield Bayes factor (WBF) as the most common Bayesian approach in GWAS. Another Bayesian approach in identifying causal SNPs is by using summaries from the posterior distribution. Using this approach helps to identify the SNPs by ranking the SNPs based on the posterior summaries. The posterior distribution of the intercept α and the effect size parameter β is given by

$$f(\alpha, \beta \mid \hat{\alpha}, \hat{\beta}) = \frac{f(\hat{\alpha}, \hat{\beta} \mid \alpha, \beta)\ \pi(\alpha, \beta)}{\int f(\hat{\alpha}, \hat{\beta} \mid \alpha, \beta)\ \pi(\alpha, \beta)\ d\beta} \tag{4.2}$$

By using the same assumption and argument to derive WBF in Section 2.3.4 (that cov($\hat{\theta}, \hat{\beta}$) = 0 and that they are jointly normally distributed), the posterior distribution for the reparameterised intercept θ and the effect size β in Equation (4.2) can be written as

$$\begin{aligned} f(\theta, \beta \mid \hat{\theta}, \hat{\beta}) &= \frac{f(\hat{\theta} \mid \theta) f(\hat{\beta} \mid \beta) \pi(\theta) \pi(\beta)}{\int\int f(\hat{\theta} \mid \theta) f(\hat{\beta} \mid \beta) \pi(\theta) \pi(\beta)\ d\theta\, d\beta} \\ &= \frac{f(\hat{\theta} \mid \theta)\, \pi(\theta)}{\int f(\hat{\theta} \mid \theta) \pi(\theta)\ d\theta}\ \frac{f(\hat{\beta} \mid \beta) \pi(\beta)}{\int f(\hat{\beta} \mid \beta) \pi(\beta)\ d\beta}. \end{aligned} \tag{4.3}$$

Our interest is in parameter β, the log odds ratio, thus by taking the integral of Equation (4.3) with

respect to θ, the marginal on β can be obtained as

$$\begin{aligned} f(\beta \mid \hat{\beta}) &= \int f(\theta, \beta \mid \hat{\theta}, \hat{\beta})\ d\theta \\ &= \frac{f(\hat{\beta} \mid \beta)\pi(\beta)}{\int f(\hat{\beta} \mid \beta)\pi(\beta) \quad d\beta}. \end{aligned} \tag{4.4}$$

The posterior density describes our posterior uncertainty about β, the log odds ratio. We have already decided on the prior for the log odds ratio to follow the Laplace distribution $\beta \sim La(\lambda)$ and the likelihood of the data to follow a Gaussian distribution, $\hat{\beta} \mid \beta \sim N(\beta, V)$. The posterior distribution can be derived from Equation (4.4). First, we derive the numerator in Equation (4.4) as follows

$$\begin{aligned} f(\hat{\beta} \mid \beta)\pi(\beta) &= \frac{1}{\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}(\hat{\beta} - \beta)^2 \bigg) \times \frac{\lambda}{2} exp\bigg(-\lambda|\beta| \bigg) \\ &= \frac{\lambda}{2\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}(\hat{\beta} - \beta)^2 - \lambda|\beta| \bigg) \\ &= \frac{\lambda}{2\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}[(\hat{\beta} - \beta)^2 + 2V\lambda|\beta|] \bigg). \end{aligned} \tag{4.5}$$

To obtain the denominator in Equation (4.4), we have

$$\begin{aligned} &\int_{-\infty}^{\infty} f(\hat{\beta} \mid \beta)\pi(\beta) \quad d\beta \\ &= \int_{-\infty}^{\infty} \frac{\lambda}{2\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}[(\hat{\beta} - \beta)^2 + 2V\lambda|\beta|] \bigg) \quad d\beta \\ &= \int_{-\infty}^{0} \frac{\lambda}{2\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}[(\beta - Q_-)^2 + (\hat{\beta}^2 - Q_-^2)] \bigg) d\beta \\ &\quad + \int_{0}^{\infty} \frac{\lambda}{2\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}[(\beta - Q_+)^2 + (\hat{\beta}^2 - Q_+^2)] \bigg) d\beta. \end{aligned} \tag{4.6}$$

The values for Q_- and Q_+ are defined as follows

$$Q_- = \hat{\beta} + V\lambda \tag{4.7a}$$

$$Q_+ = \hat{\beta} - V\lambda. \tag{4.7b}$$

Since the denominator in Equation (4.6) has two parts, integrating over the negative support gives

$$\begin{aligned} f(\hat{\beta} \mid \beta < 0) &= \int_{-\infty}^{0} \frac{\lambda}{2\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}[(\beta - Q_-)^2 + (\hat{\beta}^2 - Q_-^2)]\bigg) d\beta \\ &= \int_{-\infty}^{0} \frac{\lambda}{2\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}(\beta - Q_-)^2\bigg) \times exp\bigg(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\bigg) d\beta \\ &= \frac{\lambda}{2} exp\bigg(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\bigg)\bigg[\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi V}} exp\bigg[-\frac{1}{2V}(\beta - Q_-^2)\bigg] d\beta\bigg] \\ &= \frac{\lambda}{2} exp\bigg(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\bigg)\bigg[\Phi\bigg(\frac{-Q_-}{\sqrt{V}}\bigg)\bigg] \end{aligned} \tag{4.8}$$

where $\Phi(\cdot)$ is the distribution function of a standard normal. A similar approach by completing the squares gives the positive support as follows

$$f(\hat{\beta} \mid \beta > 0) = \frac{\lambda}{2} exp\bigg(-\frac{1}{2V}(\hat{\beta}^2 - Q_+^2)\bigg)\bigg[1 - \Phi\bigg(\frac{-Q_+}{\sqrt{V}}\bigg)\bigg]. \tag{4.9}$$

Thus, Equation (4.6) can be derived by adding Equation(4.8) and Equation (4.9). Fulfilling both positive and negative support gives

$$\begin{aligned} &\int_{-\infty}^{\infty} f(\hat{\beta} \mid \beta)\pi(\beta) \quad d\beta \\ &= \frac{\lambda}{2} exp\bigg(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\bigg)\bigg[\Phi\bigg(\frac{-Q_-}{\sqrt{V}}\bigg)\bigg] \\ &\quad + \frac{\lambda}{2} exp\bigg(-\frac{1}{2V}(\hat{\beta}^2 - Q_+^2)\bigg)\bigg[1 - \Phi\bigg(\frac{-Q_+}{\sqrt{V}}\bigg)\bigg] \\ &= \frac{\lambda}{2} D \end{aligned} \tag{4.10}$$

with

$$D = exp\Big(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\Big)\Big[\Phi\Big(\frac{-Q_-}{\sqrt{V}}\Big)\Big] + exp\Big(-\frac{1}{2V}(\hat{\beta}^2 - Q_+^2)\Big)\Big[1 - \Phi\Big(\frac{-Q_+}{\sqrt{V}}\Big)\Big].$$

To compute the posterior distribution on the negative support of β ($\beta < 0$)

$$\begin{aligned} f(\beta \mid \hat{\beta}) &= \frac{f(\hat{\beta} \mid \beta)\pi(\beta)}{\int_{-\infty}^{\infty} f(\hat{\beta} \mid \beta)\pi(\beta) \quad d\beta} \\ &= \frac{\frac{\lambda}{2\sqrt{2\pi V}} exp\Big(-\frac{1}{2V}(\beta - Q_-)^2\Big) \times exp\Big(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\Big)}{\frac{\lambda}{2}D} \\ &= \frac{E_-}{\sqrt{2\pi V}} exp\Big(-\frac{1}{2V}(\beta - Q_-)^2\Big) \quad \text{with} \quad E_- = \frac{exp\Big(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\Big)}{D} \end{aligned} \tag{4.11}$$

Following a similar approach, we can obtain the positive support $\beta \geq 0$ of the posterior distribution as follows

$$f(\beta \mid \hat{\beta}) = \frac{E_+}{\sqrt{2\pi V}} exp\Big(-\frac{1}{2V}(\beta - Q_+)^2\Big) \quad \text{with} \quad E_+ = \frac{exp\Big(-\frac{1}{2V}(\hat{\beta}^2 - Q_+^2)\Big)}{D} \tag{4.12}$$

Hence, the posterior distribution with a Laplace prior is given by combining Equation (4.11) and Equation (4.12) as follows

$$f(\beta \mid \hat{\beta}) = \begin{cases} \frac{E_-}{\sqrt{2\pi V}} exp\Big(-\frac{1}{2V}(\beta - Q_-)^2\Big) & \text{if } \beta < 0 \\ \\ \frac{E_+}{\sqrt{2\pi V}} exp\Big(-\frac{1}{2V}(\beta - Q_+)^2\Big) & \text{if } \beta \geq 0 \end{cases} \tag{4.13}$$

The posterior density is continuous at $\beta = 0$ since $\lim_{\beta \to 0^-} f(\beta|\hat{\beta}) = \lim_{\beta \to 0^+} f(\beta|\hat{\beta})$. Examples of the posterior densities are shown in Figure 4.1 and 4.2. The posterior summaries obtained from the

posterior distribution give information about the parameter of interest (β). In Bayesian analysis, point estimates (i.e. expected value and median) and interval estimate (i.e. credible interval and highest density posterior interval) are often computed as the summaries of the posterior distribution. Posterior point estimates do not identify SNPs with a true association, but can be used to rank causal SNPs. The interval estimates can be used to determine whether $\beta = 0$ is in some posterior interval.

4.2.1 Posterior expected value

The first posterior point estimate that we will derive from the posterior distribution is the expected value, $E(\beta \mid \hat{\beta})$. The posterior expected value is given by,

$$\begin{aligned} E(\beta \mid \hat{\beta}) &= \int_{-\infty}^{\infty} \beta f(\beta \mid \hat{\beta}) \quad d\beta \\ &= \int_{-\infty}^{0} \beta \left[\frac{E_- \, exp\big(-\frac{1}{2V}(\beta - Q_-)^2 \big)}{\sqrt{2\pi V}} \right] d\beta + \int_{0}^{\infty} \beta \left[\frac{E_+ \, exp\big(-\frac{1}{2V}(\beta - Q_+)^2 \big)}{\sqrt{2\pi V}} \right] d\beta. \end{aligned} \tag{4.14}$$

We need to evaluate the two integrals in Equation 4.14.To calculate the posterior expected value for $\beta < 0$

$$\begin{aligned} &\frac{E_-}{\sqrt{2\pi V}} \int_{-\infty}^{0} \beta \left[exp\big(-\frac{1}{2V}(\beta - Q_-)^2 \big) \right] \quad d\beta \\ &= \frac{E_-}{\sqrt{2\pi V}} \int_{-\infty}^{0} (\beta - Q_-) \left[exp\big(-\frac{1}{2V}(\beta - Q_-)^2 \big) \right] \quad d\beta + \frac{E_- Q_-}{\sqrt{2\pi V}} \int_{-\infty}^{0} exp\big(-\frac{1}{2V}(\beta - Q_-)^2 \big) \quad d\beta \\ &= \frac{E_-}{\sqrt{2\pi V}} \left[-V \big[exp\big(-\frac{Q_-^2}{2V} \big) \big] \right] + E_- Q_- \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi V}} \left[exp\big(-\frac{1}{2V}(\beta - Q_-)^2 \big) \right] \quad d\beta \\ &= \frac{E_-}{\sqrt{2\pi V}} \left[-V \big[exp\big(-\frac{Q_-^2}{2V} \big) \big] \right] + E_- Q_- \left[\Phi\left(\frac{-Q_-}{\sqrt{V}} \right) \right]. \end{aligned} \tag{4.15}$$

Using the same approach in calculating the posterior expected value for the negative support of β, we

can calculate $E(\beta \mid \hat{\beta})$ for $\beta \geq 0$ which gives

$$\int_0^\infty \beta \left[\frac{exp[-\frac{1}{2V}(\beta - Q_+)^2]}{D_+} \right] d\beta = \frac{E_+}{\sqrt{2\pi V}} \left[V\big[exp\big(-\frac{Q_+^2}{2V}\big)\big] \right] + E_+ Q_+ \left[1 - \Phi\left(\frac{-Q_+}{\sqrt{V}} \right) \right]. \quad (4.16)$$

By adding Equation (4.15) and Equation (4.16), the posterior expected value is

$$\begin{aligned} E(\beta \mid \hat{\beta}) = E_+ \left[\sqrt{\frac{V}{2\pi}} \big[exp\big(-\frac{Q_+^2}{2V}\big)\big] + Q_+ \left[1 - \Phi\left(\frac{-Q_+}{\sqrt{V}} \right) \right] \right] \\ - E_- \left[\sqrt{\frac{V}{2\pi}} \big[exp\big(-\frac{Q_-^2}{2V}\big)\big] - Q_- \left[\Phi\left(\frac{-Q_-}{\sqrt{V}} \right) \right] \right]. \end{aligned} \quad (4.17)$$

4.2.2 Posterior median

Another posterior summary is the posterior median, m. Since there are two parts of the posterior distribution, $f(\beta_-|\hat{\beta})$ (when $\beta < 0$) and $f(\beta_+|\hat{\beta})$ (when $\beta \geq 0$), we have to take into consideration where the median might fall, either in the negative region of the posterior or in the positive region. If the median falls in the positive region

$$\begin{aligned} \int_m^\infty f(\beta_+|\hat{\beta}) \quad d\beta &= \frac{1}{2} \\ \int_m^\infty \frac{E_+}{\sqrt{2\pi V}} exp\left(-\frac{1}{2V}(\beta - Q_+)^2 \right) \quad d\beta &= \frac{1}{2} \\ E_+ \left[1 - \Phi\left(\frac{m - Q_+}{\sqrt{V}} \right) \right] &= \frac{1}{2}. \end{aligned}$$

Thus, to compute the posterior median

$$\begin{aligned} E_+ \left[1 - \Phi\left(\frac{m - Q_+}{\sqrt{V}} \right) \right] &= \frac{1}{2} \\ m &= \Phi^{-1}\left(1 - \frac{1}{2E_+} \right) \sqrt{V} + Q_+. \end{aligned} \quad (4.18)$$

A similar approach can be used to calculate posterior median if the posterior median falls in the region where β is less than 0. This gives

$$\int_{-\infty}^{m} f(\beta_-|\hat{\beta}) \quad d\beta = \frac{1}{2}$$
$$\int_{-\infty}^{m} \frac{E_-}{\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}(\beta - Q_-)^2 \bigg) \quad d\beta = \frac{1}{2}.$$

The posterior median in a negative region is

$$m = \Phi^{-1}\bigg(\frac{1}{2E_-}\bigg)\sqrt{V} + Q_-. \tag{4.19}$$

4.2.3 Posterior credible interval

Beside the point estimates discussed in Section 4.2.1 and 4.2.2, an interval estimate, the credible interval (CI), can be obtained from the posterior distribution. A Bayesian credible interval has equal tail probabilities. A general $100(1-\alpha)\%$ credible interval for β, $[\beta_L, \beta_U]$ can be obtained as follows

$$\int_{-\infty}^{\beta_L} f(\beta|\hat{\beta}) \quad d\beta = \int_{\beta_U}^{\infty} f(\beta|\hat{\beta}) \quad d\beta = \frac{\alpha}{2}.$$

where 1 - α is the 'confidence' or 'credible' level.

Before proceeding to calculate the limits in the interval, recall that the posterior obtained in equation (4.13) consists of negative and positive regions. Thus, we lay out several cases on the regions that the interval might fall in. The cases are:

1. lower limit falls in the negative region and upper limit is in the positive region. i.e $\beta_L < 0$ and $\beta_U > 0$.

2. the interval falls in the negative region i.e $[\beta_L,\beta_U] \subset (-\infty, 0]$.

3. the interval falls in the positive region i.e $[\beta_L,\beta_U] \subset [0, \infty)$.

Case 1

We first look into case 1 where $\beta_L < 0$ and $\beta_U > 0$. This is guaranteed to be true when $\min\{\int_{-\infty}^{0} f(\beta_-|\hat{\beta}), \int_{0}^{\infty} f(\beta_+|\hat{\beta})\} > \alpha/2$. The $100(1-\alpha)\%$ credible interval for case 1 is defined as

$$\int_{-\infty}^{\beta_L} f(\beta_-|\hat{\beta}) \quad d\beta = \int_{\beta_U}^{\infty} f(\beta_+|\hat{\beta}) \quad d\beta = \frac{\alpha}{2}.$$

To obtain the lower limit,β_L, we have

$$\begin{aligned}
\int_{-\infty}^{\beta_L} f(\beta_-|\hat{\beta}) \quad d\beta &= \frac{\alpha}{2} \\
\int_{-\infty}^{\beta_L} \frac{E_-}{\sqrt{2\pi V}} exp\bigg(-\frac{1}{2V}(\beta - Q_-)^2 \bigg) \quad d\beta &= \frac{\alpha}{2} \\
E_- \Phi\bigg(\frac{\beta_L - Q_-}{\sqrt{V}} \bigg) &= \frac{\alpha}{2} \\
\beta_L &= \Phi^{-1}\bigg(\frac{\alpha}{2E_-} \bigg)\sqrt{V} + Q_-. \qquad (4.20)
\end{aligned}$$

Using the same approach, we can obtain the upper limit, β_U as follows

$$\beta_U = \Phi^{-1}\bigg(\frac{2E_+ - \alpha}{2E_+} \bigg)\sqrt{V} + Q_+. \qquad (4.21)$$

Case 2

The second case is where the interval $[\beta_L, \beta_U]$ falls in the negative region of the posterior. If $\int_{0}^{\infty} f(\beta_+|\hat{\beta}) \quad d\beta < \alpha/2$, thus the upper limit is certain to be in the negative region, (i.e $\beta_U < 0$). The $100(1-\alpha)\%$ credible interval for case 2 is

$$\int_{-\infty}^{\beta_L} f(\beta_-|\hat{\beta}) \quad d\beta = \int_{\beta_U}^{0} f(\beta_-|\hat{\beta}) \quad d\beta + \int_{0}^{\infty} f(\beta_+|\hat{\beta}) \quad d\beta = \frac{\alpha}{2}.$$

To obtain the upper limit, β_U, we have

$$\begin{aligned}
\int_{\beta_U}^{0} f(\beta_-|\hat{\beta}) \quad d\beta + \int_{0}^{\infty} f(\beta_+|\hat{\beta}) \quad d\beta &= \frac{\alpha}{2} \\
1 - \int_{-\infty}^{\beta_U} f(\beta_-|\hat{\beta}) \quad d\beta &= \frac{\alpha}{2} \\
1 - E_- \Phi\left(\frac{\beta_U - Q_-}{\sqrt{V}}\right) &= \frac{\alpha}{2} \\
\beta_U &= \Phi^{-1}\left(\frac{2-\alpha}{2E_-}\right)\sqrt{V} + Q_-. \qquad (4.22)
\end{aligned}$$

Since in this case, the lower limit, β_L falls in the same region as in case 1, thus the lower limit is the same as the one in Equation (4.20).

Case 3

In case 3, the lower limit falls in the positive region if $\int_{-\infty}^{0} f(\beta_-|\hat{\beta}) \quad d\beta < \alpha/2$. Therefore,

$$\int_{-\infty}^{0} f(\beta_-|\hat{\beta}) \quad d\beta + \int_{0}^{\beta_L} f(\beta_+|\hat{\beta}) \quad d\beta = \int_{\beta_U}^{\infty} f(\beta_+|\hat{\beta}) \quad d\beta = \frac{\alpha}{2}. \qquad (4.23)$$

is the $100(1-\alpha)\%$ credible interval when $[\beta_L,\beta_U] > 0$. The upper limit can be obtain using Equation (4.21). To have a simpler calculation for the lower limit in case 3, Equation 4.23 can be defined as

$$\begin{aligned}
1 - \alpha &= \int_{\beta_U}^{\beta_L} f(\beta_+|\hat{\beta}) \quad d\beta \\
&= E_+\left[\Phi\left(\frac{\beta_U - Q_+}{\sqrt{V}}\right) - \Phi\left(\frac{\beta_L - Q_+}{\sqrt{V}}\right)\right]. \qquad (4.24)
\end{aligned}$$

Rearranging Equation (4.24), we have

$$\beta_L = \Phi^{-1}\left[\Phi\left(\frac{\beta_U - Q_+}{\sqrt{V}}\right) - \left(\frac{1-\alpha}{E_+}\right)\right]\sqrt{V} + Q_+. \qquad (4.25)$$

Figure 4.1 shows the illustration of the 95% posterior credible interval in all cases. The posterior density using $\hat{\lambda} = 60.47$ were obtained from a SNP with different log odds ratio and variances in every cases. The red shaded area is the $\alpha/2$ area, using $\alpha = 0.05$.

4.2.4 Highest density posterior interval

Besides Bayesian credible intervals, the Highest density posterior interval (HDI) is another type of Bayesian interval estimates. The HDI does not have an equal tail probabilities. The idea of HDI is to take a horizontal line and shift it down until the area below the density is $1 - \alpha$. In our case, $100(1 - \alpha)\%$ HDI is defined as

$$\int_{\beta_L}^{\beta_U} f(\beta|\hat{\beta}) \quad d\beta = 1 - \alpha$$

and the density

$$f(\beta_L|\hat{\beta}) = f(\beta_U|\hat{\beta}).$$

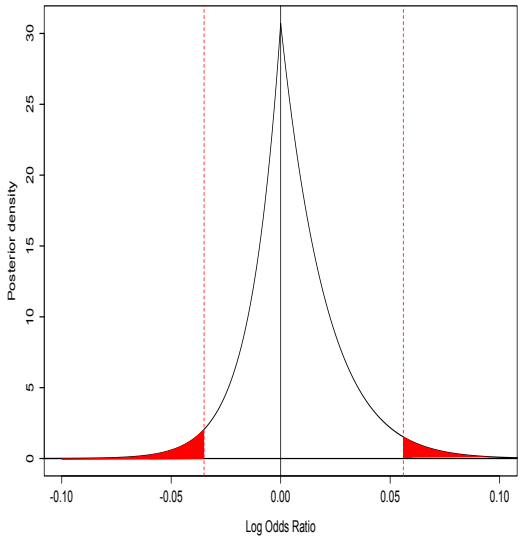
An illustration of a 83% HDI (the white area under the graph) is shown in Figure 4.2. The posterior density was plotted using $\hat{\lambda} = 64.15$, $\hat{\beta} = -0.1016$ and V=0.00091. To be able to determine the limits using HDI, we have to take into account all the cases as discussed in previous Section 4.2.3. We listed several conditions to check if these conditions fulfil all the cases in Section 4.2.3 as follows

1. Case 1: the interval falls in the negative region i.e $[\beta_L,\beta_U] \subset (-\infty, 0)$.

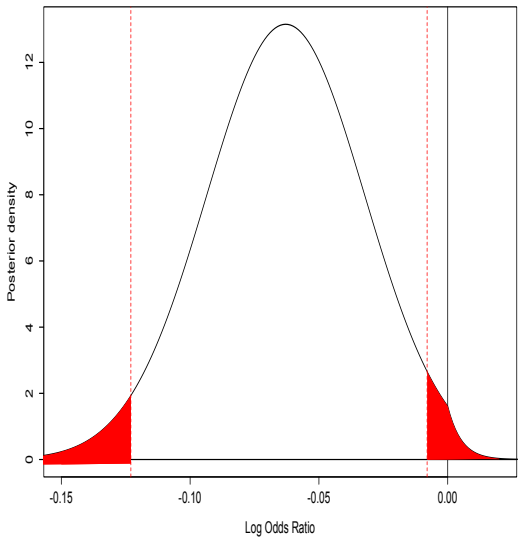
   (a) $Q_- < 0$ and $\int_{2Q_-}^{0} f(\beta_-|\hat{\beta})\ d\beta > 1 - \alpha$

2. Case 2: the interval falls in the positive region i.e $[\beta_L,\beta_U] \subset (0, \infty)$.

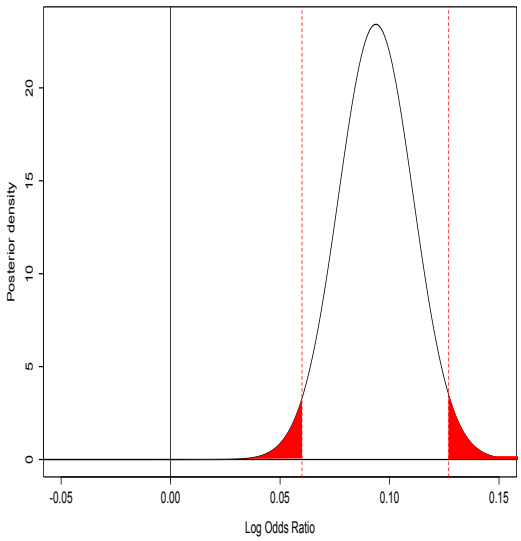
   (a) $Q_+ > 0$ and $\int_{0}^{2Q_+} f(\beta_+|\hat{\beta})\ d\beta > 1 - \alpha$



(a) The 95% posterior credible interval for a specified SNP where the lower limit, β_L is less than zero and upper limit, β_U is greater than zero.



(b) The 95% posterior credible interval for a specified SNP where both lower and upper limits is less than zero



(c) The 95% posterior credible interval for a specified SNP where both lower and upper limits is greater than zero

***Figure 4.1*:** The 95% posterior credible interval for a specified SNP. Posterior densities were plotted using $\hat{\lambda} = 60.47$ with different log odds ratio and variances for all cases. The red shaded area is the $\alpha/2$ area with $\alpha = 0.05$. (a) shows the 95% posterior credible interval for a SNP with log odd ratio, $\hat{\beta}$ = 0.0595 and variance, V = 0.00458. (b) shows the 95% posterior credible interval for a SNP with $\hat{\beta}$ = -0.11988 and V = 0.0009425 and (c) is the 95% posterior credible interval for a SNP with $\hat{\beta}$ = 0.1113 and V = 0.00029.

3. Case 3: the lower limit falls in the negative region and upper limit is in the positive region. i.e $\beta_L < 0$ and $\beta_U > 0$.

   (a) $Q_- < 0$ and $\int_{2Q_-}^{0} f(\beta_-|\hat{\beta})\ d\beta < 1 - \alpha$

   (b) $Q_+ > 0$ and $\int_{0}^{2Q_+} f(\beta_+|\hat{\beta})\ d\beta < 1 - \alpha$

   (c) $Q_- > 0, Q_+ < 0$ i.e $\mid \hat{\beta} \mid < V\lambda$

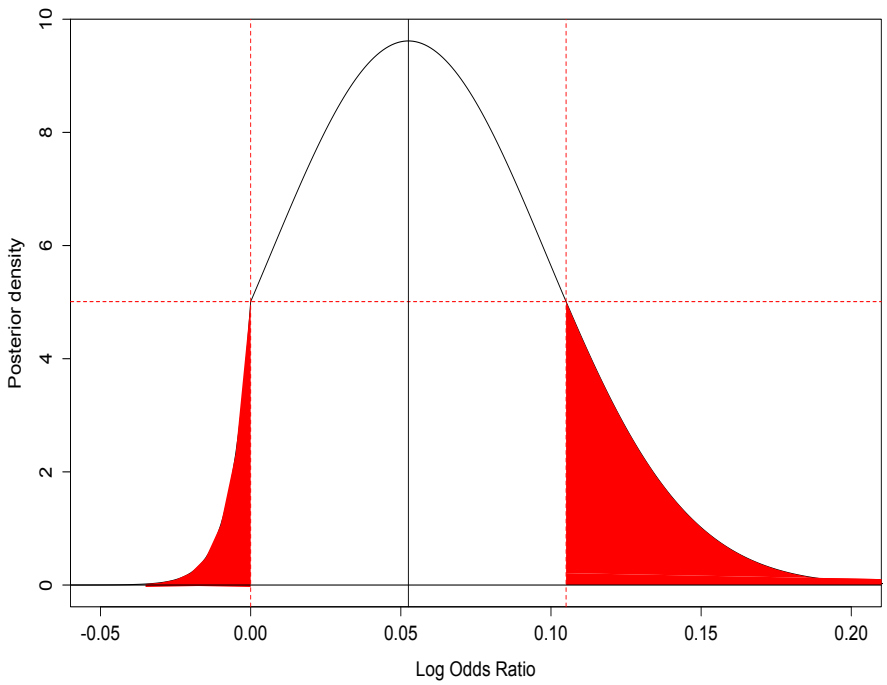


***Figure 4.2*:** The 83% highest density posterior interval for a specified SNP. The posterior density was plotted using $\hat{\lambda} = 64.15$ for a SNP with log odd ratio, $\hat{\beta}$ = -0.01016 and variance, V = 0.00091.

To start with, we first look into the first case, case 1. This case gives condition 1(a) as

$$\begin{aligned}
\int_{2Q_-}^{0} f(\beta_-|\hat{\beta}) \quad d\beta &> 1-\alpha \\
E_- \int_{2Q_-}^{0} \frac{1}{\sqrt{2\pi V}} exp\Big(-\frac{1}{2V}(\beta - Q_-)^2 \Big) \quad d\beta &> 1-\alpha \\
E_- \left[\Phi\left(\frac{0-Q_-}{\sqrt{V}} \right) - \Phi\left(\frac{2Q_- - Q_-}{\sqrt{V}} \right) \right] &> 1-\alpha \\
E_- \left[\Phi\left(\frac{-Q_-}{\sqrt{V}} \right) - \Phi\left(\frac{Q_-}{\sqrt{V}} \right) \right] &> 1-\alpha \\
E_- \left[1 - 2\Phi\left(\frac{Q_-}{\sqrt{V}} \right) \right] &> 1-\alpha. \qquad (4.26)
\end{aligned}$$

Similarly for condition 2(a) in case 2, the condition is

$$\begin{aligned}
\int_{0}^{2Q_+} f(\beta_+|\hat{\beta}) \quad d\beta &> 1-\alpha \\
E_+ \int_{0}^{2Q_+} \frac{1}{\sqrt{2\pi V}} exp\Big(-\frac{1}{2V}(\beta - Q_+)^2 \Big) \quad d\beta &> 1-\alpha \\
E_+ \left[\Phi\left(\frac{2Q_+ - Q_+}{\sqrt{V}} \right) - \Phi\left(\frac{0-Q_+}{\sqrt{V}} \right) \right] &> 1-\alpha \\
E_+ \left[\Phi\left(\frac{Q_+}{\sqrt{V}} \right) - \Phi\left(\frac{-Q_+}{\sqrt{V}} \right) \right] &> 1-\alpha \\
E_+ \left[2\Phi\left(\frac{Q_+}{\sqrt{V}} \right) - 1 \right] &> 1-\alpha. \qquad (4.27)
\end{aligned}$$

In case 3, condition 3(a) and 3(b) follows Equation (4.26) and (4.27) respectively with reversed inequalities. Thus, these conditions fully specify all cases to be true. Once we know where the limits might fall based on all the cases, we can then find the values for the lower and upper limits. We proceed by finding the limits according to cases and conditions listed above.

Condition 1(a)

$$E_{-}\left[\Phi\left(\frac{\beta_U - Q_{-}}{\sqrt{V}}\right) - \Phi\left(\frac{\beta_L - Q_{-}}{\sqrt{V}}\right)\right] = 1 - \alpha$$

we have $\beta_U - Q_{-} = -\beta_L + Q_{-}$, thus

$$E_{-}\left[\Phi\left(\frac{-\beta_L + Q_{-}}{\sqrt{V}}\right) - \Phi\left(\frac{\beta_L - Q_{-}}{\sqrt{V}}\right)\right] = 1 - \alpha.$$

Since $\Phi\left(\frac{-\beta_L + Q_{-}}{\sqrt{V}}\right) + \Phi\left(\frac{\beta_L - Q_{-}}{\sqrt{V}}\right) = 1$, we have

$$E_{-}\left[2\Phi\left(\frac{-\beta_L + Q_{-}}{\sqrt{V}}\right) - 1\right] = 1 - \alpha$$

$$\beta_L = Q_{-} - \sqrt{V}\,\Phi^{-1}\left(\frac{1 - \alpha - E_{-}}{2E_{-}}\right). \tag{4.28}$$

We mentioned $\beta_U - Q_{-} = -\beta_L + Q_{-}$, rearranging the equation will give the upper limit in this case as

$$\beta_U = -\beta_L + 2Q_{-} \tag{4.29}$$

Condition 2a

To obtain the upper and lower limit for condition 2(a), we noticed the similarity in conditioned 2(a) and condition 1(a). Therefore, the only changes needed are in Equation (4.28) in which E_{-} and Q_{-} are change to E_{+} and Q_{+} respectively.

Condition 3a

$$\int_{\beta_L}^{0} f(\beta_-|\hat{\beta}) \quad d\beta + \int_{0}^{\beta_U} f(\beta_+|\hat{\beta}) \quad d\beta = 1-\alpha$$

$$E_-\left[\Phi\left(\frac{-Q_-}{\sqrt{V}}\right) - \Phi\left(\frac{\beta_L - Q_-}{\sqrt{V}}\right)\right] + E_+\left[\Phi\left(\frac{\beta_U - Q_+}{\sqrt{V}}\right) - \Phi\left(\frac{-Q_+}{\sqrt{V}}\right)\right] = 1-\alpha. \tag{4.30}$$

To obtain the values for the interval $[\beta_L, \beta_U]$, we also have

$$f(\beta_- = \beta_L|\hat{\beta}) = \frac{E_-}{\sqrt{2\pi V}}\, exp\bigg(-\frac{1}{2V}(\beta_L - Q_-)^2\bigg) = h_1$$
$$f(\beta_+ = \beta_U|\hat{\beta}) = \frac{E_+}{\sqrt{2\pi V}}\, exp\bigg(-\frac{1}{2V}(\beta_U - Q_+)^2\bigg) = h_2$$

where h_1 is the height of the density at β_L and h_2 is the height of the density at β_U. We want a $100(1-\alpha)\%$ HDI for β at the same height, i.e $h_1 = h_2 = h$. Therefore, $f(\beta_L|\hat{\beta}) = f(\beta_U|\hat{\beta}) = h$ and hence

$$\frac{E_-}{\sqrt{2\pi V}}\, exp\bigg(-\frac{1}{2V}(\beta_L - Q_-)^2\bigg) = \frac{E_+}{\sqrt{2\pi V}}\, exp\bigg(-\frac{1}{2V}(\beta_U - Q_+)^2\bigg)$$
$$\frac{E_-}{E_+}\, exp\bigg(-\frac{1}{2V}(\beta_L - Q_-)^2\bigg) = exp\bigg(-\frac{1}{2V}(\beta_U - Q_+)^2\bigg). \tag{4.31}$$

from Equation (4.11) and (4.12), we have

$$\frac{E_-}{E_+} = \frac{exp\bigg(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\bigg)}{exp\bigg(-\frac{1}{2V}(\hat{\beta}^2 - Q_+^2)\bigg)}$$
$$= exp\bigg(-\frac{1}{2V}(Q_+^2 - Q_-^2)\bigg). \tag{4.32}$$

We substitute Equation (4.32) into Equation (4.31),

$$exp\bigg(-\frac{1}{2V}(Q_+^2-Q_-^2)\bigg)\,exp\bigg(-\frac{1}{2V}(\beta_L-Q_-)^2\bigg)=exp\bigg(-\frac{1}{2V}(\beta_U-Q_+)^2\bigg)n$$

$$(\beta_L-Q_-)^2-(Q_-^2-Q_+^2)=(\beta_U-Q_+)^2. \tag{4.33}$$

Expanding Equation (4.33) gives

$$\beta_L^2-2Q_-\beta_L-(\beta_U^2-2Q_+\beta_U)=0.$$

To obtain the lower limit, we have

$$\beta_L=\frac{2Q_-\pm\sqrt{4Q_-^2+4(\beta_U^2-2Q_+\beta_U)}}{2}$$

$$\beta_L=Q_-\pm\sqrt{Q_-^2+\beta_U^2-2Q_+\beta_U}. \tag{4.34}$$

In this case, $\beta_U>0$ and $Q_+<0$, therefore $\sqrt{Q_-^2+\beta_U^2-2Q_+\beta_U}>|Q_-|$.

If $\beta_L=Q_-+\sqrt{Q_-^2+\beta_U^2-2Q_+\beta_U}$, thus $\beta_L>0$. This is not the solution for this case. The only solution to obtain a β_L less than zero is by choosing

$$\beta_L=Q_--\sqrt{Q_-^2+\beta_U^2-2Q_+\beta_U}. \tag{4.35}$$

From Equation (4.35), we substitute $\beta_L-Q_-=-\sqrt{Q_-^2+\beta_U^2-2Q_+\beta_U}$ into Equation (4.30). This gives

$$E_-\bigg[\Phi\bigg(\frac{-Q_-}{\sqrt{V}}\bigg)-\Phi\bigg(-\sqrt{\frac{Q_-^2+\beta_U^2-2Q_+\beta_U}{V}}\bigg)\bigg]+$$

$$E_+\bigg[\Phi\bigg(\frac{\beta_U-Q_+}{\sqrt{V}}\bigg)-\Phi\bigg(\frac{-Q_+}{\sqrt{V}}\bigg)\bigg]=1-\alpha \tag{4.36}$$

To obtain β_U, Equation 4.36 can be solve numerically using uniroot in R. Using β_U, we then solve for β_L in Equation (4.35).

Condition 3(b)

Similar to condition 3(a), we have $\frac{E_-}{E_+}$ from Equation (4.31) which leads to Equation (4.33). Expanding Equation (4.33) gives

$$\beta_U^2 - 2Q_+\beta_U - (\beta_L^2 - 2Q_-\beta_L) = 0$$

and hence $\beta_U = Q_+ \pm \sqrt{Q_+ + \beta_L^2 - 2Q_-\beta_L}$. In this case, $\beta_L < 0$ and $Q_- > 0$, so $\sqrt{Q_+^2 + \beta_L^2 - 2Q_-\beta_L} > |Q_+|$.

$\beta_U = Q_+ - \sqrt{Q_+ + \beta_L^2 - 2Q_-\beta_L}$ gives a negative value for β_U which is not a solution in this case. In order to have a solution that satisfies β_U greater than zero, the only solution is

$$\beta_U = Q_+ + \sqrt{Q_+ + \beta_L^2 - 2Q_-\beta_L}. \tag{4.37}$$

To find the limits, from Equation (4.37) we substitute $\beta_U - Q_+ = \sqrt{Q_+ + \beta_L^2 - 2Q_-\beta_L}$ in Equation (4.30) to obtain values for β_L and thus solve for β_U.

Condition 3(c)

To solve for the limit in condition 3(c), we can use the same argument as in condition 3(a). We have lower limit β_L from Equation (4.34). In this case β_L can not have a positive value, hence β_L with a positive square root is not the solution for this case. The only solution is when we have β_L with a negative square root as in Equation (4.35). This gives $\beta_L < 0$. The solution for [β_L,β_U] can be obtained as the ones in condition 3(a) by solving Equation (4.35) and (4.36).

4.2.5 Receiver operating characteristic (ROC) curve.

A receiver operating characteristic (ROC) curve is an essential tool to visualise the performance of a classifier. A ROC curve is based on two basic measures; sensitivity and specificity. Sensitivity, also known as True Positive Rate (TPR), is represented on the y-axis against the False Positive Rate (FPR) on the x-axis which can be calculated as 1 - Specificity. Table 4.2 shows the 2 by 2 contingency table to illustrate the four possible outcomes from a classifier. TPR can be estimated by dividing the number of true positives by P. Whereas FPR is the number of false positives divided by N. To create a ROC curve, ROC points (a point with a pair of FPR and TPR values) of a classifier are connected by a straight line which starts at (0,0) and ends at (1,1) in the ROC space. FPR and TPR pairs are determined by varying the threshold of the classifier.

***Table 4.2*:** 2x2 contingency table to illustrate the four possible outcomes from a classifier and an instance.

		Predicted Condition		
		Positive	Negative	Total
True Condition	Positive	True Positive (TP)	False Negative (FN)	P
	Negative	False Positive (FP)	True Negative(TN)	N

Interpreting the performance of the classifier comes from understanding where the ROC curve lies in the ROC space. The diagonal line, $y = x$ represent a classifier with a random performance level which separates the ROC space into two areas. ROC curves which appear in the lower right triangle indicate a poor performance level (worse than guessing) and the ones that appear in the top left indicate good performance level. Figure 4.3 shows an example of a ROC curve plot for a single dataset with a single causal SNP where each SNP is assigned some numerical value, which depends on the analysis method.

ROC curves for multiple dataset

Averaging ROC curves is a method to obtain a ROC curve for multiple datasets. This can be done by merging the datasets together into one large dataset. However, merging these datasets does not

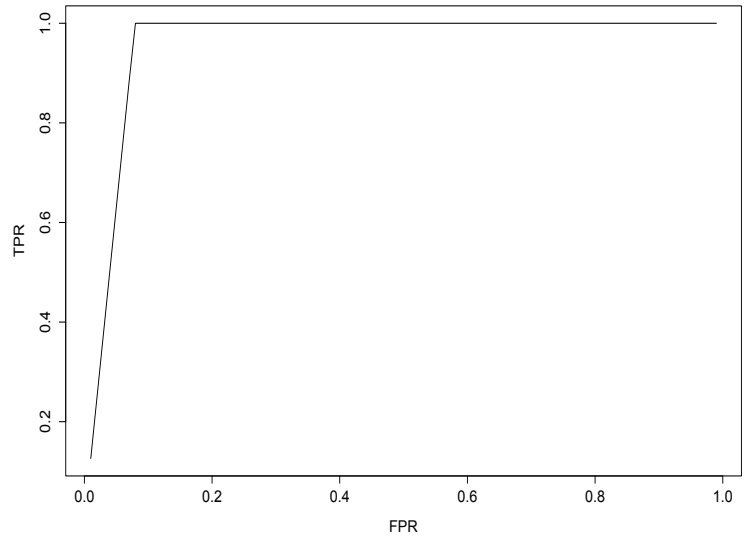


***Figure 4.3*:** Receiver operating characteristic (ROC) curve for one dataset with a single causal SNP.

provide any measure of variance among the datasets. To overcome this, Fawcett (2006) came up with methods that averages the datasets in different ways. The two methods are vertical averaging and threshold averaging. In vertical averaging, we average the TPR throughout the datasets at fixed FPRs meanwhile in threshold averaging, we average both TPR and FPR across datasets at a given fixed threshold.

4.2.6 ROC curves comparing posterior summaries in univariate analyses

The posterior summaries were obtained by using the output from a simple logistic regression analysis for a particular SNP; the parameter estimates ($\hat{\beta}$) and the standard error ($\sqrt{V}$). Since we need to specify a value for lambda, we used $\hat{\lambda} = 64.15$ obtained from Section 3.6.1 since this estimate of lambda is the maximum $\hat{\lambda}$ from the 95% CI for $\hat{\lambda}$ obtained from the top hits data with 148 samples and β_c= log 1.02.

The rankings for the interval estimates were based on the probability contained within the largest interval that does not contain 0. For posterior credible intervals, the tail probabilities must be equal,

so the interval probability is

$$1 - 2 \times \min\{P(\beta < 0|\hat{\beta}), P(\beta \geq 0|\hat{\beta})\}.$$

While in posterior HDI, the areas which does not include 0 are not necessarily the same. The area depends on the intersection points with the horizontal line when the line crosses the density where $\beta = 0$.

To evaluate the performance among the posterior expected value, posterior median, posterior credible interval and posterior HDI in ranking the SNPs, we used the 20 simulated datasets to obtain $\hat{\beta}$ and $\sqrt{V}$ for each SNP in each dataset using univariate logistic regression analysis and hence calculated values for the true positive rates (TPR) and false positive rates (FPR) for every posterior summary as the threshold was varied. The performance for each posterior summary can be observed using ROC curves using the method of vertical averaging. The scenarios we consider were given in Table 4.1.

Figure 4.4 shows the ROC curves for different posterior summaries based on the simulated data for FPR ≤ 0.04. Observing the ROC curves for a single common causal SNP (MAF=0.3), the posterior HDI shows the best performance among other posterior summaries in the scenario where we have a high OR (OR=1.15), however, the posterior summaries have very similar performance with OR=1.12 and OR=1.08.

For a single rare causal SNP with MAF=0.09, the performance of the posterior summaries has different results in performance for every scenario. We first look at the case with OR=1.15 and 1.12. It is interesting to see that the posterior summaries using the point estimates, i.e posterior mean and median, have the same performance. A similar result was also shown by the interval estimates. In the scenario with OR=1.15, the point estimates perform marginally better than the interval estimates. However, in the scenario with OR=1.12, it shows the reverse. Each posterior summary in the scenario with OR=1.08 has a different performance with the posterior HDI performing the best.

Comparing the performance of all the posterior summaries in all six scenarios, we could see very

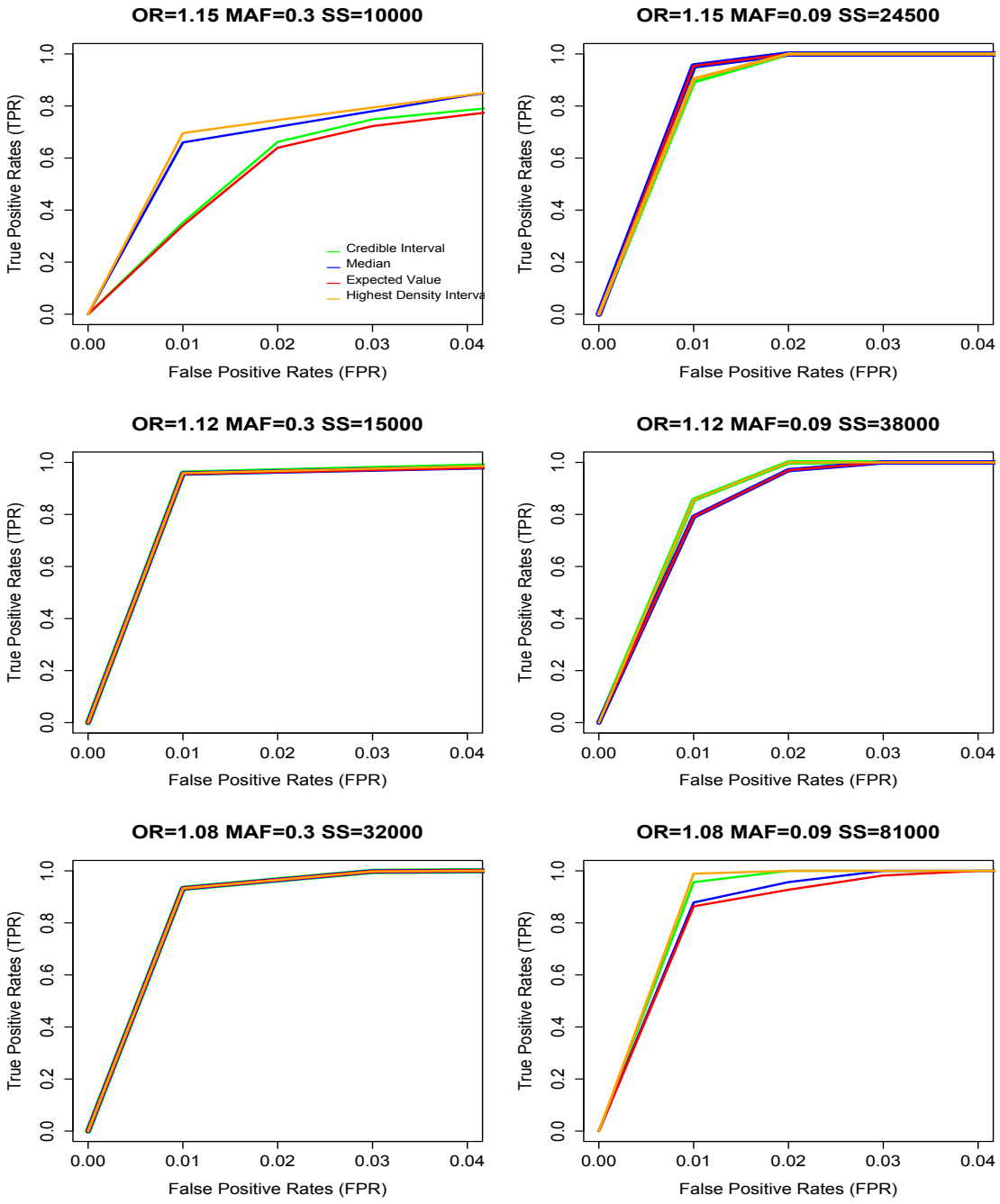


***Figure 4.4*:** Receiver operating characteristic (ROC) curves comparing the performance of posterior expected value, posterior median and posterior credible interval in ranking SNPs. The prior for the posterior distribution has $\hat{\lambda} = 64.15$. All rankings were carried out on 20 simulated datasets from HAPGEN using a single rare causal SNP (MAF=0.09) and a single common causal SNP(MAF=0.3) with three different odd ratios (OR = 1.08, 1.12, 1.15). The sample size (SS) for each scenario depends on 80% power.

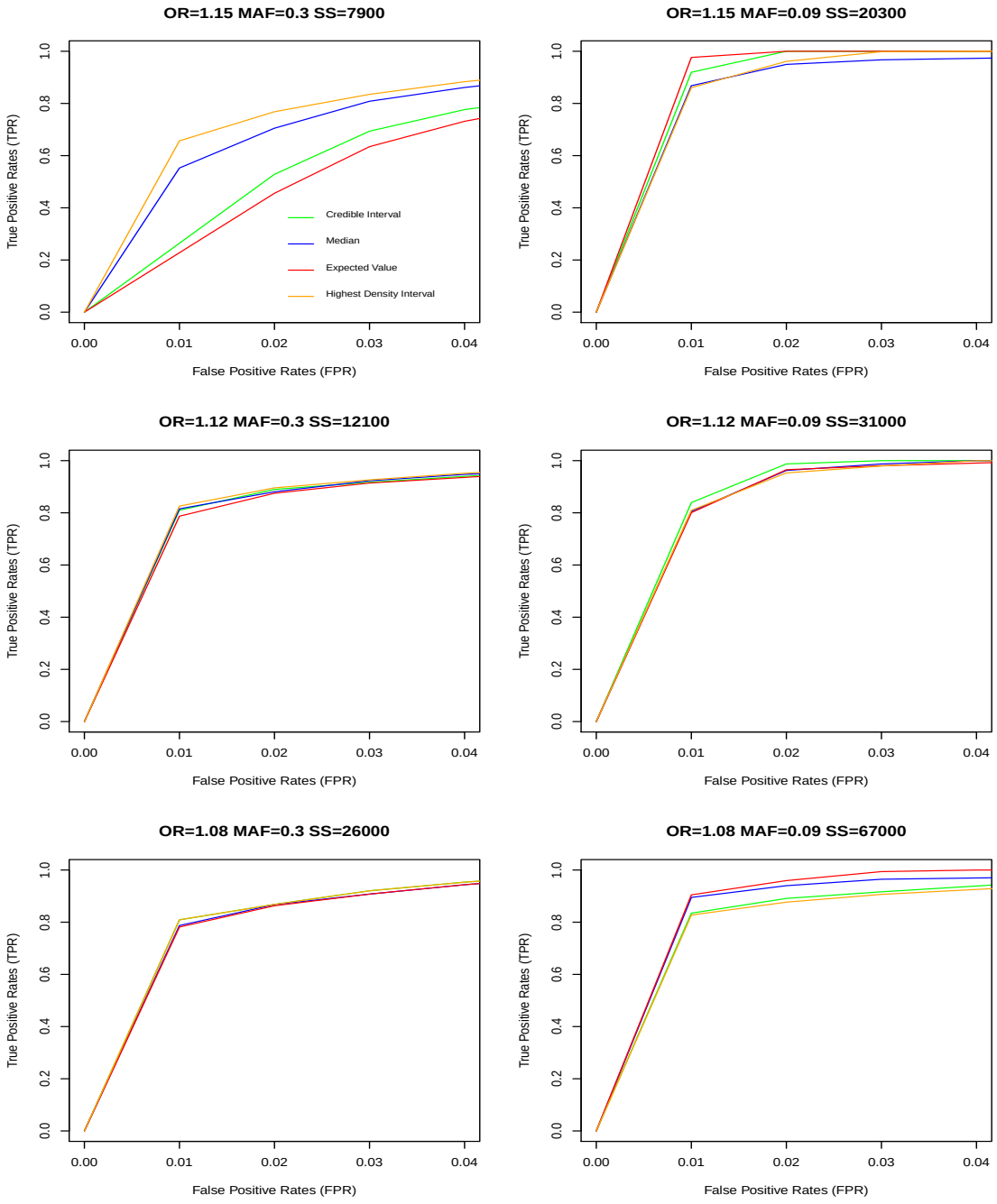


***Figure 4.5*:** Receiver operating characteristic (ROC) curves comparing the performance of posterior expected value, posterior median and posterior credible interval in ranking SNPs. The prior for the posterior distribution has $\hat{\lambda} = 64.15$. All rankings were carried out on 20 simulated datasets from HAPGEN using a single rare causal SNP (MAF=0.09) and a single common causal SNP(MAF=0.3) with three different odd ratios (OR = 1.08, 1.12, 1.15). The sample size (SS) for each scenario depends on 60% power.

little difference between them apart from the scenario with a single common causal SNP with OR=1.5. Most of the scenarios, the TPRs for all the posterior summaries go up to one (all of the causal SNP are caught) by the time FPR is 0.04. This is an exception for the first scenario (OR = 1.15 and MAF = 0.3). In this scenario, at FPR equal to 0.04, all the posterior summaries caught about 75% to 85% of the causal SNPs. Although with the same statistical power (80%) of picking up causal SNPs, there are still differences in performance for all posterior summaries.

Generally, the classification (ranking) performance of the posterior summaries is very good at 80% power in all scenarios. Among the four posterior summaries, we could see that in most scenarios, atleast two posterior summaries performed similarly. Only in two scenarios (a single common SNP with OR=1.15 and a single rare causal SNP with OR = 1.08), the performance of each posterior summary performed differently resulting in posterior HDI having the best performance among all. However, after considering all scenarios, it is difficult to come with a conclusive choice of the best posterior summaries.

In order to come up with a conclusive choice, we decided to reduce the statistical power. We predict that by reducing the statistical power, the performance of each posterior summaries might be distinguishable. Thus, we reduce the sample size for every scenario to yield 60% power. All the posterior summaries were recalculated using the new simulated data with reduced sample sizes. However, the ORs and MAFs remained the same for all 6 scenarios.

Figure 4.5 shows the ROC curves for the posterior summaries in different scenarios with new sample sizes. The performance of the posterior summaries is now marginally more different compared to 80% power but there is still much overlap. However, in every scenario, all posterior summaries take longer to reach TPR equals to one. Among the ORs with a single common causal SNP, the posterior HDI shows the best performance compared to other posterior summaries. A different result in performance is shown with a single rare causal SNP as the posterior HDI show the poorest performance. It is interesting to see that the performance for each posterior summary shows dissimilar results in every OR with a single rare causal SNP.

It appears that there is no obvious result in showing which posterior summary has the best performance in ranking the SNPs in various scenario. However, we decided to pick posterior HDI as the best posterior summary since it appears to show a better performance in many of the scenarios. We now examine the performance of the Laplace Bayes factor that places a Laplace prior on the log OR of all SNPs and compare it with the performance of the posterior HDI.

4.3 Laplace Bayes factor

We previously used posterior summaries to rank the SNPs for association with the phenotype and concluded that the best summary among them was the posterior HDI. In this section, we discuss another Bayesian approach, the Bayes factor. The choice of Laplace prior was made for the reason that we want to have a prior that led to a tractable integral in calculating the Bayes factor and that placed more mass near zero but also exhibits heavier tails. A Laplace Bayes factor (LBF) with a Laplace prior and a Gaussian likelihood is derived in this section. The LBF not only ranks the SNPs, it can be used in updating the posterior probability of association (PPA) and hence can determine which SNPs are noteworthy.

The Laplace Bayes factor is derived from the Bayes factor's general equation given in Equation (2.8). We also rely on the justification discussed in Section 2.3.4 which explained why we do not need a prior on the intercept. Using the Bayes factor in Equation (2.16), under the null Hypothesis of no effect (β = 0)

$$f(\hat{\beta} \mid \beta = 0) = \frac{1}{\sqrt{2\pi V}} exp\left(-\frac{\hat{\beta}^2}{2V}\right). \tag{4.38}$$

Under the alternative Hypothesis ($\beta \neq 0$), from Equation (4.10), we can obtain $f(\hat{\beta} \mid H_1)$ which gives

$$f(\hat{\beta} \mid H_1) = \frac{\lambda}{2} exp\Big(-\frac{1}{2V}(\hat{\beta}^2 - Q_-^2)\Big)\Big[\Phi\Big(\frac{-Q_-}{\sqrt{V}}\Big)\Big] + \frac{\lambda}{2} exp\Big(-\frac{1}{2V}(\hat{\beta}^2 - Q_+^2)\Big)\Big[1 - \Phi\Big(\frac{-Q_+}{\sqrt{V}}\Big)\Big] \tag{4.39}$$

where Q_- and Q_+ were defined in Equation (4.7a) and (4.7b) respectively. This gives the Laplace Bayes factor (LBF) as follows

$$\frac{\lambda\sqrt{2\pi V}}{2} exp\Big(\frac{Q_-^2}{2V}\Big)\Big[\Phi\Big(\frac{-Q_-}{\sqrt{V}}\Big)\Big] + \frac{\lambda\sqrt{2\pi V}}{2} exp\Big(\frac{Q_+^2}{2V}\Big)\Big[1 - \Phi\Big(\frac{-Q_+}{\sqrt{V}}\Big)\Big] \tag{4.40}$$

4.3.1 ROC curves for Laplace Bayes factor

We continue to rank the SNPs by using the new derived LBF using the same procedure on the same simulated datasets previously used in ranking the SNPs with posterior summaries. As mentioned in Section 3.6, there is a lot of $\hat{\lambda}$. We consider how sensitive the LBF is to te choice of $\hat{\lambda}$ is to the LBF. We considered two values of lambda from the 148 top hits data. These values are the lowest and highest values specified in any of the 95% confidence intervals (15.3, 64.15) obtained from Table 3.3.

TPR and FPR values for LBF with both estimated lambda values were obtained for all 6 scenarios with rare and common single causal SNP with various ORs given in Table 4.1. The performances for both LBFs were quantified using ROC curves and were plotted together with the ROC curves for the posterior HDI as shown in Figure 4.6. We restricted the ROC curves to have FPR $\leq 4\%$ to clearly see the differences in performance between both LBFs.

First, we look at the ROC curves for the scenario with a single rare causal SNP (MAF=0.09). Each approach shows different performance in all given ORs. In the scenario with OR=1.15, the LBF using both lambdas has a better performance compared to posterior HDI but in the scenario with OR=1.08, it shows the opposite result. For OR=1.12, the LBF with $\hat{\lambda} = 64.15$ and the posterior HDI perform the same and are better than the LBF with $\hat{\lambda} = 15.30$.

In scenarios with a single common causal SNP (MAF=0.3), the ROC curves for every approach in all ORs show the performances are quite similar to each other. We cannot tell which approach shows a better performance in scenarios with OR=1.12. In the scenario with OR=1.08, the LBF with $\hat{\lambda} = 64.15$ has a slightly better performance than the others. For OR=1.15, the performance has a similar result as the result in the scenario with OR=1.12 with a single rare causal SNP.

If we want to compare just the LBF, the LBF obtained using $\hat{\lambda} = 64.15$ performed better in most of the scenarios shown in Figure 4.6. We now compare the performance of the posterior HDI and both LBFs in all scenarios if we reduce the power from 80% to 60%. We recalculated both LBFs with the same estimated lambda using new simulated data keeping the ORs and MAFs the same. The only changes made in each scenario is the total sample sizes since the power had been reduced by 20%. Figure 4.7 shows the ROC curves for both LBFs and also the posterior HDI in various scenarios with new total sample sizes. Figure 4.7 shows that all the ROC curves now have a much more difference between the classifiers in all scenarios. In most of the scenarios, both LBFs perform better than the posterior HDI except for OR=1.15 with a single common causal SNP. We can also see that the LBF with a higher estimated lambda have the best performance among others.

Based on the ROC curves, we can conclude that the LBF generally performs better than the posterior summaries. The ROC curves give information about the performance for ranking the SNPs. We further look at the LBF in updating the prior to the posterior probability of association (PPA) and determine the noteworthiness of all SNPs.

4.3.2 Noteworthiness of the SNPs using Laplace Bayes factor

In addition to using the LBF for ranking the SNPs, the LBF can be used to assess whether the strength of an association is noteworthy. We observed the noteworthiness of SNPs in all scenarios using the LBF and set the prior odds of the null hypothesis (π_0) to be 0.995. As priori, we expect 99.5$\%$ of SNPs in the region not be causal and associated to disease. We also set the cost of ratio, (C_ω/C_α)= 1

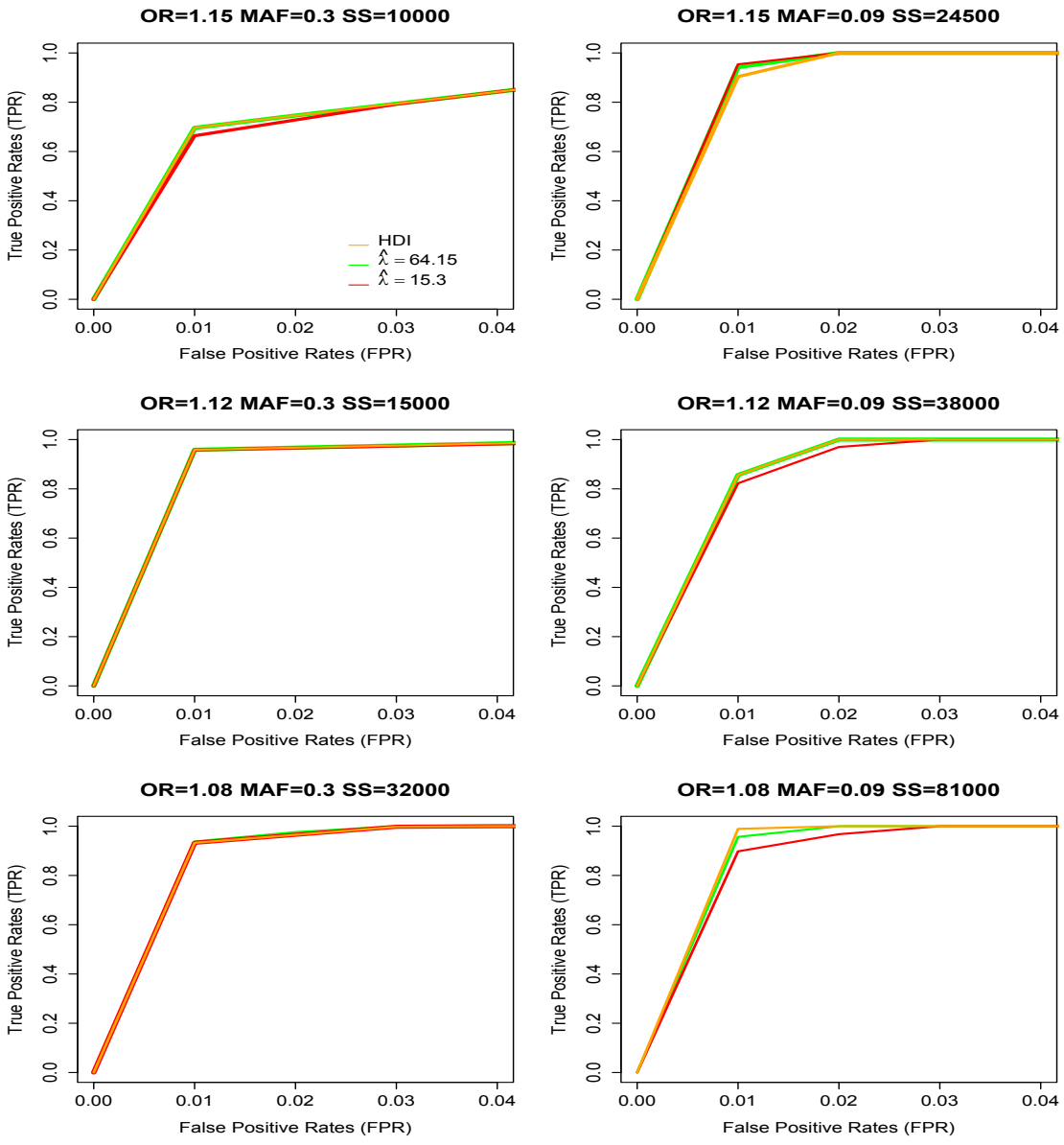


***Figure 4.6*:** Receiver operating characteristic (ROC) curves comparing the SNP ranking performance of the posterior highest density interval (HDI) with Laplace Bayes factor (LBF) with two different values of $\hat{\lambda}$. The values of $\hat{\lambda}$ used are 15.30 and 68.32. All rankings were carried out on 20 simulated datasets from HAPGEN using a single rare causal SNP (MAF=0.09) and a single common causal SNP(MAF=0.3) with three different odd ratios (OR = 1.08, 1.12, 1.15). The sample size (SS) for each scenario depends on 80% power.

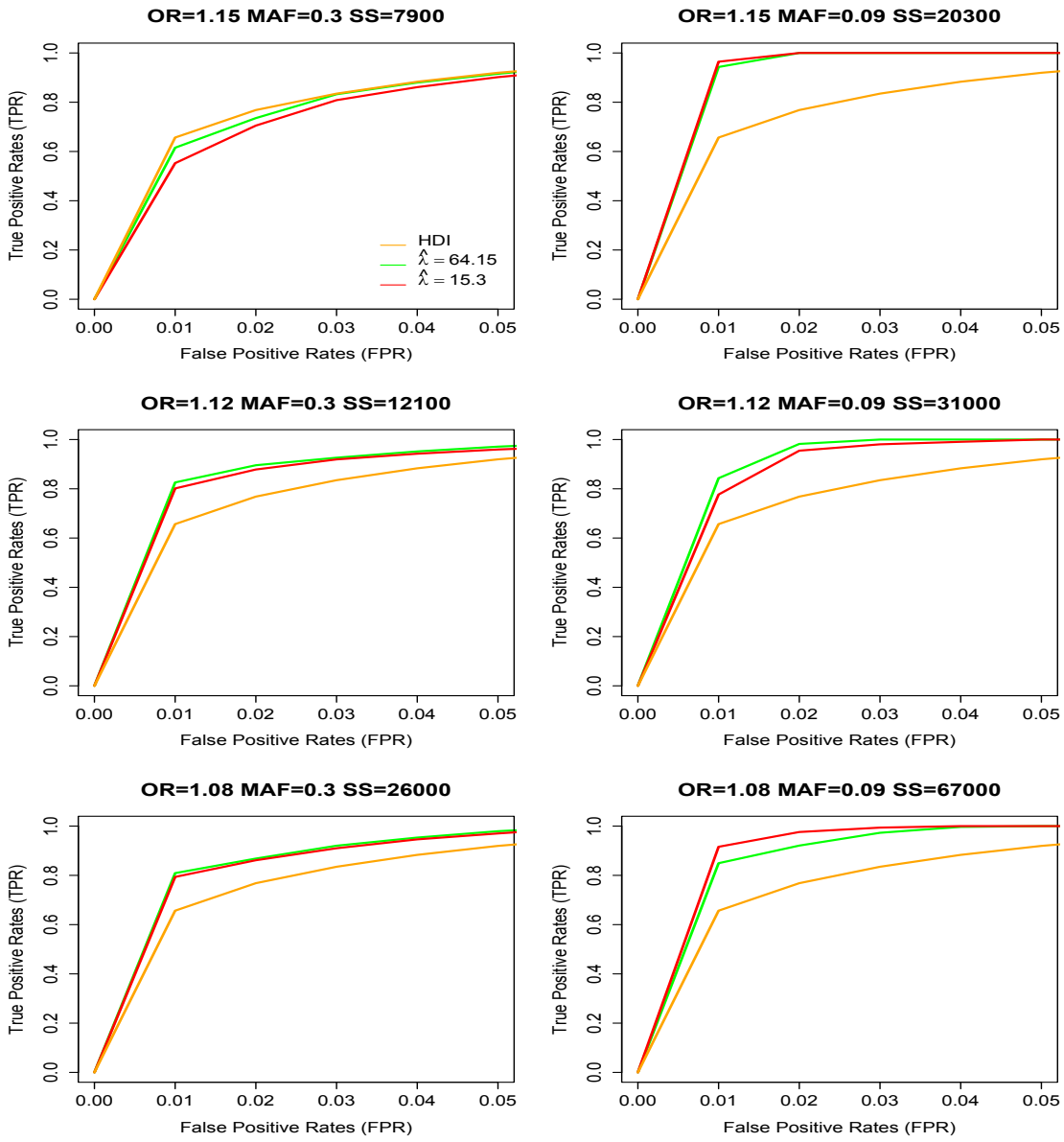


***Figure 4.7*:** Receiver operating characteristic (ROC) curves comparing the SNP ranking performance of the posterior highest density interval (HDI) with Laplace Bayes factor (LBF) with two different values of $\hat{\lambda}$. The values of $\hat{\lambda}$ used are 15.30 and 68.32. All rankings were carried out on 20 simulated datasets from HAPGEN using single a rare causal SNP (MAF=0.09) and a single common causal SNP(MAF=0.3) with three different odd ratios (OR = 1.08, 1.12, 1.15). The sample size (SS) for each scenario depends on 60% power.

(missing a discovery is equally as costly as reporting a null hypothesis). By using Equation 2.19, we will report a noteworthy SNP if the posterior probability on H_0 is less than 0.5. The posterior odds and posterior probabilities with a LBF are derived from Equation 2.20.

$$P(H_0 \mid \hat{\beta}) = \frac{\pi_0}{\pi_0 + \text{LBF}(1 - \pi_0)}. \tag{4.41}$$

The TPR for each scenario is calculated as the proportion of the twenty causal SNPs that are noteworthy (i.e the proportion over the 20 datasets). The FPR is the proportion of non causal SNPs declared noteworthy across all 20 datasets. Referring back to the ROC curves in Figure 4.6, in terms of ranking, the performance of the LBFs using $\hat{\lambda} = 15.30$ and $\hat{\lambda} = 64.15$ are slightly different in most scenarios. The decision to declare a SNP is noteworthy depends on the actual value of the Laplace Bayes factor rather than the ranks. Table 4.3 shows the TPR and FPR values for noteworthy SNPs using the LBF in the scenarios in Table 4.1 with sample sizes yielding 80% power.

***Table 4.3*:** True Positive Rates (TPR) and False Positive Rates (FPR) for declaring if the SNP is noteworthy using the Laplace Bayes factor (LBF) with $\hat{\lambda} = 64.15$ and $\hat{\lambda} = 15.30$ in various scenarios for a single causal SNP. The sample size for each scenario gives 80% power.

$\hat{\lambda}$	MAF	Odd Ratios	Sample Size	FPR	TPR
64.15	0.3	1.15	10000	0.027	0.4
		1.12	15000	0.035	0.7
		1.08	32000	0.054	1
	0.09	1.15	24500	0.0015	0.35
		1.12	38000	0.0044	0.4
		1.08	81000	0.0038	0.6
15.3	0.3	1.15	10000	0.095	1
		1.12	15000	0.083	0.95
		1.08	32000	0.078	1
	0.09	1.15	24500	0.010	1
		1.12	38000	0.020	0.7
		1.08	81000	0.0083	0.6

From Table 4.3, generally, more of the common causal SNPs are detected to have noteworthy

associations compared to the rare causal SNPs. For the LBF with estimated lambda equal to 64.15, if we examine the TPRs according to the ORs, we can see that as the OR gets smaller, more of the causal SNPs are picked up as having a noteworthy association. In the case where the LBF uses an estimated lambda of 15.30, all the causal SNPs in the 20 datasets with MAF=0.3 have a noteworthy association except for one causal SNP with OR=1.12. For rare causal SNPs, as the OR gets smaller, the number of noteworthy causal SNPs detected decreases, contradicting the pattern for λ = 64.15.

Declaring a SNP to have an association that is noteworthy is determined by the posterior odds on the null hypothesis which depends on the LBF. The differences in the number of causal SNPs having a noteworthy association in every scenario are therefore strongly related to the LBF obtained for the causal SNP in each dataset. A large value of LBF will reduce the posterior odds on the null and hence increase the chance that the SNP will have a noteworthy association. From Equation 4.41, the SNP Bayes factor should be more than 199 in order for the SNP to be claimed as noteworthy. The pattern of the TPR in Table 4.3 can be explained by observing the values of the LBF for causal SNP in each scenario.

Thus, boxplots are plotted to represent the distribution of the LBF of the 20 causal SNPs from the 20 datasets in all six scenarios. Figures 4.8 and 4.9 are boxplots for distribution of the LBF with $\hat{\lambda} = 15.30$ and $\hat{\lambda} = 64.15$ respectively. As mentioned above, we require the LBF to exceed 199 for the SNP to be noteworthy in our analysis. Generally, in both the LBFs with $\hat{\lambda} = 15.30$ and $\hat{\lambda} = 64.15$, the boxplots with a single common causal SNP (MAF=0.3) show larger values compared to a single causal SNP with MAF=0.09. Hence, this explains why more of the common causal SNPs were detected to have a noteworthy association compared to the rare causal SNPs across the 20 datasets.

We observed the boxplots for scenarios having TPR=1 (all 20 causal SNPs are noteworthy) to understand the distribution of the LBFs. From these boxplots, the LBFs in these scenarios are very large and have distributions with large median. This shows that all the causal SNPs across the 20 datasets have large LBFs which exceed the LBF threshold in our analysis. In scenarios with TPR less than equal to 0.4, the boxplots observed have median less than the LBF=199 explaining more than

50% of the causal SNPs are not noteworthy. In scenario with a single common SNP with OR=1.12, the distribution of the LBF with $\hat{\lambda} = 15.30$ shows very large median, however there is one SNP with small value of LBF and hence contributes to TPR=0.95.

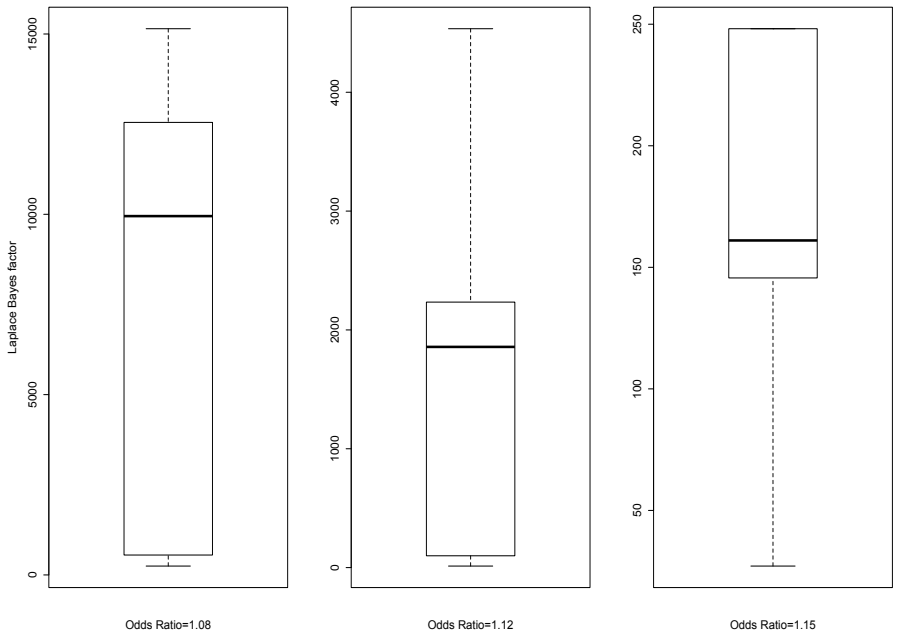
The distribution of the LBF described the pattern of TPRs in Table 4.3. Thus, the decision on the noteworthiness of SNPs depends on the LBF of each SNP. A large LBF would decrease the posterior probability on H_0 and hence increase the chance of declaring the SNP as noteworthy. However, other factors such as the SNP's odds ratio and minor allele frequency and also the estimated λ affect the size of the LBF.

As a result, the decision on the noteworthiness of SNPs depends on the LBF of each SNP. The chance of declaring a SNP as noteworthy increases when the LBF is large. The size of the LBF depends on other factors such as the SNP's odds ratio and minor allele frequency and also the estimated λ. As shown in Table 4.3, the LBF with a smaller estimated lambda ($\hat{\lambda} = 15.30$) appears to declare more noteworthy causal SNPs. This demonstrates that the LBF is sensitive to the choice of the $\hat{\lambda}$ and this can also be observed in Section 4.3.1. Hence, we propose a Bayes factor that allow for the uncertainty in the estimated λ.

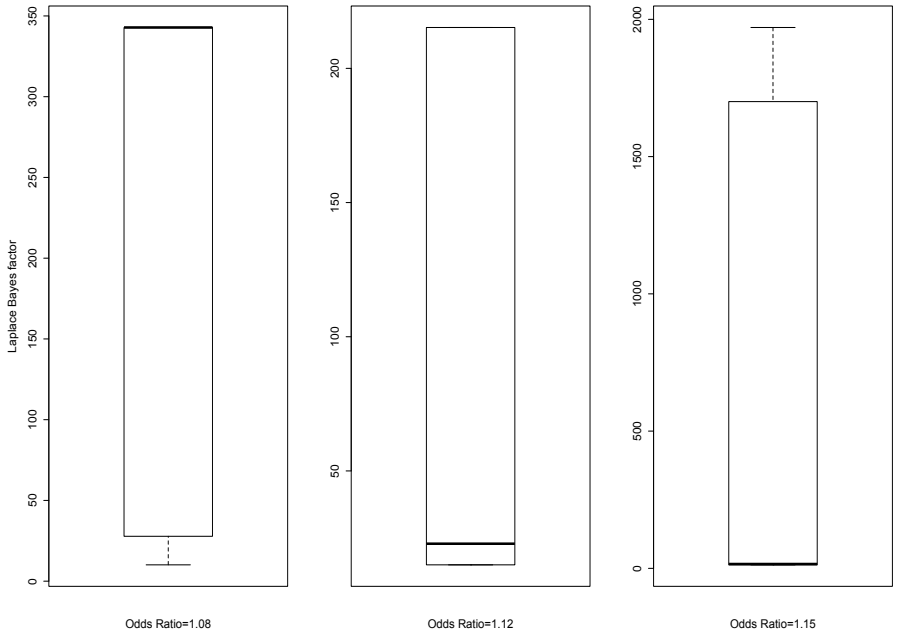
4.4 Laplace Gamma Bayes factor

The calculation for the posterior summaries and the LBFs depends on the value of $\hat{\lambda}$ that we choose. The choice we make on $\hat{\lambda}$ can make a significant difference in the performance of posterior summaries and the LBF in ranking the SNPs. The choices also affect the PPA and hence affect the noteworthiness of the SNPs. The value of $\hat{\lambda}$ was estimated from the top hits data taking account the number of YTBD SNPs. Hence, the value of $\hat{\lambda}$ varies according to how many SNPs are estimated to be unidentified. Figure 4.10 shows the relationship between $\hat{\lambda}$ and the number of YTBD SNPs. The uncertainty in $\hat{\lambda}$ in the plot is based on the 95% confidence interval from Table 3.3.

Since there is uncertainty in the estimated value of λ, we should also include the uncertainty in

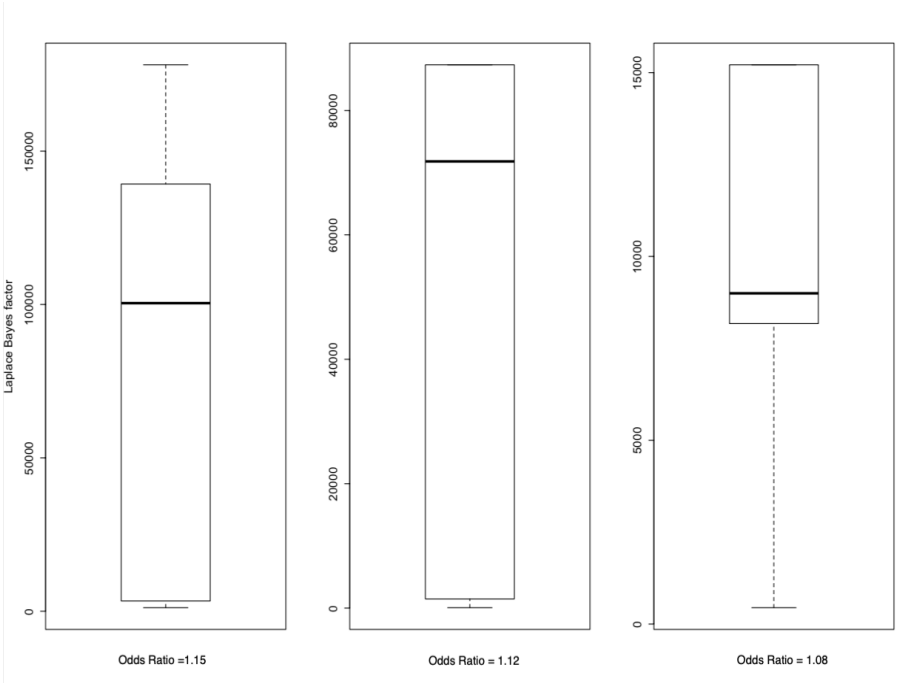


(a) $\hat{\lambda} = 64.15$, MAF=0.3.

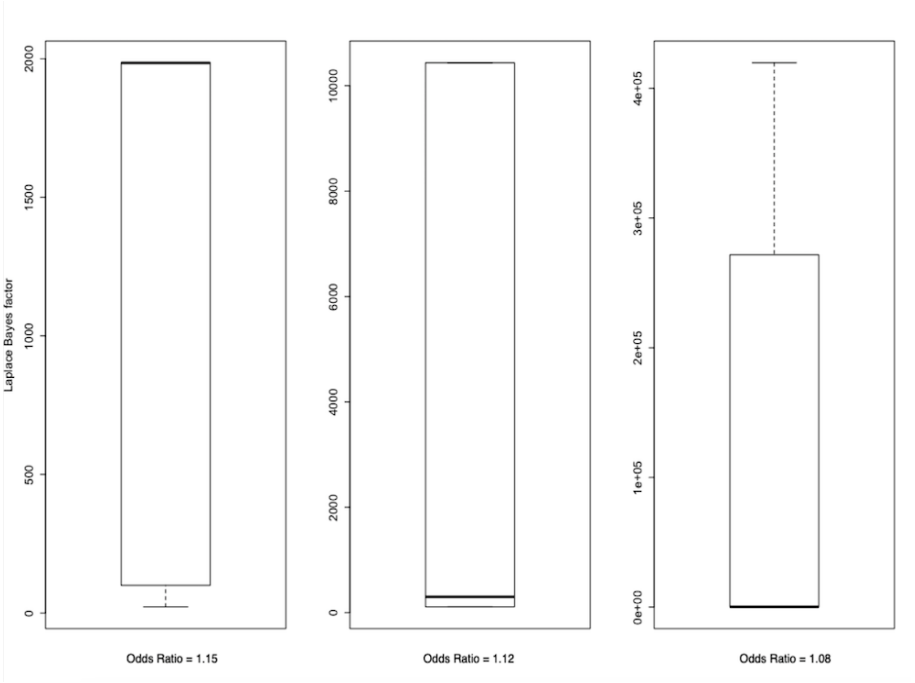


(b) $\hat{\lambda} = 64.15$, MAF=0.09

***Figure 4.8*:** Boxplots representing the distribution of the Laplace Bayes factor (LBF) of the 20 causal SNPs in six scenarios. The MAF and value of λ are given in the caption. The odds ratio is given on the x-axis of each plot.



(a) $\hat{\lambda} = 15.30$, MAF=0.3



(b) $\hat{\lambda} = 15.30$, MAF=0.09.

***Figure 4.9*:** Boxplots representing the distribution of the Laplace Bayes factor (LBF) of the 20 causal SNPs in six scenarios. The MAF and value of λ are given in the caption. The odds ratio is given on the x-axis of each plot.

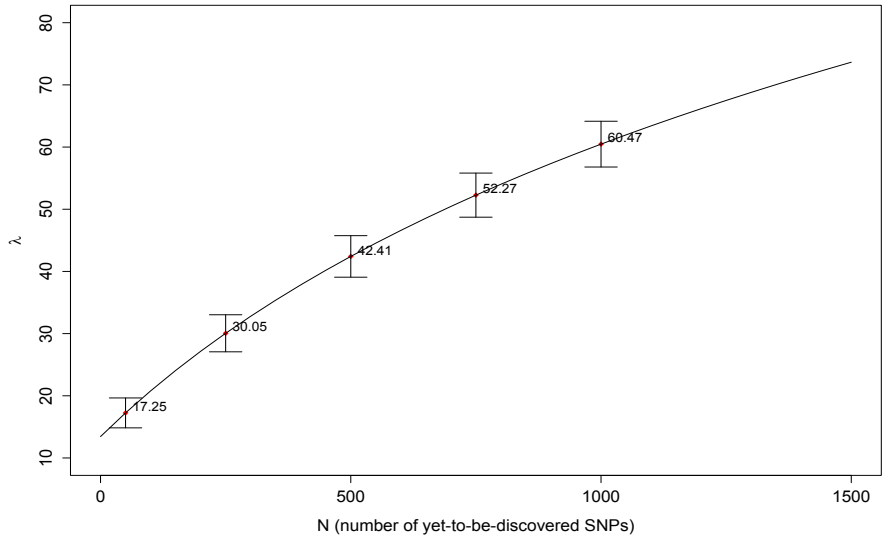


***Figure 4.10*:** The curve is the estimated relationship between $\hat{\lambda}$ and the number of yet-to-be-discovered SNPs, N. The error bars are ± 2 standard errors of $\hat{\lambda}$ taken from Table 3.3.

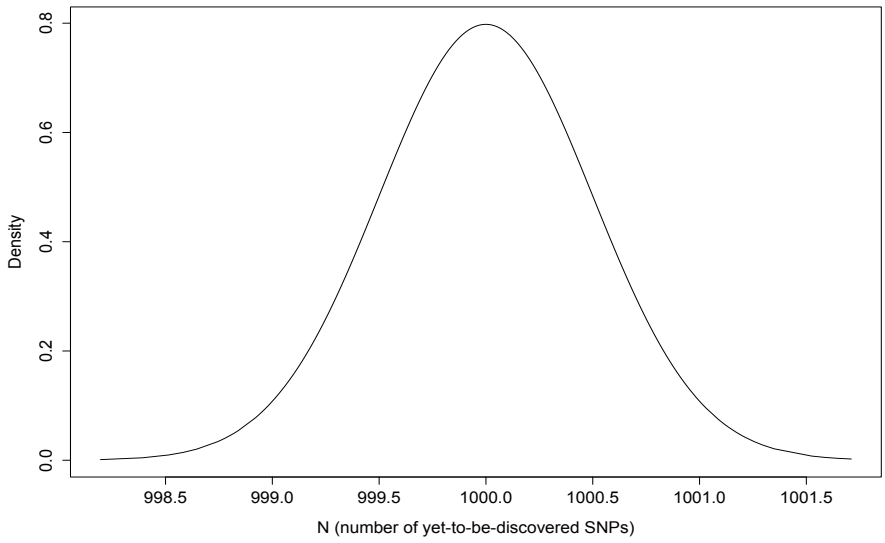


***Figure 4.11*:** Plot shows the Gamma probability density function (PDF) for the number of yet-to-be-discovered SNPs, N with $\theta = 4 \times 10^6$ and $\phi = 4000$

the number of YTBD SNPs when calculating the Bayes factor. Let N represent the random variable for the number of YTBD SNPs (N must be a positive value). Subsequently, we put a prior on N that follows a Gamma distribution since the support on Gamma Distribution (θ, ϕ) only allows value to be more than zero. Although N is a discrete random variable, we decide to use a continuous distribution to represent N in order to have a simpler calculation.

The parameters chosen must lead to a sensible prior and have high mass around $N = 1000$. Hence, the parameters are estimated from the mean ($\mathbb{E}(N) = \theta/\phi$) and variance ($Var(N) = \theta/\phi^2$) of the Gamma distribution with mean = 1000 and variance =0.25. Therefore, N follows a Gamma distribution with $\theta = 4 \times 10^6$ and $\phi = 4000$. Figure 4.11 shows the plot of the Gamma density function. The probability density function (PDF) is given as follows

$$f_N(n) = \begin{cases} \frac{1}{\Gamma(\theta)}\ \phi^\theta n^{\theta-1} e^{-\phi n} & \text{if } n > 0 \\ 0 & \text{if otherwise} \end{cases} \tag{4.42}$$

Let λ be the derived random variable and $g(\lambda)$ be the new PDF associated with λ. The transformation function, $\lambda = 2\sqrt{N}$ is obtained from the relationship between λ and N shown in Figure 4.10. From the transformation function, we obtain $N = \lambda^2/4$. The new PDF associated with λ is given as follows

$$g(\lambda) = f_N(\lambda)\ .\ \left|\frac{dN}{d\lambda}\right|$$

$$= \frac{\phi^\theta (4\lambda^2)^{\theta-1} e^{-4\phi\lambda^2}}{\Gamma(\theta)} (8\lambda)$$

$$= \frac{2(4\phi)^\theta\ \lambda^{2\theta-1}\ e^{-4\phi\lambda^2}}{\Gamma(\theta)} \tag{4.43}$$

This is the Nakagami Distribution, NK(λ; θ,$\frac{\theta}{4\phi}$) and thus, $\lambda \sim NK(\theta, \frac{\theta}{4\phi})$.

To obtain the Laplace Gamma Bayes factor (LGBF), we now have

$$\lambda \sim NK\left(\theta, \frac{\theta}{4\phi}\right)$$
$$\beta \mid \lambda \sim La(\lambda)$$
$$\hat{\beta} \mid \beta \sim N(\beta, V).$$

We are comparing the hypothesis of $H_0 : \beta = 0$ against $H_1 = \beta \neq 0$. Under the null hypothesis, we have

$$f(\hat{\beta} \mid \beta = 0) = \frac{1}{\sqrt{2\pi V}} exp\big(\frac{-\hat{\beta}^2}{2V}\big) \tag{4.44}$$

and under the alternative hypothesis,

$$f(\hat{\beta} \mid \beta \neq 0) = \int_{\beta}\int_{\lambda} \Pi(\lambda) f(\beta \mid \lambda) f(\hat{\beta} \mid \beta) \quad d\lambda \; d\beta. \tag{4.45}$$

Therefore, using the general Bayes factor formula in Equation (2.8), the LGBF is given by

$$\text{LGBF} = \frac{\int_{\beta}\int_{\lambda} \Pi(\lambda) f(\beta \mid \lambda) f(\hat{\beta} \mid \beta) \quad d\lambda \, d\beta}{\frac{1}{\sqrt{2\pi V}} exp\big(\frac{-\hat{\beta}^2}{2V}\big)}. \tag{4.46}$$

The integrals in the numerator cannot be solve analytically, hence, the calculation for LGBF is per-

formed using Monte Carlo Integration by re-expressing the integral as an expectation.

$$\begin{aligned}
&\int_{\beta} f(\hat{\beta} \mid \beta)\left[\int_{\lambda} \Pi(\lambda) f(\beta \mid \lambda)\ d\lambda\right] d\beta \\
&= \int_{\beta} f(\hat{\beta} \mid \beta)\, f(\beta)\ d\beta \\
&= \mathbb{E}_{\beta}\ f(\hat{\beta} \mid \beta) \\
&\approx \frac{1}{n} \sum_{i=1}^{n} f(\hat{\beta} \,|\, \beta_i)
\end{aligned} \tag{4.47}$$

where β is sampled by sequential Monte Carlo with large n.

4.4.1 ROC curves for Laplace Gamma Bayes factor

In order to evaluate the performance of LGBF, we calculate the TPR and FPR values for every scenario using the same procedure we did in obtaining for posterior summaries and LBF. Figure 4.12 shows the ROC curves for the LGBF plotted with the ROC curves for LBF with two estimated λ ($\hat{\lambda}$=64.15 and $\hat{\lambda}$=15.30). The ROC curves are zoom into FPR $\leq$ 4%. It is predicted that the LGBF will have a ROC curve in between the ROC curves of the maximum and minimum estimated lambda used in obtaining LBF. This predicted does not apply to all scenarios. Most of the LGBF have the same performance as LBF with $\hat{\lambda} = 64.15$. This is because this LBF used $\hat{\lambda} = 64.15$ which were estimated considering 1000 YTBD SNPs. As mentioned before, the parameters for the Gamma prior in LGBF were estimated to have mean=1000. However, the results for LGBF and LBF ($\hat{\lambda} = 64.15$) shows a better performance compared to LBF with $\hat{\lambda} = 15.30$.

To see more significant differences among the Bayes factor, the sample size for all scenario were reduced to depend on 60% power. Figure 4.13 shows the ROC curves with FPR $\leq$ 4% of all three Bayes factors obtained from 20 simulated datasets with the same scenario of single causal SNPs but with reduced sample size with 60% power. The prediction on LGBF's ROC curves to be in between both LBFs can be only be seen in scenario with OR=1.15 for a single common causal SNPs. Others

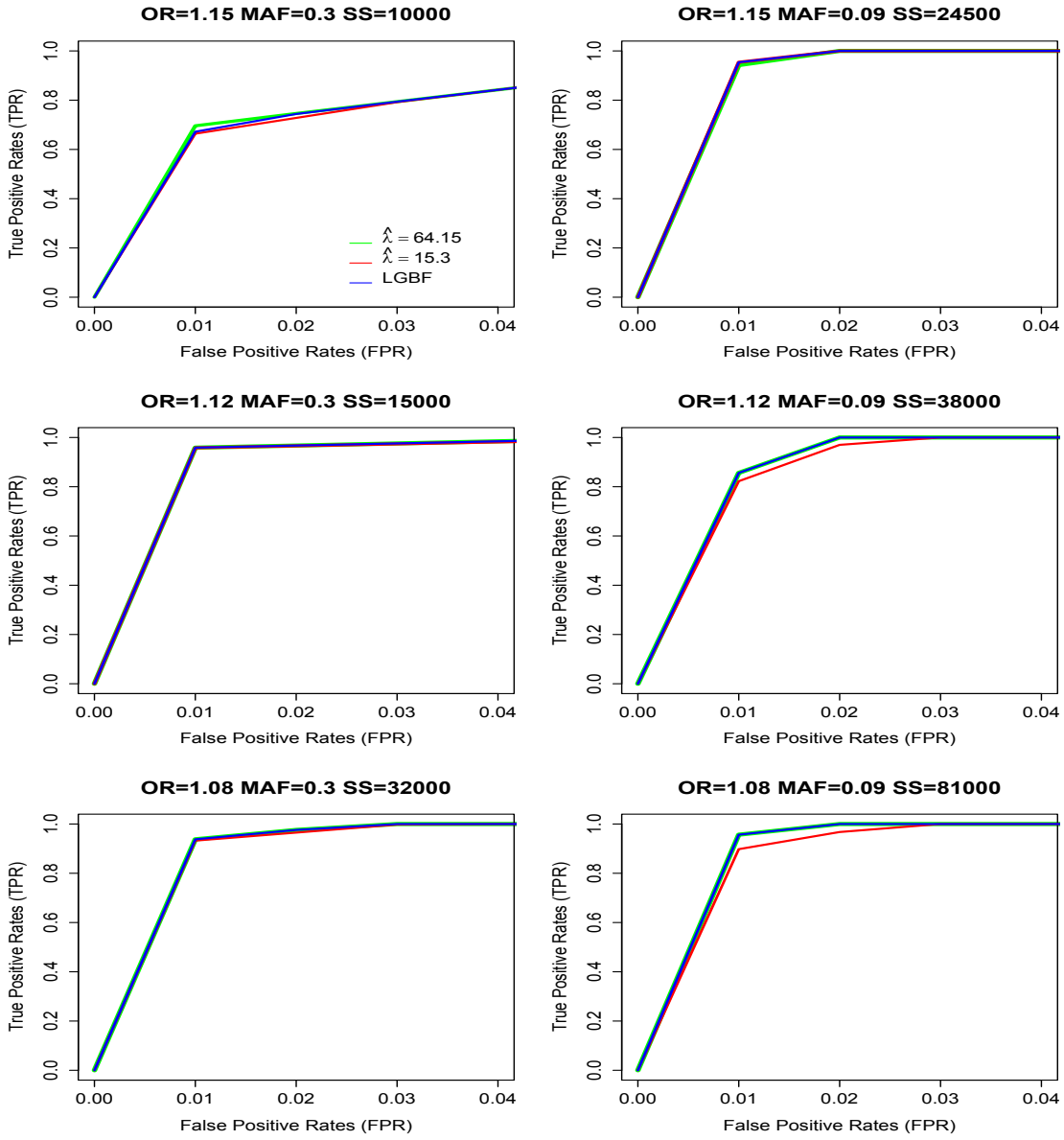


***Figure 4.12*:** Receiver operating characteristic (ROC) curves shows the results of SNP ranking using Laplace Gamma Bayes factor (LGBF) and Laplace Bayes factor with two different MLE ($\hat{\lambda}$=64.15 and $\hat{\lambda}$=15.30). The SNPs ranking were carried out on 20 simulated datasets from HAPGEN having 80% power with a single causal SNP of various scenarios.

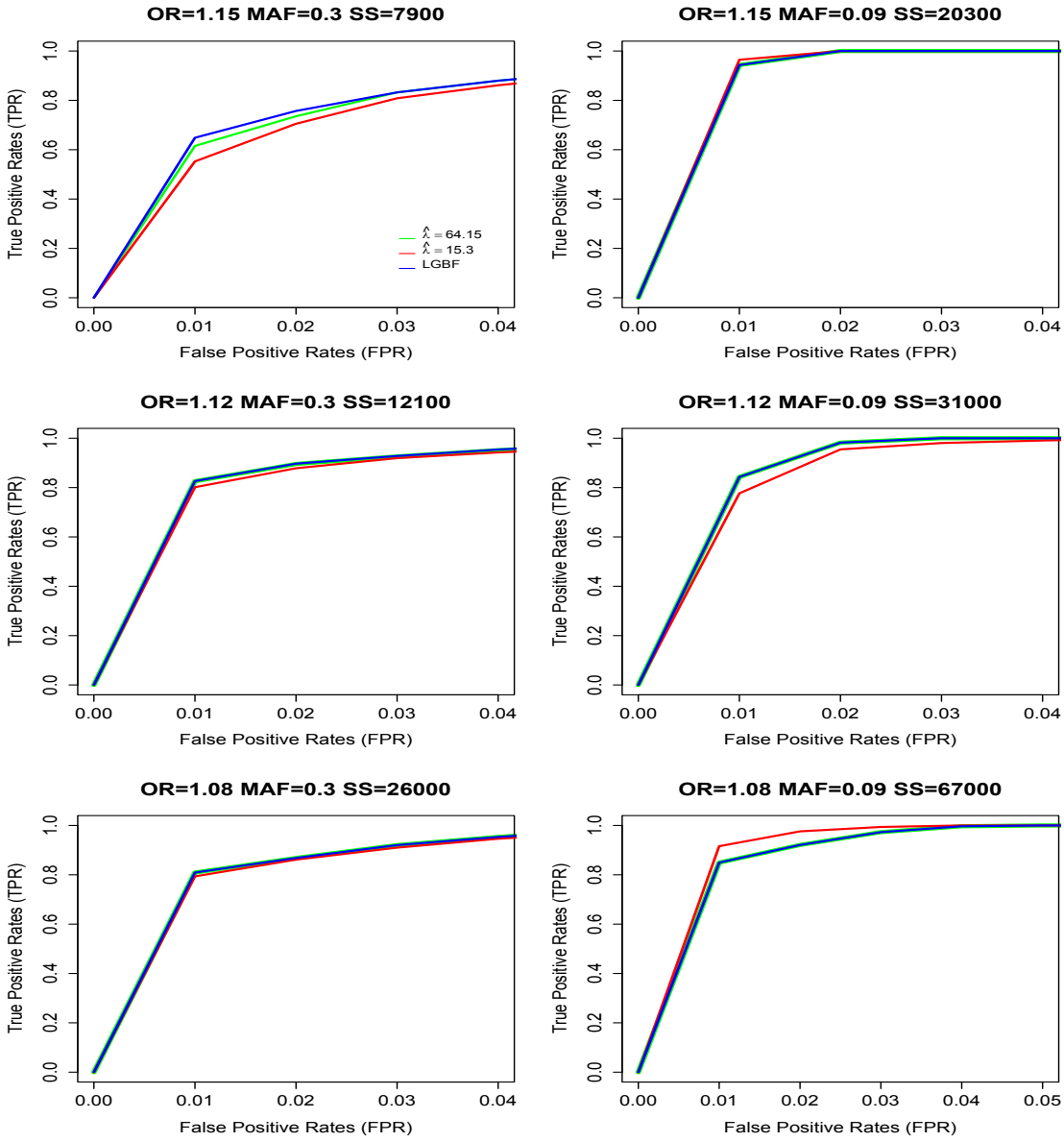


***Figure 4.13*:** Receiver operating characteristic (ROC) curves shows the results of SNP ranking using Laplace Gamma Bayes factor (LGBF) and Laplace Bayes factor with two different MLE ($\hat{\lambda}$=64.15 and $\hat{\lambda}$=15.30). The SNPs ranking were carried out on 20 simulated datasets from HAPGEN having 60% power with a single causal SNP of various scenarios.

show that LGBF and LBF ($\hat{\lambda} = 64.15$) have the same performance.

Although the LGBF and LBF ($\hat{\lambda} = 64.15$) have similar performance in ranking SNP, we further look into how LGBF affect PPA and therefore affect the decision to call a SNP as noteworthy.

4.4.2 Noteworthiness of the SNPs using Laplace Gamma Bayes factor

The notewortiness of SNPs by using LGBF used the same concept as the noteworthiness using LBF in section 4.3.2. An association is detect as noteworthy is based on equation 4.41. The only different is the Bayes factor in the posterior odds. Instead of using LBF, we now used LGBF. The posterior odds is given as follow

$$\text{Posterior Odds of } H_0 = \text{prior odds of } H_0 \text{ / LGBF}$$
$$Pr(H_0 \mid \hat{\beta}) = \frac{\pi_0}{\pi_0 + \text{LGBF}(1 - \pi_0)}. \tag{4.48}$$

Table 4.4: True Positive Rates (TPR) and False Positive Rates (FPR) for decision made if the SNP is noteworthy using Laplace Gamma Bayes factor (LGBF) in various scenarios of single causal SNP.

MAF	Odd Ratios	Sample Size	FPR	TPR
0.3	1.15	10000	0.0273	0.4
	1.12	15000	0.0355	0.7
	1.08	32000	0.0537	1
0.09	1.15	24500	0.0015	0.35
	1.12	38000	0.0044	0.35
	1.08	81000	0.0038	0.6

Table 4.4 shows the TPR and FPR for decision made if the SNP is noteworthy using (LGBF) in various scenarios with single causal SNP. The decision in identifying SNPs to have a noteworthy associations using LGBF are almost identical to the results for LBF ($\hat{\lambda} = 64.15$) in table 4.3. Over the 20 datasets, the causal SNP with MAF=0.3 appears more to have association that is noteworthy compared to rare causal SNP (MAF=0.09). Since we are interested in detecting a noteworthy association in smaller ORs, from Table 4.4, the number of SNPs appeared as noteworthy increases when the ORs

gets smaller.

We predicted LGBF to give a way of weighting the values between LBF with $\hat{\lambda} = 64.15$ and $\hat{\lambda} = 15.3$. However, ranking SNPs and identifying SNPs that have noteworthy associations using LGBF does not support our prediction. LGBF shows an almost identical results as using LBF with $\hat{\lambda} = 64.15$ (N=1000). This is because the parameters chosen for the Gamma distribution for the prior on N have high mass around N = 1000 as shown in Figure 4.11. This is saying that N is 1000. One way to improve this is to have a more sensible prior on N by spreading the mass between 50 to 1000 (the number of YTBD SNPs considered in this thesis). However, in our thesis, this might not give any significant results since there are no big difference (in most scenarios) between the performance of LBF with $\hat{\lambda} = 64.15$ and $\hat{\lambda} = 15.3$ as shown in Figure 4.12 and 4.13. Although this method was not very sensible, we had shown that the method works since it gives the same result as LBF with $\hat{\lambda} = 64.15$.

Chapter 5

Current multivariate Bayesian statistical approaches to fine mapping

In previous chapters we discussed statistical approaches to single-SNP analysis. Bayes factors were calculated for each SNP, one at a time, to test for association with the phenotype. However, Bayes factor only explain the association and do not detect causality, which means a non-causal SNP could obtain a similar Bayes factor to a causal SNP if these SNPs are in strong LD. This makes it difficult to distinguish the real causal SNP. Furthermore, there could be more than one causal SNP in the GWAS associated region.

To overcome this problem, we use multivariate (multi-SNP) analysis by using many SNPs in the region as explanatory variables in the regression. Problems arise however in selecting the causal SNPs from the set of SNPs available in the region. This problem can be addressed by using variable selection in which a small subset of the variables is chosen to be included in the model. In recent years, Bayesian variable selection is commonly used in fine mapping. PiMASS (Guan and Stephens, 2011), CAVIARBF (Chen et al., 2015) and FINEMAP (Benner et al., 2016) are among the methods applying multiple regression with Bayesian variable selection in fine mapping. We next briefly describe multiple logistic regression and then go on to review Bayesian variable selection.

5.1 Multiple logistic regression

Multivariate logistic regression can be extended from logistic regression in a univariate setting discussed in Section 2.2.1. Equation (2.1) refers to the probability of an individual with a disease. Thus, extending equation (2.1) into a multivariate setting gives

$$\begin{aligned} p_i &= \frac{exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)} \\ &= \frac{exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{1 + exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}. \end{aligned} \tag{5.1}$$

where p_i is the probability of having disease, $\boldsymbol{\beta} = (\alpha, \beta_1, ..., \beta_p)$ is the vector of beta coefficient and $\boldsymbol{x}_i$ is a vector of explanatory variables with the first value, $x_i = 1$ as the intercept. The odds of disease occurring are

$$\frac{p_i}{1 - p_i} = exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right). \tag{5.2}$$

To have a linear relationship between the odds and the explanatory variables, we use logit transformation by taking the logarithm of both sides of equation (5.2). Thus, the multivariate logistic model is expressed as

$$log\left[\frac{p_i}{1 - p_i}\right] = \boldsymbol{\beta}^T \boldsymbol{x}_i. \tag{5.3}$$

5.2 Bayesian variable selection

Recent developments in statistical approaches have allowed researchers to use available data to identify SNPs that might have causal association with disease. Multiple logistic regression in section 5.1 describes the basic model used to test for association with p explanatory variables. However, not all of

the explanatory variables are useful to predict disease outcome. Thus, selecting which explanatory variables are good predictors of disease is key. The explanatory variables selected are known as potential causal SNPs and are to be tested in follow up studies to identify all the true causal SNPS. One way to do this is to choose an optimal model from the 2^p possible models of different combinations of selected SNPs. However, it would not be computationally practical to consider all 2^p possible models individually. There are frequentist approaches to tackling this problem, such as stepwise regression and Lasso but our focus is only on Bayesian methods.

A Bayesian approach overcomes this problem by treating the parameters as random variables often with some point mass at zero. Bayesian variable selection (BVS) estimates the posterior probability for all models. BVS aims to select a small subset from all measured variables with high posterior probability since we can assume that only a small number of variables have an effect on disease outcome. The variables with no effect have an effect size of zero. Kuo and Mallick (1998) introduced a binary indicator variable to determine weather a predictor is considered or not considered to be included in the model. This is the approach taken by many multi-SNP fine mapping methods (Park and Casella, 2008; Griffin and Brown, 2010; Guan and Stephens, 2011; Bhattacharya et al., 2012; Chen et al., 2015; Benner et al., 2016; Alenazi et al., 2019) and is the approach we use.

Standard BVS searches over models containing different variables. To perform model selection, effect size priors that rely on shrinkage can be used. The specification of the prior can be a difficult task as it requires meaningful interpretation of the effect sizes. It is also challenging because the choice of prior will affect posterior. One way of implementing BVS is deciding which regression coefficients have values equal to zero (O'Hara and Sillanpää, 2009). Thus, specifying prior probabilities to the regression coefficients can force the regression coefficients to zero or close to zero.

Continuous prior distribution can represent an interesting prior for the regression coefficients. Continuous priors aim to shrink the regression coefficients with small values to zero while keeping the true large values. Commonly, the regression coefficient follows a normal distribution with a suitable prior on its variance to give appropriate shrinkage. This became the motivation for other researchers to use

a scale mixture of normal distributions for the effect size prior. (Park and Casella, 2008; Griffin and Brown, 2010; Bhattacharya et al., 2012; Alenazi et al., 2019).

Another way of specifying priors in BVS is using mixture priors also termed spike and slab priors (Mitchell and Beauchamp, 1988). In contrast to a continuous prior, there are two components in a spike and slab prior in which the spike is a point mass at zero whilst the slab specifies some other distribution elsewhere. A priori, the regression coefficients are assumed to be independent of each other. George and McCulloch (1993) modified the spike and slab prior by introducing a latent variable to distinguish zero and non-zero regression coefficients. They then used a mixture of two normal distributions on both components and fixed a very small variance for the normal distribution on the spike component to have it centered around zero.

As well as priors on the effect size, we also need to specify priors for each possible model (combination of SNPs). According to Forte et al. (2018), commonly used priors on each model in the model space follow a Binomial distribution with probability ρ. The hyperparameter ρ can either be a fixed or a random value. The typical choice for a fixed value, $\rho = 1/2^p$, assigns equal probability to each model. When treated as a random variable, the default choice for ρ has a uniform distribution.

How easy it is to derive the posterior depends on the choice of the prior. It is convenient to choose a prior that gives a tractable integral when we derive the marginal likelihood. Since the model space is vast with p being large, this will affect the choice of the computational method in variable selection. The standard approach used with Bayesian models for posterior evaluation and exploration is Markov chain Monte Carlo (MCMC). An MCMC algorithm samples from the posterior distribution in the model space but it is challenging with large p since it is difficult to be sure that the posterior over 2^p possible models is sufficiently explored.

5.3 Current Bayesian variable selection method used in fine mapping

There are numerous different approaches used in fine mapping to identify multiple causal SNPs in a GWAS associated region. In recent years, researchers have developed a number of statistical software packages to perform fine mapping based on a Bayesian approach. In general, they use a Bayesian variable selection method to prioritize potential causal variants by selecting a variant or a set of variants. The outputs from these different software are typically Bayes factor and posterior inclusion probability (PIP). These outputs help in deciding which variant or variants should be prioritized.

One of the earliest piece of software developed to tackle the limitations of single-SNP analyses is PiMASS (Guan and Stephens, 2011) in which they applied Bayesian Variable Selection Regression (BVSR) to GWAS. However, BVSR is also practical for fine mapping with hundred (or thousands) of covariates. This method focuses on the analysis of quantitative phenotypes but also allows a binary phenotype by introducing the probit link function. The genotype data is used as the independent variables in this analysis.

The prior specification in PiMASS focuses on appropriate priors for the hyperparameters they considered important which relates to the sparsity of the model and the effect sizes. They specified continuous prior on both hyperparameters with the sparsity hyperparameter follows a uniform distribution. Guan and Stephens (2011) propose a novel prior on the effect size hyperparameter which relates the prior to the total proportion of variance explained (PVE) in the outcome. In previous studies, the specification of priors for the sparsity and the effect size assumed the hyperparameters were independent random variables. This implies that the effect size hyperparameter and the indicator variable are also independent. This assumption can lead to the PVE increasing as the number of covariates rises.

For this reason, Guan and Stephens (2011) did not use this assumption in PiMASS as they believe more covariates will have smaller PVE, or a higher PVE may be achieved with a small number of

covariates. Therefore, a prior for the effect size hyperparameter given the indicator variable is specified to yield a flat prior on the PVE. This leads to a prior distribution on the effect sizes that leads to less shrinkage than other priors used before. As in most of Bayesian analysis, PiMASS used MCMC for the computation of the posterior distrubution for the parameters by sampling using a Metropolis-Hasting algorithm. They applied a Rao-Blackwellization technique to reduce the sampling variance when calculating the PIP.

In fine mapping, a number of studies have utilized GWAS summary statistics to perform multiple regression with variable selection. This is to allow analysis of GWAS data summary statistics without the need for the actual genotype data which may not be available. Kichaev et al. (2014) introduce PAINTOR, one of the first pieces of software that made use of the observed summary statistics (the association Z-scores) in modelling multiple causal variants. On top of using the association Z-scores, they also incorporate functional annotation data into their analysis. A standard way to obtain Z-scores is from the Wald statistics taken from regressing the phenotype on the SNPs.

Instead of specifying a prior distribution on the effect sizes of the SNPs, Kichaev et al. (2014) developed a method implemention in CAVIAR which introduced a prior probability through a standard logistic model applied to the effect size of the annotation given a SNP is causal. This is the method used by Kichaev et al. (2014) to incorporate the functional annotation data into the model. PAINTOR uses an EM algorithm to infer the parameters of the model and fixes the value for the effect sizes of the SNPs. They also restricted the number of causal SNPs in the region to be at most three potential causal SNPs. The posterior probabilities for each SNP are computed via exhaustive search.

Without the functional annotation, PAINTOR is comparable to CAVIAR (Hormozdiari et al., 2014). PAINTOR and CAVIAR do not specify a prior distribution for the effect size of the SNPs. They allow the observed summary statistics to follow a multivariate normal distribution. The effect size of the SNP is fixed and is included in the calculation of the mean for the observed summary statistics. Chen et al. (2015) described the relationship of CAVIAR with a Bayesian method for fine mapping called BIMBAM (Servin and Stephens, 2007). They showed that these methods are in fact

identical which led them to come up with another method which specifies a prior distribution on the effect size of all SNPs. Chen et al. (2015) called this method CAVIARBF.

CAVIARBF applied the standard BVS approach to computing posterior probabilities of SNPs being causal. To reduce the total number of models in the model space, CAVIARBF allow users to fix the maximum number of causal variants in the model. Thus, it requires for them to specify a prior for each model. They used the binomial distribution discussed in Section 5.2. As for the prior on the effect sizes, they used a Gaussian prior with a small variance of 0.01. CAVIARBF calculates Bayes factors analytically by comparing each model with the null model. They further use Bayes factor to calculate the PIP to prioritize SNPs.

One of the main factors to consider when choosing methods to perform fine mapping is the computational time. Benner et al. (2016) suggests that all the methods using summary statistics discussed above are too slow or even impossible to run when users set the number of causal variants to be more than three. This became a motivation for Benner et al to develop a novel software, FINEMAP, by introducing another computational algorithm whilst retaining the statistical model used in CAVIARBF. In FINEMAP, they fix the number of causal SNPs in the model to be up to five which is an improvement over CAVIARBF which allows for at most three causal SNPs. However, FINEMAP does not take into account the functional annotations in its model.

FINEMAP implement a Shotgun Stochastic Search (SSS) algorithm which uses a similar procedure as an MCMC algorithm to explore the model space. Rather than exploring the model space sequentially, as MCMC would, SSS explores the model in a parallel fashion at each iteration. For a given current model in an iteration, the proposed models are defined by deleting, changing and adding a causal SNP to the current model. The new current model in the next iteration is sampled from the proposed models in the previous iteration. All proposed models in each iteration are evaluated, and their posterior probabilities are saved in a list. This list grows to contain the posterior probabilties of models ever proposed. Thus, if there appear any already-evaluated models in the next iterations, its posterior probabilities would not be recomputed. This is the advantage of the SSS algorithm which

makes the method run quickly. The posterior probabilities in the list can be further used in calculating the Bayes factor and PIP of each SNP. The processing time is not the only factor that gives an advantage to FINEMAP, Benner et al claim that FINEMAP is more accurate especially when they increase the number of causal SNP to more than the maximum possible in CAVIARBF. A possible major drawback of FINEMAP is the same as for MCMC in BVS. It is difficult to know whether FINEMAP adequately explores the model space or whether it gets stuck in a local posterior mode.

All software discussed in this section is summarised in Table 5.1. The similarities and differences of each method are highlighted. Typically, the prior for the effect sizes in a GWAS or fine-mapping study is Gaussian distribution. In Section 3.4, we already obtained a prior driven by the GWAS top hits data and used that distribution as a prior in single-SNP analysis in Chapter 4. Consequently in the coming chapters, we develop a multi-SNP Bayesian approach using the Laplace prior discussed in Chapter 3 and compare the results with those using a Gaussian prior.

***Table 5.1*:** A summary of current software used in Bayesian fine mapping

Software	Type of phenotype	Type of input data	Maximum number of causal SNPs allowed	Prior on effect size (distribution)	Computation	Functional data	Output
piMASS	Quantitative, Binary	Genotype data	Unlimited	Continuous (Gaussian)	MCMC	No	Bayes factor and PIP
PAINTOR	Quantitative, Binary	Summary statistic	3	Empirical Bayes prior	Exhaustive enumeration	Yes	Bayes factor and PIP
CAVIARBF	Quantitative, Binary	Summary statistic	5	Spike and slab (Gaussian)	Exhaustive enumeration	Yes	Bayes factor and PIP
FINEMAP	Quantitative, Binary	Summary statistic	5	Spike and slab (Gaussian)	Shotgun stochastic search	No	Bayes factor and PIP

Chapter 6

A multivariate Bayesian approach with Gaussian and Laplace priors

Bayesian model selection in a fine map setting calculates the posterior probability of a specific model ($\mathcal{M}$). A model in fine mapping is formed from indicator variables for each SNP which are then organized in a vector $\mathbf{c}$. The indicator variables take the value of 1 for causal SNPs and 0 otherwise. For p SNPs, there are 2^p possible models ranging from having all 0 values (no causal SNP present) to having all values equal to 1 which means all SNPs are causal.

We restrict the space of models to those with at most K causal SNPs. This is equivalent to those models $\mathcal{M}_{\boldsymbol{c}}$, such that $||\boldsymbol{c}||_1 \leq K$, where $||\cdot||_1$ is the L_1 norm. Using a Bayesian approach, we can compute the posterior probability of a specific model by combining the prior probability of the model, the prior density of the effect size and the likelihood of the observed data, $\hat{\boldsymbol{\beta}}$ derived from the phenotype and genotypes. The calculation for the posterior probability of $\mathcal{M}_c$ is given as follows

$$P(\mathcal{M}_{\mathbf{c}}) = \frac{P(\mathcal{M}_{\mathbf{c}},\, \hat{\boldsymbol{\beta}})}{P(\hat{\boldsymbol{\beta}})} = \frac{P(\mathcal{M}_c)\, P(\hat{\boldsymbol{\beta}}|\mathcal{M}_c)}{P(\hat{\boldsymbol{\beta}})}. \tag{6.1}$$

Where $\mathcal{M}_c$ represents the model containing only those SNPs where the elements of $\mathbf{c}$ are 1.

6.1 Defining the prior probability of the model $P(\mathcal{M}_{\mathbf{c}})$

The prior probability for each model, $\mathcal{M}$, can be defined by letting the number of causal SNPs, k, follow a binomial distribution with probability ω. The prior probability of k causal SNPs in the model is

$$P(\mathcal{M}_{\mathbf{c}}|\, k, p, \omega) \;=\; \binom{p}{k}\, \omega^k\, (1-\omega)^{p-k} \qquad \text{for } \; k \;=\; 0, 1, ..., p. \tag{6.2}$$

Since we restrict the space of models to have at most K causal SNPs, we need to normalise the prior model. Thus, for each model, the prior is

$$P(\mathcal{M}_{\mathbf{c}}) \;=\; \frac{\binom{p}{k}\, \omega^k\, (1-\omega)^{p-k}}{\sum_{k=1}^{K} \binom{p}{k}\, \omega^k\, (1-\omega)^{p-k}} \qquad \text{for } \; k \;=\; 0, 1, ..., K. \tag{6.3}$$

where $K = ||\boldsymbol{c}||_1$.

6.2 Deriving the marginal likelihood $P(\hat{\boldsymbol{\beta}}|\mathcal{M}_{\mathbf{c}})$

$P(\hat{\boldsymbol{\beta}}|\mathcal{M}_{\mathbf{c}})$ is the probability density of the observed data under a specific model, $\mathcal{M}_{\mathbf{c}}$. In a Bayesian setting, we treat $P(\hat{\boldsymbol{\beta}}|\mathcal{M}_{\mathbf{c}})$ as a marginal likelihood evaluated by integrating $f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}|\mathcal{M}_{\mathbf{c}})$ over $\boldsymbol{\beta_c}$ where $\boldsymbol{\beta_c} = \{\beta_j \text{ for which } c_j = 1\}$. If we let $\boldsymbol{\beta_N} = \{\beta_j \text{ for which } c_j = 0\}$ then we have

$$\begin{aligned} P(\hat{\boldsymbol{\beta}}|\mathcal{M}_{\mathbf{c}}) &= \int_{\boldsymbol{\beta_c}} f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta_c}\,|\mathcal{M}_{\mathbf{c}}) \;\; d\boldsymbol{\beta_c} \\ &= \int_{\boldsymbol{\beta_c}} \frac{f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta_c}\,, \mathcal{M}_{\mathbf{c}})}{P(\mathcal{M}_{\mathbf{c}})} \;\; d\boldsymbol{\beta_c} \\ &= \int_{\boldsymbol{\beta_c}} \frac{P(\mathcal{M}_{\mathbf{c}})\, f(\boldsymbol{\beta_c}\,|\mathcal{M}_{\mathbf{c}})\, f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta_c}\,, \mathcal{M}_{\mathbf{c}})}{P(\mathcal{M}_{\mathbf{c}})} \;\; d\boldsymbol{\beta_c} \\ &= \int_{\boldsymbol{\beta_c}} f(\boldsymbol{\beta_c}\,|\mathcal{M}_{\mathbf{c}})\, f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta_c}\,, \mathcal{M}_{c}) \;\; d\boldsymbol{\beta_c} \end{aligned} \tag{6.4}$$

where $f(\boldsymbol{\beta}\,|\mathcal{M}_{\mathbf{c}})$ is the joint prior distribution for the effect sizes of SNPs given the model and $f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta_c}\,, \mathcal{M}_{\mathbf{c}})$ is the likelihood of the data given all the effect sizes.

6.3 The likelihood of the data, $f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta_c}, \mathcal{M}_{\mathbf{c}})$

In our case the data used is the estimated $\boldsymbol{\beta}$ from the fitted multiple linear regression, along with the covariance of the estimated effect sizes, $\boldsymbol{V}$. $\hat{\boldsymbol{\beta}}$ is conditioned on $\boldsymbol{\beta}$ and distributed as normal distribution. Thus,

$$\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{\beta} \;\sim N_p(\boldsymbol{\beta}, \boldsymbol{V})$$

where $\boldsymbol{V}$ is the variance covariance matrix from the fitted linear regression. In Section 5.2, we defined $\boldsymbol{\beta} = (\alpha, \beta_1, ..., \beta_p)$ where α is the intercept in the logistic regression model in Equation (5.3). For

computational reasons, we would not to include the intercept. However, we need to have a justification for not needing to use the intercept in the model. We have shown how a transformed intercept can yield cov($\hat{\alpha}, \hat{\beta}$) = 0 for a scalar $\hat{\beta}$ (Section 4.2) but we need to justify it for a vector $\boldsymbol{\beta}$ (refer Appendix A).

Given a model, $\mathcal{M}_c$, the likelihood is introduced through a partitioned matrix by grouping the SNPs into causal ($\boldsymbol{\beta_c}$) and non-causal ($\boldsymbol{\beta_N}$) as follows

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{c}} \,|\, \boldsymbol{\beta_c} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{N}} \,|\, \boldsymbol{\beta_N} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\beta_c} \\ \mathbf{0} \end{bmatrix}, \boldsymbol{V} \right).$$

The pdf is given as

$$f(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{\beta_c}, \mathcal{M}_c) = (2\pi)^{-p/2} |\boldsymbol{V}|^{-1/2} \; exp\left(-\frac{1}{2} \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{\beta_c} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{N}} - \mathbf{0} \end{bmatrix}^T \boldsymbol{V}^{-1} \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{\beta_c} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{N}} - \mathbf{0} \end{bmatrix} \right). \tag{6.5}$$

We let,

$$\begin{bmatrix} \boldsymbol{\Sigma_c} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}^T & \boldsymbol{\Sigma_N} \end{bmatrix} = \boldsymbol{V}^{-1} \tag{6.6}$$

where $\boldsymbol{\Sigma_c}$ is a $||c||_1 \times ||c||_1$ variance covariance matrix of the causal SNPs, $\boldsymbol{\Sigma_N}$ is a $(p - ||c||_1) \times (p - ||c||_1)$ variance covariance matrix of the non-causal SNPs and $\boldsymbol{\Sigma}$ is a $||c||_1 \times p$ SNPs variance matrix. Expanding the exponent term in Equation (6.11) gives

$$-\frac{1}{2}\Big[(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{\beta}_c)^T \boldsymbol{\Sigma_c} (\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{\beta}_c) + (\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{\beta}_c)^T \boldsymbol{\Sigma}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{N}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{\beta}_c) + \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T \boldsymbol{\Sigma_N} \hat{\boldsymbol{\beta}}_{\boldsymbol{N}} \Big]. \tag{6.7}$$

$\boldsymbol{\Sigma}$ is a scalar matrix, therefore, $(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{\beta}_c)^T \boldsymbol{\Sigma}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{N}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} - \boldsymbol{\beta}_c)$. Thus, Equation (6.7) can be

written as

$$
\begin{aligned}
&-\frac{1}{2}\Big[(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}-\boldsymbol{\beta}_c)^T\,\boldsymbol{\Sigma_c}\,(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}-\boldsymbol{\beta}_c)+2(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}-\boldsymbol{\beta}_c)^T\,\boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}+\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma_N}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}\Big]\\
&=-\frac{1}{2}\Big[\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}-\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c}-\boldsymbol{\beta}_c^T\,\boldsymbol{\Sigma_c}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}+2\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}{}^T\,\boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}-\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}\big)+\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma_N}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}\Big]. \qquad (6.8)
\end{aligned}
$$

To simplify Equation (6.8) we note that $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c}=\boldsymbol{\beta}_c^T\,\boldsymbol{\Sigma_c}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ because $\boldsymbol{\Sigma_c}$ is a symmetric matrix. Thus, the exponent term is

$$
\begin{aligned}
&-\frac{1}{2}\Big[\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}-2\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c}+\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c}+2\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}{}^T\,\boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}-\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}\big)+\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma_N}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}\Big]\\
&=-\frac{1}{2}\Big[\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c}-2(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,+\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T)\boldsymbol{\beta_c}+\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}+2\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}+\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma_N}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}\Big]\\
&=-\frac{1}{2}\Big[\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c}-2(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}{}^T\,\boldsymbol{\Sigma_c}\,+\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T)\boldsymbol{\beta_c}+\boldsymbol{Q}\Big] \qquad (6.9)
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{Q}&=\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}+2\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}+\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma_N}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}\\
&=\begin{bmatrix}\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}\\ \hat{\boldsymbol{\beta}_{\boldsymbol{N}}}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{\Sigma_c} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma}^T & \boldsymbol{\Sigma_N}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}\\ \hat{\boldsymbol{\beta}_{\boldsymbol{N}}}\end{bmatrix}\\
&=\hat{\boldsymbol{\beta}}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}. \qquad (6.10)
\end{aligned}
$$

The pdf for the likelihood of the data is therefore

$$
f(\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{\beta}\,,\mathcal{M}_c)=\Big[(2\pi)^p\;|\boldsymbol{V}|\Big]^{-1/2}\;exp\Big(-\frac{1}{2}\Big[\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c}-2(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}{}^T\,\boldsymbol{\Sigma_c}\,+\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T)\boldsymbol{\beta_c}+\boldsymbol{Q}\Big]\Big). \qquad (6.11)
$$

6.4 Spike and slab prior

There are several ways to specify the prior distribution of the effect size. We assume a maximum number of causal SNPs (K) in the model. In this research we chose a mixture prior with two components; the spike and slab. The spike component is a point mass at zero and the slab component is a continuous distribution with a mass spread over all possible real values of β excluding $\beta = 0$. Thus, the general prior distribution for the effect size of SNP j has pdf

$$f(\beta_j|\, c_j) = (1-g)(1-c_j)\, + g\, h(\beta_j)$$

where c_j is the j^{th} elements of c and $h(\beta_j)$ is the slab component of the prior. When $c_j = 0$, the prior is a point mass of $(1-g)$ at $\beta_j = 0$. When $c_j = 1$, the prior is the density $h(\beta_j)$ scaled by g.

g is the probability of the effect sizes is non zero whilst $(1-g)$ is the probability of the effect size is zero. The choice we made on g is the same as how we choose ω in Section 6.1. The reason we chose $\omega = 1/p$ is because we want to get the expected value of 1 which is a prior (Benner et al., 2016). We consider two different priors for the slab component in this research. The priors are the Gaussian prior and the Laplace prior which will be discussed in Section 6.5 and 6.6 respectively.

6.5 Deriving $P(\hat{\boldsymbol{\beta}}|\mathcal{M}_{\mathbf{c}})$ using the Gaussian prior

We derive $P(\hat{\boldsymbol{\beta}}|\mathcal{M}_{\mathbf{c}})$ using a Gaussian prior on the slab component as follows

$$\boldsymbol{\beta}_c|\mathcal{M}_{\mathbf{c}} \sim\ N_{||c||_1}(\mathbf{0}, \boldsymbol{W})$$

where, $\boldsymbol{W} = \text{diag}(w_1, w_2, ..., w_{||c||_1})$. Hence the pdf is

$$f(\boldsymbol{\beta}_c|\,\mathcal{M}_{\mathbf{c}}) = \left[(2\pi)^{||c||_1}\ |\boldsymbol{W}|\right]^{-1/2} exp\Big(\ -\frac{1}{2}\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{W}^{-1}\,\boldsymbol{\beta}_{\boldsymbol{c}}\Big). \tag{6.12}$$

From Equation (6.4) we want to integrate $f(\boldsymbol{\beta_c}\,|\mathcal{M}_{\mathbf{c}})\,f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta_c}\,,\mathcal{M}_{\mathbf{c}})$ over $\boldsymbol{\beta_c}$. Using the Gaussian prior in Equation (6.12) and the likelihood in Equation (6.11), we have

$$\begin{aligned}
&f(\boldsymbol{\beta_c}\,|\mathcal{M}_{\mathbf{c}})\,f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta_c}\,,\mathcal{M}_{\mathbf{c}})\\
&= \left[(2\pi)^{p+||c||_1}\,|\boldsymbol{V}|\ |\boldsymbol{W}|\right]^{-1/2} exp\Big(\ -\frac{1}{2}\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{W}^{-1}\,\boldsymbol{\beta_c} - \frac{1}{2}\Big[\boldsymbol{\beta}_{\boldsymbol{c}}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c} - 2(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^{\ T}\,\boldsymbol{\Sigma_c}\ +\ \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T)\boldsymbol{\beta_c} + \boldsymbol{Q}\Big]\Big)\\
&= \left[(2\pi)^{p+||c||_1}\,|\boldsymbol{V}|\ |\boldsymbol{W}|\right]^{-1/2} exp\Big(\ -\frac{1}{2}\Big[\boldsymbol{\beta}_{\boldsymbol{c}}^T\,(\boldsymbol{W}^{-1} + \boldsymbol{\Sigma_c})\,\boldsymbol{\beta_c} - 2(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^{\ T}\,\boldsymbol{\Sigma_c}\ +\ \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T)\boldsymbol{\beta_c} + \boldsymbol{Q}\Big]\Big).
\end{aligned} \tag{6.13}$$

The exponent term in Equation (6.13) can be expressed as

$$\begin{aligned}
&\boldsymbol{\beta}_{\boldsymbol{c}}^T\,(\boldsymbol{W}^{-1} + \boldsymbol{\Sigma_c})\,\boldsymbol{\beta_c} - 2(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^{\ T}\,\boldsymbol{\Sigma_c}\ +\ \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T)\boldsymbol{\beta_c} + \boldsymbol{Q}\\
&= \boldsymbol{\beta}_{\boldsymbol{c}}^T\boldsymbol{\Omega}\boldsymbol{\beta_c} - 2\mathcal{H}\boldsymbol{\beta_c} + \boldsymbol{Q}
\end{aligned} \tag{6.14}$$

where $\boldsymbol{\Omega} = \boldsymbol{W}^{-1} + \boldsymbol{\Sigma_c}$ and $\mathcal{H} = \hat{\boldsymbol{\beta}}_{\boldsymbol{c}}^{\ T}\,\boldsymbol{\Sigma_c}\ +\ \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T$.

We need to put Equation (6.9) into the form $(\boldsymbol{\beta_c} - \boldsymbol{\mu})^T\,\boldsymbol{\Omega}\,(\boldsymbol{\beta_c} - \boldsymbol{\mu})\ +\ \boldsymbol{R}$ in order to express Equation (6.13) as a Gaussian Kernel. Now we have,

$$\begin{aligned}
&(\boldsymbol{\beta_c} - \boldsymbol{\mu})^T\,\boldsymbol{\Omega}\,(\boldsymbol{\beta_c} - \boldsymbol{\mu})\ +\ \boldsymbol{R}\\
&= \boldsymbol{\beta}_{\boldsymbol{c}}^T\boldsymbol{\Omega}\boldsymbol{\beta_c} - 2\boldsymbol{\mu}^T\boldsymbol{\Omega}\boldsymbol{\beta_c} + \boldsymbol{\mu}^T\boldsymbol{\Omega}\boldsymbol{\mu} + \boldsymbol{R}.
\end{aligned} \tag{6.15}$$

Equating Equation (6.9) and (6.15) gives,

$$\begin{aligned}\mathcal{H} &= \boldsymbol{\mu}^T \boldsymbol{\Omega} \\ \mathcal{H}^T &= \boldsymbol{\Omega}^T \boldsymbol{\mu} \\ &= \boldsymbol{\Omega}\boldsymbol{\mu} \quad \text{since} \quad \boldsymbol{\Omega}^T = \boldsymbol{\Omega}.\end{aligned}$$

Therefore

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{\Omega}^{-1}\, \mathcal{H}^T \\ \boldsymbol{\mu}^T &= \mathcal{H}\, \boldsymbol{\Omega}^{-1} \quad \text{since} \quad \boldsymbol{\Omega}^{-1} \quad \text{is symmetric.}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\mu}^T \boldsymbol{\Omega} \boldsymbol{\mu} &= (\mathcal{H}\, \boldsymbol{\Omega}^{-1}) \boldsymbol{\Omega} (\boldsymbol{\Omega}^{-1}\, \mathcal{H}^T) \\ &= \mathcal{H}\, \boldsymbol{\Omega}^{-1}\, \mathcal{H}^T \end{aligned} \tag{6.16}$$

Since $\boldsymbol{Q} = \boldsymbol{\mu}^T \boldsymbol{\Omega} \boldsymbol{\mu} + \boldsymbol{R}$, we have,

$$\boldsymbol{R} = \boldsymbol{Q} - \boldsymbol{\mu}^T \boldsymbol{\Omega} \boldsymbol{\mu}.$$

We substitute $\boldsymbol{Q}$ from Equation (6.10) and $\boldsymbol{\mu}^T \boldsymbol{\Omega} \boldsymbol{\mu}$ from Equation (6.16) into $\boldsymbol{R}$ which gives

$$\boldsymbol{R} = \hat{\boldsymbol{\beta}}^T\, \boldsymbol{V}^{-1} \hat{\boldsymbol{\beta}} \, - \mathcal{H}\, \boldsymbol{\Omega}^{-1}\, \mathcal{H}^T. \tag{6.17}$$

We can express Equation (6.13) as

$$\left[(2\pi)^{p+||c||_1}\,|\boldsymbol{V}|\;|\boldsymbol{W}|\right]^{-1/2}\frac{|\boldsymbol{\Omega}^{-1}|^{1/2}}{|\boldsymbol{\Omega}^{-1}|^{1/2}}\,exp\Bigg(-\frac{1}{2}\Big[(\boldsymbol{\beta_c}-\boldsymbol{\mu})^T\,\boldsymbol{\Omega}\,(\boldsymbol{\beta_c}-\boldsymbol{\mu})+\boldsymbol{R}\Big]\Bigg). \tag{6.18}$$

Where,

$$\boldsymbol{\Omega} = \boldsymbol{W}^{-1}+\boldsymbol{\Sigma_c}. \tag{6.19}$$

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{\Omega}^{-1}\,\mathcal{H}^T\\ &= \boldsymbol{\Omega}^{-1}\,(\boldsymbol{\Sigma_c}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}\;+\;\boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}})\\ &= \boldsymbol{\Omega}^{-1}\,(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}} \qquad \text{where } \boldsymbol{V}_*^{-1}=\boldsymbol{V}^{-1}_{p\times||c||_1}.\end{aligned} \tag{6.20}$$

$\boldsymbol{V}^{-1}_{p\times||c||_1}$ is a submatrix of $\boldsymbol{V}^{-1}$ containing all rows of $\boldsymbol{V}^{-1}$ but only columns of $\boldsymbol{V}^{-1}$ where $c_j = 1$. From Equation (6.17)

$$\begin{aligned}\boldsymbol{R} &= \hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}\;-\mathcal{H}\,\boldsymbol{\Omega}^{-1}\,\mathcal{H}^T\\ &= \hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}\;-\;\hat{\boldsymbol{\beta}}^T\boldsymbol{V}_*^{-1}\,(\boldsymbol{\Omega}^{-1})\,(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}}\\ &= \hat{\boldsymbol{\beta}}^T\left[\boldsymbol{V}^{-1}-\boldsymbol{V}_*^{-1}\,(\boldsymbol{\Omega}^{-1})\,(\boldsymbol{V}_*^{-1})^T\right]\,\hat{\boldsymbol{\beta}}.\end{aligned} \tag{6.21}$$

By integrating Equation (6.18) over $\boldsymbol{\beta_c}$, we can obtain the marginal likelihood with a Gaussian prior as follows

$$
\begin{aligned}
& P(\hat{\boldsymbol{\beta}}|\mathcal{M}_c) \\
&= \int_{\boldsymbol{\beta}_c} f(\boldsymbol{\beta}_c\,|\mathcal{M}_c)\, f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_c\,, \mathcal{M}_c) \;\; d\boldsymbol{\beta}_{\boldsymbol{c}} \\
&= \int_{\boldsymbol{\beta}_c} \left[(2\pi)^{p+||c||_1}\,|\boldsymbol{V}|\;|\boldsymbol{W}|\right]^{-1/2} \frac{|\boldsymbol{\Omega}^{-1}|^{1/2}}{|\boldsymbol{\Omega}^{-1}|^{1/2}}\, exp\bigg(-\frac{1}{2}\Big[(\boldsymbol{\beta}_{\boldsymbol{c}}-\boldsymbol{\mu})^T\,\boldsymbol{\Omega}\,(\boldsymbol{\beta}_{\boldsymbol{c}}-\boldsymbol{\mu})+\boldsymbol{R}\Big]\bigg) d\boldsymbol{\beta}_{\boldsymbol{c}} \\
&= \left[(2\pi)^p\,|\boldsymbol{V}|\;|\boldsymbol{W}|\,|\boldsymbol{\Omega}|\right]^{-1/2} \exp\bigg(-\frac{\boldsymbol{R}}{2}\bigg) \int_{\boldsymbol{\beta}_c} \sqrt{(2\pi)^{-||c||_1}\,|\boldsymbol{\Omega}|}\, exp\bigg(-\frac{1}{2}\Big[(\boldsymbol{\beta}_{\boldsymbol{c}}-\boldsymbol{\mu})^T\,\boldsymbol{\Omega}\,(\boldsymbol{\beta}_{\boldsymbol{c}}-\boldsymbol{\mu})\Big]\bigg) d\boldsymbol{\beta}_{\boldsymbol{c}} \\
&= \left[(2\pi)^p\,|\boldsymbol{V}|\;|\boldsymbol{W}|\,|\boldsymbol{\Omega}|\right]^{-1/2} \exp\bigg(-\frac{\boldsymbol{R}}{2}\bigg) \qquad (6.22)
\end{aligned}
$$

6.6 Deriving $P(\hat{\boldsymbol{\beta}}|\mathcal{M}_{\mathbf{c}})$ using the Laplace prior

A Laplace prior on the slab component gives

$$\boldsymbol{\beta}_{\mathbf{c}}|\mathcal{M}_{\mathbf{c}} \sim \prod_{j:c_j=1} La\,(\beta_j;\lambda)$$

with pdf

$$f(\boldsymbol{\beta}_c|\,\mathcal{M}_c) = \left(\frac{\lambda}{2}\right)^{||c||_1} exp\,(-\lambda \boldsymbol{A}^T\boldsymbol{\beta}_{\mathbf{c}}) \qquad (6.23)$$

where $\boldsymbol{A}$ is a $||c||_1 \times 1$ vector with j^{th} element

$$A_j = \begin{cases} -1 & \text{if } \beta_j < 0 \\ 1 & \text{if } \beta_j > 0. \end{cases} \qquad (6.24)$$

From Equation (6.4) we integrate $f(\boldsymbol{\beta_c}\,|\mathcal{M}_\mathbf{c})\,f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta_c}\,,\mathcal{M}_\mathbf{c})$ over $\boldsymbol{\beta_c}$ to obtain $P(\hat{\boldsymbol{\beta}}|\mathcal{M}_\mathbf{c})$. Using the Laplace prior in Equation (6.23) and the likelihood in Equation (6.11), we have

$$\begin{aligned}
&f(\boldsymbol{\beta_c}\,|\mathcal{M}_\mathbf{c})\,f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta_c}\,,\mathcal{M}_\mathbf{c}) \\
&= \left(\frac{\lambda}{2}\right)^{||c||_1} \left[(2\pi)^p\,|\boldsymbol{V}|\right]^{-1/2} exp\left(-\frac{1}{2}\left[\boldsymbol{\beta_c}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c} - 2(\hat{\boldsymbol{\beta}}_\mathbf{c}^{\,T}\,\boldsymbol{\Sigma_c} + \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T)\boldsymbol{\beta_c} + \boldsymbol{Q}\right] - \lambda \boldsymbol{A}^T\boldsymbol{\beta_c}\right) \\
&= \left(\frac{\lambda}{2}\right)^{||c||_1} \left[(2\pi)^p\,|\boldsymbol{V}|\right]^{-1/2} exp\left(-\frac{1}{2}\left[\boldsymbol{\beta_c}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c} - 2(\hat{\boldsymbol{\beta}}_\mathbf{c}^{\,T}\,\boldsymbol{\Sigma_c} + \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T - \lambda\boldsymbol{A}^T)\boldsymbol{\beta_c} + \boldsymbol{Q}\right]\right).
\end{aligned} \tag{6.25}$$

The exponent term in Equation (6.25) can be expressed as,

$$\begin{aligned}
&\boldsymbol{\beta_c}^T\,\boldsymbol{\Sigma_c}\,\boldsymbol{\beta_c} - 2(\hat{\boldsymbol{\beta}}_\mathbf{c}^{\,T}\,\boldsymbol{\Sigma_c} + \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T - \lambda\boldsymbol{A}^T)\boldsymbol{\beta_c} + \boldsymbol{Q} \\
&= \boldsymbol{\beta_c}^T\boldsymbol{\nu}\boldsymbol{\beta_c} - 2\mathcal{K}\boldsymbol{\beta_c} + \boldsymbol{Q}
\end{aligned} \tag{6.26}$$

where $\boldsymbol{\nu} = \boldsymbol{\Sigma_c}$ and $\mathcal{K} = \hat{\boldsymbol{\beta}}_\mathbf{c}^{\,T}\,\boldsymbol{\Sigma_c} + \hat{\boldsymbol{\beta}}_{\boldsymbol{N}}^T\,\boldsymbol{\Sigma}^T - \lambda\boldsymbol{A}^T$. Thus, to express Equation (6.25) as a Gaussian Kernel, we put Equation (6.26) into the form

$$\begin{aligned}
&(\boldsymbol{\beta_c} - \boldsymbol{\mu})^T\,\boldsymbol{\nu}\,(\boldsymbol{\beta_c} - \boldsymbol{\mu}) + \boldsymbol{T} \\
&= \boldsymbol{\beta_c}^T\boldsymbol{\nu}\boldsymbol{\beta_c} - 2\boldsymbol{\mu}^T\boldsymbol{\nu}\boldsymbol{\beta_c} + \boldsymbol{\mu}^T\boldsymbol{\nu}\boldsymbol{\mu} + \boldsymbol{T}.
\end{aligned} \tag{6.27}$$

Equating Equation (6.26) and Equation (6.27) gives

$$\begin{aligned}
\mathcal{K} &= \boldsymbol{\mu}^T\boldsymbol{\nu} \\
\mathcal{K}^T &= \boldsymbol{\nu}^T\boldsymbol{\mu} \\
&= \boldsymbol{\nu}\boldsymbol{\mu} \quad \text{since} \quad \boldsymbol{\nu}^T \text{ is symmetric.}
\end{aligned}$$

Therefore,

$$\boldsymbol{\mu} = \boldsymbol{\nu}^{-1}\,\mathcal{K}^T$$

$$\boldsymbol{\mu}^T = \mathcal{K}\,\boldsymbol{\nu}^{-1}$$

and

$$\begin{aligned}\boldsymbol{\mu}^T\boldsymbol{\nu}\boldsymbol{\mu} &= (\mathcal{K}\,\boldsymbol{\nu}^{-1})\boldsymbol{\nu}\boldsymbol{\nu}^{-1}\,\mathcal{K}^T \\ &= \mathcal{K}\,\boldsymbol{\nu}^{-1}\,\mathcal{K}^T. \end{aligned} \tag{6.28}$$

Since $\boldsymbol{Q} = \boldsymbol{\mu}^T\boldsymbol{\nu}\boldsymbol{\mu} + \boldsymbol{T}$, it follows that

$$\boldsymbol{T} = \boldsymbol{Q} - \boldsymbol{\mu}^T\boldsymbol{\nu}\boldsymbol{\mu}.$$

Substituting $\boldsymbol{Q}$ from Equation (6.10) and $\boldsymbol{\mu}^T\boldsymbol{\nu}\boldsymbol{\mu}$ from Equation (6.28) into $\boldsymbol{T}$ gives

$$\boldsymbol{T} = \hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}\; - \mathcal{K}\,\boldsymbol{\nu}^{-1}\,\mathcal{K}^T. \tag{6.29}$$

Equation (6.25) can be expressed in Gaussian Kernel form as follows

$$\left(\frac{\lambda}{2}\right)^{||c||_1}\;\left[(2\pi)^p\,|\boldsymbol{V}|\right]^{-1/2}\;\frac{|\boldsymbol{\nu^{-1}}|^{1/2}}{|\boldsymbol{\nu^{-1}}|^{\mathbf{1/2}}}\,exp\left(\;-\frac{1}{2}\left[(\boldsymbol{\beta_c}-\boldsymbol{\mu})^T\,\boldsymbol{\nu}\,(\boldsymbol{\beta_c}-\boldsymbol{\mu})\;+\;\boldsymbol{T}\right]\right). \tag{6.30}$$

Where,

$$\boldsymbol{\nu} = \boldsymbol{\Sigma_c} \tag{6.31}$$

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{\nu}^{-1}\mathcal{K}^T \\ &= \boldsymbol{\nu}^{-1}\left(\boldsymbol{\Sigma_c}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} + \boldsymbol{\Sigma}\,\hat{\boldsymbol{\beta}}_{\boldsymbol{N}} - \lambda\boldsymbol{A}\right) \\ &= \boldsymbol{\nu}^{-1}\left[(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}} - \lambda\boldsymbol{A}\right]\end{aligned} \tag{6.32}$$

and from Equation (6.29)

$$\begin{aligned}\boldsymbol{T} &= \hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}} - \mathcal{K}\,\boldsymbol{\nu}^{-1}\,\mathcal{K}^T \\ &= \hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}} - \left[\hat{\boldsymbol{\beta}}^T\boldsymbol{V}_*^{-1} - \lambda\boldsymbol{A}^T\right](\boldsymbol{\nu}^{-1})\left[(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}} - \lambda\boldsymbol{A}\right] \\ &= \hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}} - \left[\hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}_*^{-1}\,\boldsymbol{\nu}^{-1}\,(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^T\boldsymbol{V}_*^{-1}\,\boldsymbol{\nu}^{-1}\,\lambda\,\boldsymbol{A} - \lambda\,\boldsymbol{A}^T\,\boldsymbol{\nu}^{-1}\,(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}} + \lambda^2\boldsymbol{A}^T\,\boldsymbol{\nu}^{-1}\,\boldsymbol{A}\right].\end{aligned}$$

$\hat{\boldsymbol{\beta}}^T\boldsymbol{V}_*^{-1}\,\boldsymbol{\nu}^{-1}\,\boldsymbol{A}$ is equal to $\boldsymbol{A}^T\,\boldsymbol{\nu}^{-1}\,(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}}$ since they are scalars and the matrix $\boldsymbol{\nu}^{-1}$ symmetric. This gives

$$\begin{aligned}\boldsymbol{T} &= \hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}} - \left[\hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}_*^{-1}\,\boldsymbol{\nu}^{-1}\,(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}} - 2\lambda\hat{\boldsymbol{\beta}}^T\boldsymbol{V}_*^{-1}\,\boldsymbol{\nu}^{-1}\,\boldsymbol{A} + \lambda^2\boldsymbol{A}^T\,\boldsymbol{\nu}^{-1}\,\boldsymbol{A}\right] \\ &= \hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^T\,\mathcal{U}\,(\boldsymbol{V}_*^{-1})^T\,\hat{\boldsymbol{\beta}} + 2\lambda\hat{\boldsymbol{\beta}}^T\mathcal{U}\,\boldsymbol{A} - \lambda^2\boldsymbol{A}^T\,\boldsymbol{\nu}^{-1}\,\boldsymbol{A} \\ &= \hat{\boldsymbol{\beta}}^T[\boldsymbol{V}^{-1} - \mathcal{U}\,(\boldsymbol{V}_*^{-1})^T]\hat{\boldsymbol{\beta}} + 2\lambda\hat{\boldsymbol{\beta}}^T\mathcal{U}\,\boldsymbol{A} - \lambda^2\boldsymbol{A}^T\,\boldsymbol{\nu}^{-1}\,\boldsymbol{A}\end{aligned} \tag{6.33}$$

where $\mathcal{U} = \boldsymbol{V}_*^{-1}\,\boldsymbol{\nu}^{-1}$.

The marginal likelihood using a Laplace prior can be obtained by integrating Equation (6.30) over $\boldsymbol{\beta}_c$ as follows

$$P(\hat{\boldsymbol{\beta}}|\mathcal{M}_c) = \int_{\boldsymbol{\beta_c}} f(\boldsymbol{\beta}_c\,|\mathcal{M}_c)\, f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_c\,, \mathcal{M}_c) \quad d\boldsymbol{\beta_c}$$

$$= \int_{\beta_1 \in \mathbb{R}} \int_{\beta_2 \in \mathbb{R}} \cdots \int_{\beta_{||c||_1} \in \mathbb{R}} f(\boldsymbol{\beta}_c\,|\mathcal{M}_c)\, f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_c\,, \mathcal{M}_c) \quad d\beta_1 d\beta_2 \cdots d\beta_{||c||_1} \tag{6.34}$$

Equation (6.34) integrates over the space, $\mathbb{R}^{||c||_1}$. However, in our case, the likelihood with a Laplace prior depends on the vector $\boldsymbol{A}$ in Equation (6.24). This affects the limits of the integration. For $A_j = 1$, the integration is over all the positive real numbers, $\mathbb{R}_{\geq 0}$ while for $A_j = -1$, the limits are from negative infinity to zero. For a specific model, $\mathcal{M}_{\mathbf{c}}$, there are $2^{||c||_1}$ combinations of $\boldsymbol{A}$. Thus, the marginal likelihood is

$$P(\hat{\boldsymbol{\beta}}|\mathcal{M}_c) = \sum_{\boldsymbol{A} \in \{-1,1\}^{||c||_1}} \int_{\beta_1 \in \theta_1} \int_{\beta_2 \in \theta_2} \cdots \int_{\beta_{||c||_1} \in \theta_{||c||_1}} f(\boldsymbol{\beta}_c\,|\mathcal{M}_c)\, f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_c\,, \mathcal{M}_c) \quad d\beta_1 d\beta_2 \cdots d\beta_{||c||_1} \tag{6.35}$$

where

$$\theta_j = \begin{cases} (-\infty, 0) & \text{if } A_j = -1 \\ (0, \infty) & \text{if } A_j = 1. \end{cases}$$

and $f(\boldsymbol{\beta}_c \,|\, \boldsymbol{\beta}_c, \mathcal{M}_c)$ depends on $\boldsymbol{A}$. From Equation (6.30), we factor out expressions which do not

depend on $\boldsymbol{\beta}_c$ in order to simplify the integration. Hence,

$$
\begin{aligned}
& f(\boldsymbol{\beta}_c \,|\mathcal{M}_c)\, f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_c\,, \mathcal{M}_c) \\
&= \left(\frac{\lambda}{2}\right)^{||c||_1} \left[(2\pi)^p \,|\boldsymbol{V}|\right]^{-1/2} \frac{|\boldsymbol{\nu^{-1}}|^{\mathbf{1/2}}}{|\boldsymbol{\nu^{-1}}|^{\mathbf{1/2}}}\, exp\left(-\frac{1}{2}\left[(\boldsymbol{\beta_c}-\boldsymbol{\mu})^T\, \boldsymbol{\nu}\,(\boldsymbol{\beta_c}-\boldsymbol{\mu}) \,+\, \boldsymbol{T}\right]\right) \\
&= \left(\frac{\lambda}{2}\right)^{||c||_1} \left[(2\pi)^{p+||c||_1-||c||_1}\,|\boldsymbol{V}|\,\frac{|\boldsymbol{\nu}|}{|\boldsymbol{\nu}|}\right]^{-\frac{1}{2}} \; exp\left(-\frac{1}{2}\left[(\boldsymbol{\beta_c}-\boldsymbol{\mu})^T\, \boldsymbol{\nu}\,(\boldsymbol{\beta_c}-\boldsymbol{\mu}) \,+\, \boldsymbol{T}\right]\right) \\
&= \left(\frac{\lambda}{2}\right)^{||c||_1} \left[(2\pi)^{p-||c||_1}|\boldsymbol{V}||\boldsymbol{\nu}|\right]^{-\frac{1}{2}} exp\left(\frac{-\boldsymbol{T}}{2}\right)\sqrt{(2\pi)^{-||c||_1}\,|\boldsymbol{\nu}|}\; exp\left(-\frac{1}{2}\left[(\boldsymbol{\beta_c}-\boldsymbol{\mu})^T\, \boldsymbol{\nu}\,(\boldsymbol{\beta_c}-\boldsymbol{\mu})\right]\right)
\end{aligned}
\tag{6.36}
$$

We integrate $f(\boldsymbol{\beta}_c \,|\mathcal{M}_c)\, f(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_c\,, \mathcal{M}_c)$ in Equation (6.35) by using the expression in Equation (6.36). This gives the marginal likelihood for a model using a Laplace prior as follows

$$
P(\hat{\boldsymbol{\beta}}|\mathcal{M}_c) = \;\boldsymbol{J} \sum_{\boldsymbol{A}\in\{-1,1\}^{||c||_1}} \boldsymbol{Y} \int_{\beta_1\in\theta_1}\int_{\beta_2\in\theta_2}\cdots\int_{\beta_{||c||_1}\in\theta_{||c||_1}} \boldsymbol{Z}\; d\beta_1 d\beta_2\cdots d\beta_{||c||_1} \tag{6.37}
$$

where

$$
\begin{aligned}
\boldsymbol{J} &= \left(\frac{\lambda}{2}\right)^{||c||_1}\left[(2\pi)^{p-||c||_1}|\boldsymbol{V}||\boldsymbol{\nu}|\right]^{-\frac{1}{2}}, \\
\boldsymbol{Y} &= \exp\left(-\frac{\boldsymbol{T}}{2}\right)
\end{aligned}
$$

and

$$
\boldsymbol{Z} = \sqrt{(2\pi)^{-||c||_1}\,|\boldsymbol{\nu}|}\; exp\left(-\frac{1}{2}\left[(\boldsymbol{\beta_c}-\boldsymbol{\mu})^T\, \boldsymbol{\nu}\,(\boldsymbol{\beta_c}-\boldsymbol{\mu})\right]\right).
$$

Integration in Equation (6.37) can be evaluated directly in R using the mvtnorm package. Using pmvnorm in this package allows us to compute multivariate normal probabilities functions with dif-

ferent limits.

6.7 Calculate posterior probability for each model $P(\mathcal{M}_{\mathbf{c}} \mid \hat{\boldsymbol{\beta}})$

In Section 6.2, we described how to calculate $P(\hat{\boldsymbol{\beta}}|\mathcal{M}_{\mathbf{c}})$, the probability density of the observed data under a specific model, $\mathcal{M}_c$. We use this to calculate $P(\mathcal{M}_{\mathbf{c}} \mid \hat{\boldsymbol{\beta}})$. In the beginning of this chapter, we mentioned there are 2^p possible models including the null model with all 0 values in $\boldsymbol{c}$, i.e when there are no causal SNP present in the model. Computing the posterior probability for the null model, $\mathcal{M}_0$ is straight forward. Since there is no causal SNP, the joint prior probability of all the zero effect sizes is equal to 1 because the prior takes the spike component. For the null model, $\mathcal{M}_0$, all effect sizes of the SNPs are zero. For this reason, $P(\hat{\boldsymbol{\beta}}|\mathcal{M}_0)$ is simply the probability density of the observed data under the null model where $\boldsymbol{\beta} = \mathbf{0}$. The observed data is normally distributed as mentioned in Section 6.3 with $\mathbf{0}$ mean and gives

$$P(\hat{\boldsymbol{\beta}}|\mathcal{M}_0) = \sqrt{(2\pi)^{-p}\,|\boldsymbol{V}^{-1}|}\; exp\bigg(-\frac{1}{2}\hat{\boldsymbol{\beta}}^T\,\boldsymbol{V}^{-1}\,\hat{\boldsymbol{\beta}}\bigg). \tag{6.38}$$

The posterior probability, $P(\mathcal{M}_c \mid \hat{\boldsymbol{\beta}})$ for each model is shown in Equation 6.1 which requires the joint probability in the numerator and the probability of the data, $P(\hat{\boldsymbol{\beta}})$, in the denominator. The numerator joint probability can be obtained by multiplying the prior probability of the model as discussed in Section 6.1, with the marginal likelihood of the data as discussed in Section 6.5 for the Gaussian prior or in Section 6.6 for the Laplace prior. For the null model, $\mathcal{M}_0$ the prior model, $P(\mathcal{M}_0)$ can be obtained by letting $k = 0$ in Equation (6.3).

To obtain the denominator, $P(\hat{\boldsymbol{\beta}})$, we sum over the joint probability of all models including the null model. The posterior probability of a specific model, $\mathcal{M}_c$ can be extended from Equation (6.1) as

follows

$$P(\mathcal{M}_c \,|\, \hat{\boldsymbol{\beta}}) = \frac{P(\mathcal{M}_c)\, P(\hat{\boldsymbol{\beta}}|\mathcal{M}_c)}{P(\hat{\boldsymbol{\beta}})}$$
$$= \frac{P(\mathcal{M}_c)\, P(\hat{\boldsymbol{\beta}}|\mathcal{M}_c)}{P(\mathcal{M}_0)\, P(\hat{\boldsymbol{\beta}}|\mathcal{M}_0) + \sum_{\mathcal{M}_c \in \mathcal{M}} P(\mathcal{M}_c)\, P(\hat{\boldsymbol{\beta}}|\mathcal{M}_c)} \tag{6.39}$$

where $\mathcal{M}$ is the set of all models allowed which depends on the maximum number of causal SNPs allowed in the model.

Chapter 7

Application of the multivariate approaches to simulated data

7.1 A description of the simulated data used

Similar to testing the methods in the univariate approach in Chapter 3, we simulate data from HAPGEN2 (Su et al., 2011) in order to test the approaches we discussed in Chapter 6. In our analysis, we look into a scenario which includes two causal SNPs in the region. The elements that we considered in simulating the data are the odds ratio and the MAF of the two causal SNPs, the sample size and the marginal power. By specifying the values of the OR and MAF, the sample size needed to achieve a given power for the analysis could be determined following the calculation to compute power in Equation (4.1).

The first element we considered is the odds ratio for the two causal SNPs. The selection of the odds ratio was based on the intersection between the Gaussian and Laplace distribution as shown in Figures 7.1 and 7.2. We wanted to assess whether the Laplace prior would actually yield much of an improvement over the Gaussian prior for log ORs that have similar prior probability densities under both priors. This is not guaranteed since it depends on the MLEs and likelihood variance. To plot the

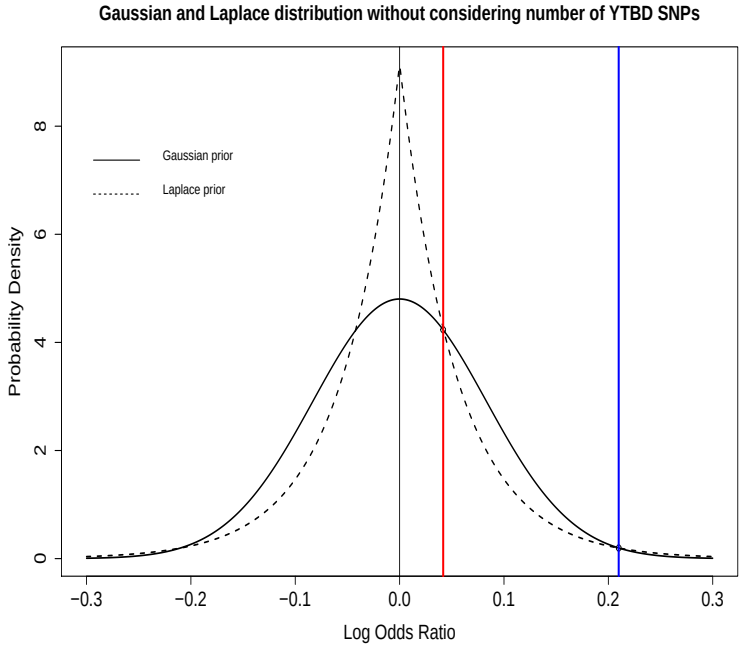


***Figure 7.1*:** The minimum OR (in red) and the maximum OR (in blue) where the Gaussian prior has higher density than the Laplace prior distributions plotted using MLEs without considering the number of yet-to-be-discovered SNPs. Only positive log odds ratios are considered.

distributions, we used the MLE of the parameters by varying the number of YTBD SNPs as given in Table 3.1 for the Gaussian distribution and Table 3.3 for the Laplace distribution. In Figure 7.1, the distributions are plotted using the MLEs without considering the number of YTBD SNPs meanwhile in Figure 7.2, the MLEs used do consider the number of YTBD SNPs (chosen to be 250, 500,750 and 1000).

In each plot, we determine the odds ratio using the points where both distributions intersect. The minimum OR and the maximum OR where the Gaussian prior has a higher density for each plot are summarised in Table 7.1. As we vary the number of YTBD SNPs, the minumum ORs range from 1.018 up to 1.043, meanwhile the maximum ORs have the highest value of 1.234 and the lowest value of 1.135. We do not want to choose odds ratios too small or too large otherwise it be difficult to choose a sample size that yields reasonable power for both SNPs. From the range of the maximum and minimum ORs, we select the OR to be 1.03 and 1.13 for each of the causal SNPs. These ORs are

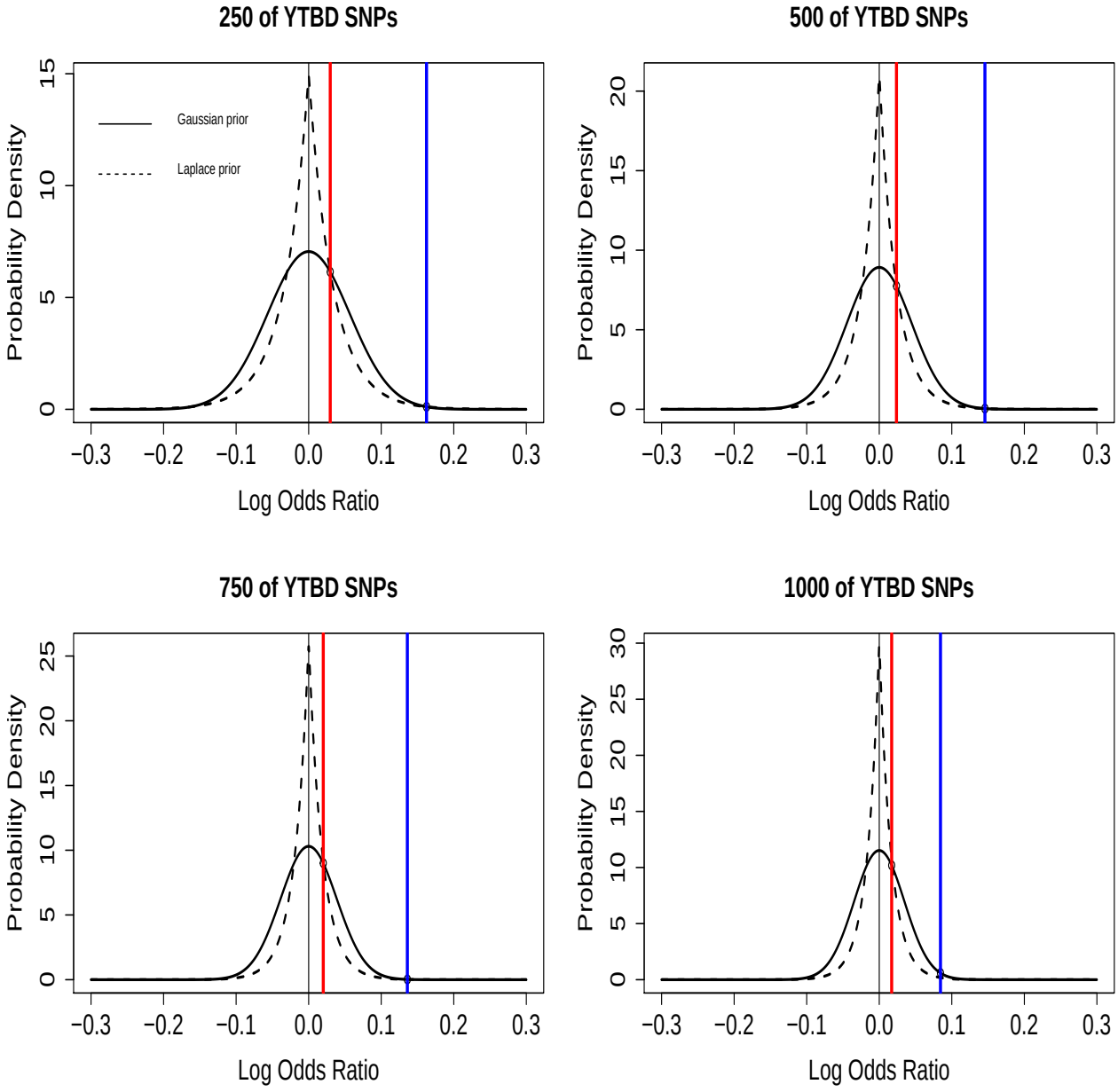


***Figure 7.2*:** The minimum OR (in red) and the maximum OR (in blue) where the Gaussian prior has higher density than the Laplace prior distributions plotted using MLEs by taking into account 250, 500,750 and 1000 of yet-to-be-discovered SNPs. Only positive log odds ratios are considered.

between the minimum and the maximum ORs in the scenarios in the last four rows of Table 7.1.

***Table 7.1*:** The maximum and the minimum odds ratio from the intersection of Gaussian and Laplace distribution by varying the number of yet-to-be-discovered SNPs.

no. YTBD SNPS	$\hat{W}$	$\hat{\lambda}$	**minimum OR**	**maximum OR**
not considered	0.0069	18.31	1.0426	1.2336
250	0.0032	30.05	1.0302	1.1764
500	0.002	42.41	1.024	1.1568
750	0.0015	52.27	1.0203	1.1457
1000	0.0012	60.47	1.0179	1.1348

The next element to consider is the MAF for the two causal SNPs. We defined one of the causal SNP as a common SNP with a MAF of 0.5. The other causal SNP has a MAF of 0.03 and is termed as a rare SNP. To achieve reasonable power, based on Equation (4.1), we assigned the smaller OR to the common SNP and the larger OR to the rare SNP. Thus, the common causal SNP has an OR of 1.03 with a MAF equal to 0.5. Meanwhile, the rare causal SNP has an OR equal to 1.13 and a MAF of 0.03.

After specifying the ORs and MAFs for both causal SNPs, we need to decide on the sample size for our simulated data. We defined the total sample size as the total number of cases and the total number of controls, with equal numbers for both cases and controls. Thus, our simulated data has a sample size of 70000 cases and 70000 controls. We wanted the sample size to be large enough to be able to detect association at both loci without being too large to give too high a power at either locus.

In our simulated data, the marginal power for the common causal SNP is 0.53 while for the rare SNP, the power is 0.82. We simulated 10 datasets each containing 412 SNPs. We use the same region as for our univariate analysis. However, there are extremely rare and monomorphic SNPs which we need to remove from the simulated data. We only retain SNPs that are neither monomorphic nor rare (MAF ≤ 0.01) in every dataset. Another factor that we observed is multicollinearity. If two SNPs are highly correlated ($r^2 \geq 0.99$), one of the SNP is eliminated from the data. We also reduce the number of SNPs in the data by removing SNPs having very small Wakefield Bayes factor. This is to

reduce the computational time of our method. SNPs with very small Bayes factor are not going to be selected so there seems little point in including them and these Bayes factors are not going to be generated from SNPs with moderate effect sizes. According to Kass and Raftery (1995), they interpret a Bayes factor of more than three to mean strong evidence against the null hypothesis. However, if we use a Bayes factor of three as our threshold, the number of SNPs reduced to only 10 SNPs, which is too small. Thus, we reduced the Bayes factor threshold to a smaller value (0.4) to allow more SNPs to be included. The number of SNPs remaining in the data is 50. A Bayes factor of 0.4 makes the null hypothesis two and a half times more likely than the alternative hypothesis so this seems like a reasonable threshold to apply. The causal SNPs in the simulated data are summarized in Table 7.2.

***Table 7.2*:** The odds ratio, minor allele frequency (MAF) and marginal power for the two causal SNPs specified in the simulated data with a sample size of 70000 cases and 70000 controls

Elements	Causal SNP 1	Causal SNP 2
Odd Ratio	1.03	1.13
MAF	0.5	0.02
Marginal power	0.53	0.82
SNP number	1	50

7.2 Comparing the performance of the Laplace and Gaussian priors with FINEMAP

We are interested in comparing the performance of the Bayesian model selection method using the Gaussian and Laplace prior derived in Chapter 6. Another interest is to compare the performance of the two methods with FINEMAP (Benner et al., 2016), one of the current packages available for fine-mapping. FINEMAP has been shown to be the gold standard in many fine mapping situations (Benner et al., 2016). The FINEMAP software requires the Z-scores from each SNP and the LD between SNPs as the input. FINEMAP allows us to specify the maximum number of causal SNPs from 1 to 5. Users can obtain the posterior probabilities for each SNP and also its Bayes factor. However, in this

section, we only focus on using the posterior probabilities for ranking purposes. We use $\hat{\lambda} = 64.15$ to calculate the posterior probabilities using Laplace prior since using this value in single-SNP analysis in Section 4.3.1 had shown that the Laplace Bayes factor has the best performance compared to other values. $\hat{\lambda} = 64.15$ corresponds to the upper limit of the confidence level of $\hat{\lambda}$ when considering 1000 YTBD SNPs. Thus, to calculate the posterior probabilities using Gaussian prior, we specify W equal to 0.0011 as this is the MLE for W when considering 1000 YTBD SNPs. We also used the same value of W in FINEMAP.

In Chapter 6, we showed how to calculate the posterior probability of each model (Equation 6.39). To calculate the posterior probability of causal association for each SNP, we sum the posterior probabilities of all models containing that SNP. By using the posterior probability, we can examine the ranking performance of the Gaussian prior, Laplace prior and FINEMAP. We are able to compute the posterior probabilities of each SNP for every method using the simulated data discussed in Section 7.1. The maximum number of causal SNPs allowed in the model are varied from one to five when computing the posterior probabilities to assess the effect of varying this limit. The probability ω in Equation 6.3 is taken to be one fiftieth since we have 50 SNPs in the data. Table 7.3 shows the prior probabilities of the number of causal SNPs given $\omega = 1/50$. We illustrate the ranking performance of each method by using ROC curves. The true positive rates (TPRs) and false positive rates (FPRs) are calculated using the posterior probability and plotted in a ROC curve. Since we simulated 10 datasets, we used the vertical averaging method introduced by Fawcett (2006) to plot the ROC curves. The ROC curves for each method are compared on the same plot.

Figure 7.3 shows five comparisons of ROC curves by varying the maximum number of causal SNPs. All the ROC curves shown in Figure 7.3 are plotted using FPRs up to 20 $\%$. In the case when we allowed only one causal SNP in the model, the performance of each method is not very distinctive. However, when more causal SNPs are allowed in the model, the performance of every method is distinguishable. In most cases, the Gaussian prior shows poor performance compared to the Laplace prior and FINEMAP. When we increase the maximum number of causal SNPs to two, the Laplace

***Table 7.3*:** The prior probabilities of the number of causal SNPs in the model

		Prior probability ($\omega = 1/50$)					
	Number of SNPs	**0**	**1**	**2**	**3**	**4**	**5**
Maximum number of causal SNPs	**1 causal**	0.4949	0.5051	0	0	0	0
	2 causals	0.3952	0.4031	0.2016	0	0	0
	3 causals	0.3708	0.3783	0.1892	0.0618	0	0
	4 causals	0.3653	0.3728	0.1864	0.0609	0.0146	0
	5 causals	0.3643	0.3718	0.1859	0.0607	0.01456	0.002733

prior starts to perform better than FINEMAP and the Gaussian prior. The Laplace prior continues to have better performance than the other two methods when the maximum number of causal SNPs increases to three, four and five. As a result, we conclude that the Laplace prior is the best method to rank SNPs based on posterior probability of each SNP for this particular scenario.

7.3 Comparing the posterior probabilities using the Laplace prior according to the maximum number of causal SNPs specified

The method that used the Laplace prior appeared to be the best among the three methods in ranking SNPs based on posterior probabilities. Consequently, in this section, we continue to evaluate the performance of the Laplace prior. The posterior probabilities of each SNP are computed as we vary the maximum number of causal SNPs allowed in the model. We continue using the range from one to five for the maximum number of causal SNPs in the model, the same as in Section 7.2. Using ROC curves, we can observe the ranking performance of the Laplace prior as we change the maximum number of causal SNPs allowed in the model.

Figure 7.4 shows the ranking performance of the posterior probabilities using the Laplace prior (with $\hat{\lambda} = 64.15$) by allowing one, two, three, four and five causal SNPs in the model. The ROC curve when allowing only one causal SNP in the model shows a poor performance. The ranking performance gets better when the maximum number of causal SNP increases to two. This is presumably because

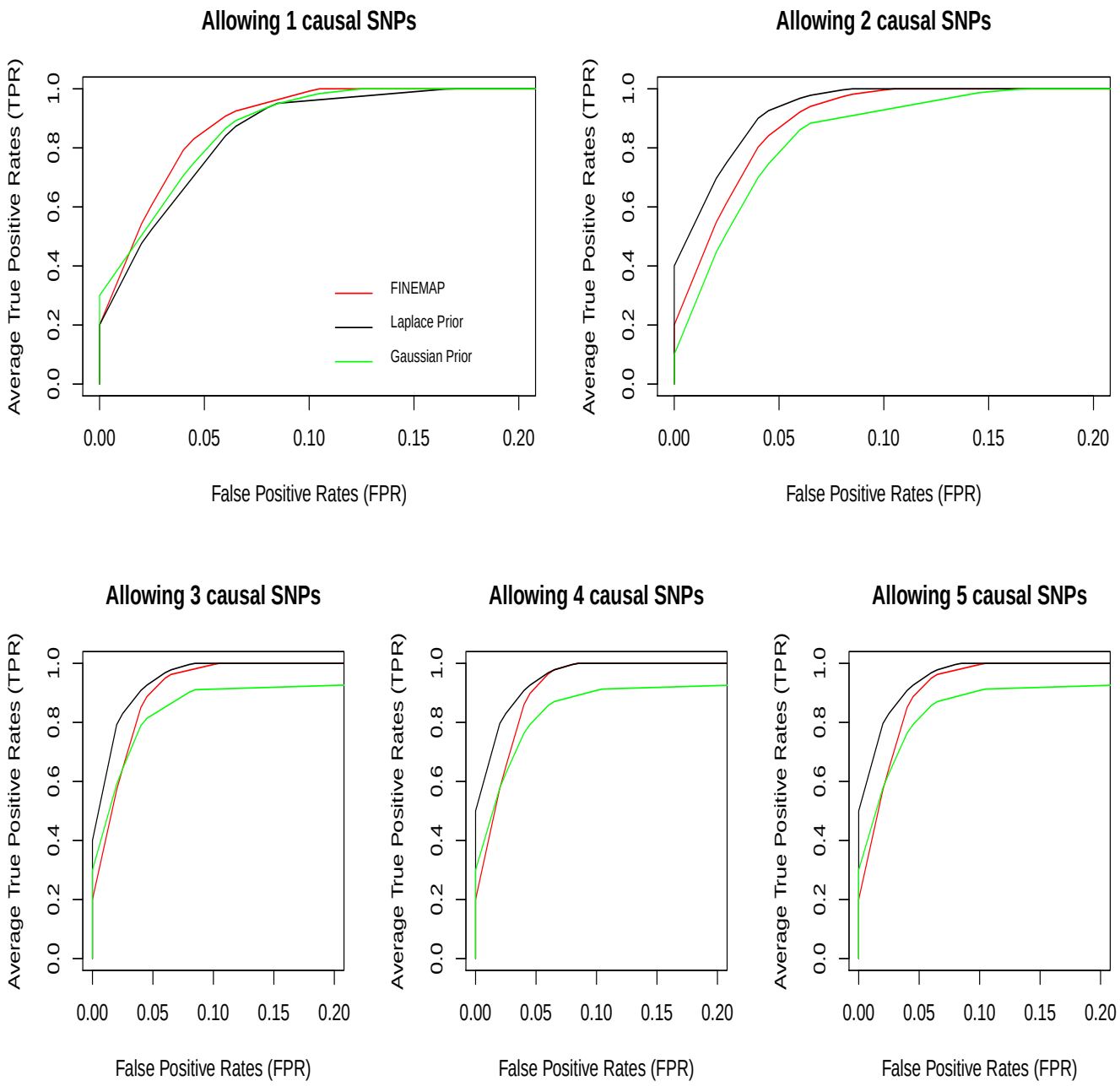


***Figure 7.3*:** ROC curves comparing the SNPs ranking performance of the posterior probabilities using three approaches; the Laplace prior, the Gaussian prior and FINEMAP. The maximum number of causal SNPs allowed in the model are varied from one to five to calculate the posterior probabilties in all approaches. The Laplace prior used $\lambda = 64.15$. The Gaussian prior and FINEMAP used $W = 0.0011$

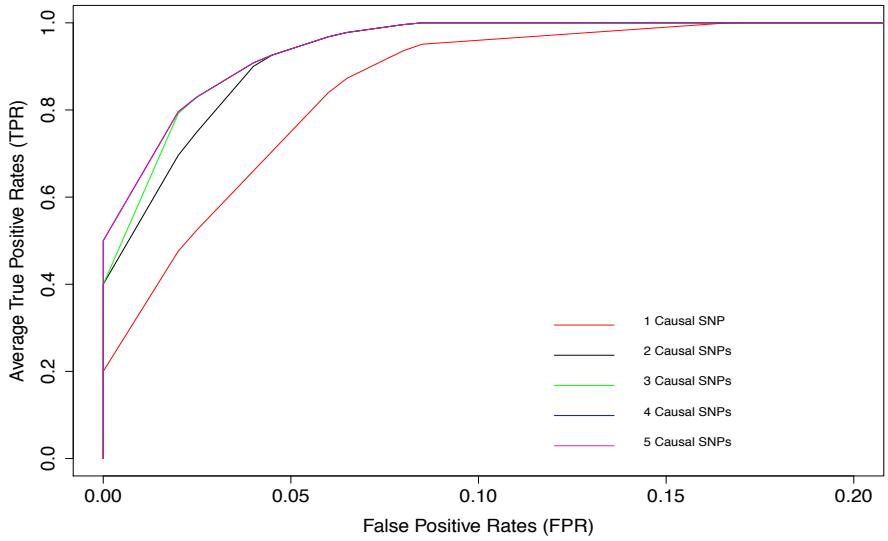


***Figure 7.4*:** ROC curves comparing the ranking performance of the posterior probabilities obtained with a Laplace prior as the maximum number of causal SNPs allowed in the model varies. The value of λ is 64.15.

there are 2 causal SNPs in the simulated data. As we increase the maximum number of causal SNPs to three, the ROC curve shows a better performance compared to the ROC curve when we specify a maximum of two causal SNPs. As we increase to a maximum of four causal SNPs, the ranking performance improves only marginally. However, as we increase it to five, the ROC curve overlap the ROC curve when we allow four causal SNPs. This shows that the ranking performance when we allow five causal SNPs remains the same as the performance when we allow four causal SNPs in the model. It seems as though there is no decrease in performance as we increase the maximum number of causal SNPs increase.

The ROC curve only considers the ranks. We further investigate if allowing three or more causal SNPs in the model changes the posterior probabilities of each SNP by very much. Datasets one to ten results in different posterior probabilities for each SNP. Thus, we average the posterior probabilities of each SNP across the 10 datatsets. The average posterior probabilities are presented in Table 7.4 as the maximum number of causal SNPs varies from three to five.

From the ROC curves in Figure 7.4, we see that the ranking of SNPs based on the SNPs individual

***Table 7.4*:** The posterior probabilities for the 50 SNPs selected based on the univariate Bayes factor when allowing a maximum of three, four and five causal SNPs in the model.

	Maximum no. of causal SNPs allowed in the model		
	3causal	**4causal**	**5causal**
SNP1	0.47416	0.48004	0.48180
SNP2	0.02295	0.02908	0.03147
SNP3	0.01411	0.01811	0.01974
SNP4	0.01409	0.01774	0.01921
SNP5	0.20760	0.21185	0.21333
SNP6	0.01667	0.02056	0.02206
SNP7	0.01350	0.01711	0.01858
SNP8	0.01279	0.01623	0.01762
SNP9	0.15100	0.15375	0.15485
SNP10	0.01410	0.01808	0.01971
SNP11	0.01558	0.01922	0.02067
SNP12	0.01181	0.01481	0.01603
SNP13	0.01309	0.01660	0.01803
SNP14	0.08079	0.08442	0.08581
SNP15	0.01179	0.01467	0.01583
SNP16	0.01158	0.01468	0.01595
SNP17	0.02085	0.02599	0.02803
SNP18	0.01553	0.01954	0.02113
SNP19	0.01620	0.02077	0.02268
SNP20	0.01438	0.01804	0.01951
SNP21	0.01319	0.01680	0.01827
SNP22	0.01795	0.02058	0.02165
SNP23	0.01456	0.01824	0.01975
SNP24	0.01502	0.01909	0.02072
SNP25	0.01241	0.01574	0.01709

	Maximum no. of causal SNPs allowed in the model		
	3 causals	**4 causals**	**5 causals**
SNP26	0.01746	0.02109	0.02251
SNP27	0.01539	0.01900	0.02045
SNP28	0.01386	0.01741	0.01884
SNP29	0.01410	0.01766	0.01911
SNP30	0.01435	0.01796	0.01941
SNP31	0.01287	0.01618	0.01753
SNP32	0.01225	0.01542	0.01670
SNP33	0.01614	0.02014	0.02171
SNP34	0.01207	0.01528	0.01659
SNP35	0.01468	0.01841	0.01994
SNP36	0.01744	0.02197	0.02377
SNP37	0.14291	0.15108	0.15383
SNP38	0.01430	0.01803	0.01953
SNP39	0.02356	0.02797	0.02967
SNP40	0.01347	0.01722	0.01876
SNP41	0.01365	0.01725	0.01871
SNP42	0.01284	0.01611	0.01744
SNP43	0.01301	0.01626	0.01757
SNP44	0.01700	0.02122	0.02291
SNP45	0.06163	0.06956	0.07244
SNP46	0.01484	0.01861	0.02012
SNP47	0.01247	0.01574	0.01709
SNP48	0.01339	0.01693	0.01839
SNP49	0.01276	0.01595	0.01723
SNP50	0.24001	0.25295	0.25704

posterior probabilities are indistinguishable in scenarios when we allow a maximum of four and five causal SNPs in the model. When analysing the posterior probabilities of every SNP across the three columns in Table 7.4, the posterior probabilities are different. Generally, as the maximum number of causal SNPs in the model increases, the posterior probability of a SNP also increases. We focus on one of the two causal SNPs in the simulated data, SNP 1, which is ranked first in all scenarios. The posterior probability of causal association of SNP 1 when the maximum number of causal SNP is three is 0.47416 and with a maximum of four causal SNPs in the model is 0.48004. The absolute difference between these two probabilities is 0.005878. However, when the maximum causal SNP in the model is five, the posterior probability of SNP 1 to be included only increases by 0.001761.

A similar result is obtained when we observe the posterior probabilities for all SNPs. The average absolute difference between posterior probabilities when allowing three and four causal SNPs in the model is 0.004. As we increase the number of maximum causal SNPs from four to five, on average, the absolute difference in posterior probabilities is 0.002. The difference is small, half the difference between allowing three and four causal SNPs in the model. These very small differences in the posterior probabilities show that either by allowing a maximum of four or five causal SNPs in the model, the posterior probabilities are still comparable.

Throughout this section, we worked with the posterior probabilities of causal association of each SNPs when we vary the maximum number of causal SNPs in the model. By observing the ranking performance, we could see that there was no difference in the ranks of the SNPs based on the posterior probabilities when there are four and five causal SNPs in the model. However, there are differences in the actual values of the posterior probabilities of each SNP as the maximum number of causal SNPs allowed increases from three to four and to five. Another quantity we can examine is the Bayes factor of each SNP since these can tell us about the noteworthiness of the SNPs.

7.4 Comparing the noteworthiness of SNPs using the Laplace prior according to the maximum number of causal SNPs specified

In addition to computing the posterior probabilities for each SNP, a Bayes factor can be computed for a SNP to quantify the evidence that the SNP has an association with the disease. The Bayes factor can be derived from the posterior probability of a SNP being causal and its prior probability. The Bayes factor is given as

$$\text{Bayes factor} = \frac{P(c_j = 1|\,\hat{\boldsymbol{\beta}}\,)}{P(c_j = 0|\,\hat{\boldsymbol{\beta}}\,)} \Big/ \frac{P(c_j = 1)}{P(c_j = 0)} \tag{7.1}$$

where $P(c_j = 1|\,\hat{\boldsymbol{\beta}}\,)$ is the posterior probability of a SNP being causal. The prior probability of as SNP being causal is given as

$$P(c_j = 1) = \sum_{k=1}^{K} \frac{{}^{p-1}C_{k-1}}{{}^{p}C_{k}}\ P_k. \tag{7.2}$$

P_k is the prior probability of k causal SNP in the model given by Equation (6.2).

Using the posterior probabilities obtained from Section 7.3, we compute the Bayes factor (Equation (7.1)) for each SNP in every scenario of the maximum number of causal SNPs allowed in the model. The Bayes factor can be used to asses the noteworthiness of a SNP. This can be defined by Equation (2.19). The posterior probability of H_0 can be computed using Equation (2.20) by specifying the prior odds of H_0 and by using the Bayes factor obtained. We vary the ratio of costs of making incorrect decisions ($R = C_\omega / C_\eta$) to explore the noteworthiness SNPs at different values of R. Table 7.5 summarises the SNPs picked up as noteworthy as we vary the values of R and the maximum number of causal SNPs allowed in the model.

The common causal SNP specified in datasets, (SNP 1) is shown to be noteworthy in all scenarios.

***Table 7.5*:** The noteworthy SNPs at different values of the maximum number of causal SNPs and ratios of costs of making incorrect decisions corresponds to the Bayesian false-discovery probabilities (BFDP).

Ratio of costs, R	**4**	**10**	**19**	**49**
BFDP	**0.8**	**0.91**	**0.95**	**0.98**
1 causal	{1,5,9,50}	{1,5,9,14,50}	{1,5,9,14,37,50}	{1,5,9,14,37,50}
2 causals	{1,5,9,50}	{1,5,9,14,37,50}	{1,5,9,14,37,50}	{1,5,9,14,37,39,45,50}
3 causals	{1,5,9,50}	{1,5,9,37,50}	{1,5,9,37,50}	{1,5,9,14,37,40,50}
4 causals	{1,9}	{1,5,9,37,50}	{1,5,9,37,50}	{1,5,9,14,37,50}
5 causals	{1}	{1,5,9,50}	{1,5,9,37,50}	{1,5,9,14,37,50}

As for SNP 50, the rare causal SNP specified, it appears to be noteworthy in all scenarios except the scenarios allowing four and five causal SNPs in the model with $R = 4$. In addition to these two specified causal SNPs, there are a number of SNPs appeared to be noteworthy. SNP 5 appeared to be noteworthy in all scenarios where SNP 50 is noteworthy. As for SNP 9, the only scenario where SNP 9 is non-noteworthy is when there are five causal SNPs in the model and $R = 4$. In most cases we can see that SNP 14 and SNP 37 are noteworthy alongside SNP 1, SNP 50, SNP 5 and SNP 9. There are three more SNPs declared to be noteworthy when $R = 49$. The SNPs are SNP 39 and SNP 45 in the model where there are 2 causal SNPs and SNP 40 in model with three causal SNPs.

From Table 7.5, we could observe an increasing pattern in the number of SNPs declared to be noteworthy when the ratio of cost, R increases regardless how many causal SNPs are allowed in the model. However, the pattern is not consistent across the maximum number of causal SNPs allowed in the model. In cases where $R = 10$ and $R = 49$, as the maximum number of causal SNP increases from one to two, the number of noteworthy SNPs increases but not in cases where $R = 4$ and $R = 19$ where the number remains the same. Allowing three, four and five causal SNPs in the model with $R = 10$, 19 and 49 has captured a smaller number of noteworthy SNPs compared to having two causal SNPs in the model. In cases where R equals 4 or 10, the number of noteworthy SNPs reduce by one as we increase the maximum number of causal SNPs from four to five. However, the noteworthy SNPs remain the same in cases where R is 19 or 49.

It is interesting to see the results of the noteworthiness of SNPs using Laplace prior. Our first interest turns to observing SNP 5 and SNP 9 being picked up as noteworthy in 90% of the scenarios. We suspect that these two SNPs are in high LD with either one of the causal SNP specified in the dataset. However, that is not true. Hence, we observe the Bayes factor for SNP 5 and SNP 9 in all cases of the maximum number of causal SNPs allowed in the model. This is because, the Bayes factor affect the decision of declaring a SNP is noteworthy as discussed in Section 4.3.2. In almost all cases, the Bayes factor for SNP 5 and SNP 9 exceeds the Bayes factor threshold in deciding if SNPs are noteworthy.

As mentioned before, as R increases, more SNPs are declared to noteworthy as shown in Table 7.5. From Equation 2.19, increasing the ratio of cost (R) would increase the threshold and hence more SNPs are declare to be noteworthy. The value of R depends on the cost of false non-discovery C_ω and cost of a false discovery C_η specified. According to Wakefield (2008), it is considered to be more costly to not identify a causal SNP as noteworthy compared to identifying a non-causal SNP as noteworthy. Hence, it is better to have the cost of false non-discovery to be larger than the cost of a false discovery.

Chapter 8

Application to Breast Cancer Consortium Data

8.1 iCOGs data

In Chapters 4 and 7, we compared our multi-SNP Laplace prior method with other fine mapping methods by applying them to the simulated data from HAPGEN (Su et al., 2011). A real dataset is available from the Collaborative Oncological Gene-environment Study (COGS) for use in fine mapping. COGS develop a genotyping array named iCOGs focusing on a large number of target SNPs highly associated with breast, prostate and ovarian cancer (Michailidou et al., 2013). The iCOGs data available for our thesis concentrate on breast cancer, in which the Breast Cancer Association Consortium (BCAC) selected a region in Chromosome two with base position between 201500074 and 202569992. The selected region contains the CASP8 gene, a gene that could affect susceptibility to cancer.

BCAC selected 585 SNPS to be genotyped. However, after quality control checks, only 501 SNPs were selected. Due to missing genotype, another 1232 SNPs were imputed using IMPUTE2 (Marchini and Howie, 2010) adding to the 501 SNPs genotyped. The data consists of 46450 cases and 42500 controls which contribute to a total sample size of 89050. To reduce the computational time, we further

reduce the number of SNPs in the data according to the marginal Wakefield Bayes factor. The same approach used in Section 7.1 is applied by specifying a Bayes factor of 3 as a threshold (Kass and Raftery, 1995). From 1733 SNPs, we managed to reduce the SNPs to 120.

8.2 Comparing methods by ranking SNPs using iCOGs data

In Chapter 7, the Laplace prior was compared to the Gaussian prior and FINEMAP via posterior probabilities of causal asscoiation for SNPs in the simulated data. The SNPs were ranked based on their posterior probabilities and the ranking performance of all methods were compared using ROC curves. In this Chapter, we are interested in evaluating the performance of the Laplace prior versus the p-value, the Gaussian prior and FINEMAP when applied to the iCOGs data. We include p-value simply out of interest to see the difference between frequentist and Bayesian approaches. The p-value comes from a univariate logistic regression with a Wald test. This can be examined by ranking the SNPs based on the posterior probabilities using Laplace prior and comparing the ranks of the top ranked Laplace prior SNPs in other methods. Our analysis is based on the 30 top ranked SNPs, using the Laplace prior.

Table 8.2 shows the top 30 SNPs ranked using the posterior probabilities computed using the Laplace prior. In this table, we could observe the ranks of the top 30 SNPs based on the p-value, the Gaussian prior and FINEMAP. To compute posterior probabilities using the Laplace prior, the Gaussian prior and FINEMAP, the maximum number of causal SNP allowed in the model is one initially. The same information is presented in Table 8.3, 8.4, 8.5 in which we change the maximum number of causal SNPs to two, three and four respectively. In Chapter 7, we concluded that allowing five causal SNPs resulted in similar posterior probabilities as allowing four causal SNP. Thus, in analysis the iCOGs data, we compare the methods only allowing up to four causal SNPs in the model. Allow five causal SNPs also takes a long time to run.

When allowing one causal SNP in the model, based on Table 8.2, SNP 31 is ranked first with a

posterior probability of 0.4827, with about 50% chance of being a causal SNP. We can observe that the top 10 SNPs ranked by the Laplace prior are consistently ranked in the top 10 across all methods. Other SNPs are ranked among the top 50 in other methods. Most of the SNPs selected in the top 30 are common SNPs with only four rare SNPs (ranked first, 18th, 20th and 29th). SNPs ranked 8th onwards have very small posterior probabilities (less than 1%) which indicates that these SNPs are unlikely to be causal. We further evaluate the performance of the Laplace prior when allowing more SNPs in the model.

In Table 8.3, SNP 31, SNP2, SNP 1, SNP 3, SNP 16 and SNP 24 are ranked the top 6 SNPs when allowing a maximum of two causal SNPs in the model. However, SNP 31 although still ranked first, the posterior probability is smaller than when allowing one causal SNP in the model. The same rare SNPs (SNP 31, SNP 602, SNP 1639 and SNP 681) are selected in the top 30 SNPs in this scenario. Some of the top 30 SNPs using the Laplace prior are ranked very low in other method. As an example, SNP 602 ranked 10th and 11th in three methods, is ranked 100th for the Gaussian prior. There is more variation in the SNP ranks across methods with a maximum of two causal SNPs compared to one.

As we increase the maximum number of causal SNPs in the model to three, the posterior probabilities of all the SNPs using the Laplace prior increase compared to allowing two causal SNPs in the model. SNP 31 which still ranked as first in all methods, has a posterior probability of 0.4728. Other SNPs have posterior probability less than 16%. More rare SNPs in the table are being selected in the top 30 SNPs ranked by the Laplace prior. All the rare SNPs are ranked above 15 except for SNP 1656 which is ranked 27th. Generally, these rare SNPs have lower rank in other methods. The same observation is true for other more common top ranked SNPs.

As expected, SNP 31 ranked first when allowing four causal SNPs in the model and is consistent in other methods. SNPs 2, SNP 1 and SNP 3 remained ranked as second, third and fourth respectively using the Laplace prior. In terms of posterior probabilities, the top 30 SNPs have higher posterior probabilities (at least 2% higher) compared to the top 30 SNPs when allowing three causal SNPs. In this scenario, there are three more rare SNPs selected in the top 30 using the Laplace prior adding to

6 rare SNPs being selected in scenarios when we allow one, two and three causal SNPs in the model. The additional three rare SNPs (SNP 811, SNP 816 and SNP 812) are already among the top 30 SNPs in other methods. There are few SNPs which are ranked very low using other methods but which are selected to be in the top 30 ranked SNPs using the Laplace prior. These SNPs are SNP 342, SNP 251, SNP 256 and SNP 244.

There are 19 SNPs consistently ranked as the top 30 in all scenarios. These SNPs are SNP 1, SNP 2, SNP 3, SNP 7, SNP 16, SNP 24, SNP 31, SNP 602, SNP 1022, SNP 1056, SNP 1062, SNP 1067, SNP 1069, SNP 1087, SNP 1088, SNP 1090, SNP 1091 and SNP 1096, SNP 1639. In all scenarios, SNPs 31, 2, 1 and 3 are consistently being in the top 4 SNPs. The top 4 SNPs have consistent posterior probability across all scenarios. SNP 31 has a posterior probability of around 0.5. SNP 2, SNP 1 and SNP 3 have posterior probabilities between 0.13 and 0.16. Among the top 4 SNPs, the only rare SNP is SNP 31.

A Spearman's correlation was run to determine the strength of the relationship between the Laplace prior and the other three methods. Table 8.1 shows the Spearman's correlation coefficient for p-value, the Gaussian prior and FINEMAP in relation to the Laplace prior. We consider running the Spearman's correlation of every method in each maximum number of causal SNPs allowed (row). When a maximum of one causal SNP is allowed in the model, all methods show a very strong correlation with the Laplace prior. However, when we allow a maximum of two causal SNPs in the model, the Spearman's correlations decrease showing a moderate correlation between the Laplace prior and all three methods. The relationship continues to have a moderate relationship when we increase the maximum number of causal SNP. All three methods (the Gaussian prior and FINEMAP) show a positive correlation meaning the higher the SNPs are ranked in Laplace prior, the higher the SNPs are rank in p-value, the Gaussian prior and FINEMAP.

***Table 8.1*:** The Spearman's correlation between Laplace prior and the other three methods (p-value, the Gaussian prior and FINEMAP) in all cases of maximum number of causal SNPs allowed in the model.

		Methods		
		p-value	**Gaussian prior**	**FINEMAP**
Maximum number of causal SNPs allowed in the model	**1**	0.9428	0.9745	0.9428
	2	0.9428	0.6236	0.4105
	3	0.9428	0.6292	0.5356
	4	0.9428	0.5526	0.5761

8.3 Comparing the noteworthiness of SNPs using the Laplace prior in iCOGs data

In this section, we examine the noteworthiness of the SNPs in the iCOGs data. This can be achieved by the decision made using Equation (2.19). To calculate the posterior probability of H_0, Equation (2.20) requires us to specify the prior odds of H_0 and the Bayes factor (Equation (7.1)). We use four different values of R (the ratio of costs of making incorrect decision) to examine the noteworthiness of SNPs when we allow up to a maximum of four causal SNPs in the model. Table 8.6 presents the SNPs identified as noteworthy when varying the values of R and the maximum number of causal SNPs allowed in the model.

Initially, specifying $R = 4$ results in no noteworthy SNPs when we allow one and two causal SNPs in the model. However, as the maximum number of causal SNPs increase to three and four, SNP 31 is picked up to be noteworthy. When specifying R equals to 10, SNP 31 appears to be the only noteworthy SNP in all scenario. As we increase the value of R to 19, SNP 2 appears to be noteworthy alongside SNP 31 as we allow four causal SNPs in the model. More SNPs are captured as noteworthy in the case where $R = 49$. We noticed that, all noteworthy SNPs are among the 19 SNPs consistently ranked as top 30 in Section 8.2 which also includes the top 4 SNPs. Out of seven noteworthy SNPs, three of them are rare SNPs (SNP 31, SNP 602 and SNP 1639).

***Table 8.2*:** The 30 top ranked SNPs in the iCOGs data based on posterior probabilities using the Laplace prior. The ranking based on the p-value, the posterior probabilities using the Gaussian prior and FINEMAP are also included. The posterior probabilities calculated for all Bayesian methods allow one causal SNP in the model.

			Ranking			
SNPs	**MAF**	**Posterior probabilities (Laplace prior)**	**Laplace prior**	**p-value**	**Gaussian prior**	**FINEMAP**
31	0.081	4.83e-1	1	1	1	1
2	0.124	1.38e-1	2	2	2	2
1	0.124	1.27e-1	3	3	3	3
3	0.123	1.25e-1	4	4	4	4
16	0.124	6.07e-2	5	5	5	5
24	0.122	2.81e-2	6	6	6	6
7	0.121	1.42e-2	7	7	7	7
29	0.171	8.61e-3	8	8	8	8
27	0.170	5.69e-3	9	9	9	9
10	0.258	5.80e-4	10	10	10	10
23	0.257	5.02e-4	11	12	11	14
14	0.258	5.02e-4	12	13	12	13
6	0.257	4.91e-4	13	14	13	12
9	0.257	4.70e-4	14	15	14	16
8	0.257	4.68e-4	15	16	15	15
15	0.254	3.74e-4	16	17	16	17
4	0.256	3.01e-4	17	18	18	18
602	0.074	2.14e-4	18	11	17	11
1096	0.486	1.46e-4	19	22	20	22
1639	0.087	1.34e-4	20	19	19	19
1056	0.426	1.31e-4	21	23	21	26
1087	0.446	1.22e-4	22	27	23	29
1022	0.495	8.54e-5	23	37	30	37
1062	0.498	8.50e-5	24	38	31	38
1069	0.498	8.34e-5	25	39	32	41
1090	0.497	8.18e-5	26	40	33	43
1067	0.484	8.11e-5	27	42	35	40
1088	0.497	8.04e-5	28	41	36	42
681	0.071	7.94e-5	29	20	22	20
1091	0.497	7.82e-5	30	43	39	44

***Table 8.3*:** The 30 top ranked SNPs in the iCOGs data based on posterior probabilities using the Laplace prior. The ranking based on the p-value, the posterior probabilities using the Gaussian prior and FINEMAP are also included. The posterior probabilities calculated for all Bayesian methods allow two causal SNP in the model.

			Ranking			
SNPs	**MAF**	**Posterior probabilities (Laplace prior)**	**Laplace prior**	**p-value**	**Gaussian prior**	**FINEMAP**
31	0.081	0.45657	1	1	1	1
2	0.124	0.14908	2	2	3	2
1	0.124	0.13782	3	3	4	3
3	0.123	0.13406	4	4	5	4
16	0.124	0.06838	5	5	6	5
24	0.122	0.03267	6	6	21	6
1639	0.087	0.03170	7	19	2	19
1087	0.446	0.02444	8	27	22	30
1056	0.426	0.02443	9	23	23	33
602	0.074	0.02342	10	11	100	10
1096	0.486	0.02072	11	22	25	22
7	0.121	0.01814	12	7	26	7
1022	0.495	0.01794	13	37	27	37
1067	0.484	0.01697	14	42	28	54
1062	0.498	0.01648	15	38	30	40
1069	0.498	0.01615	16	39	31	43
1090	0.497	0.01602	17	40	32	45
1014	0.488	0.01587	18	45	34	57
1088	0.497	0.01572	19	41	35	44
1012	0.488	0.01556	20	47	36	60
1091	0.497	0.01531	21	43	38	46
1078	0.482	0.01494	22	50	39	61
1021	0.495	0.01482	23	48	40	42
1010	0.491	0.01469	24	49	37	32
1055	0.499	0.01406	25	46	41	49
1020	0.495	0.01372	26	53	42	47
1007	0.497	0.01318	27	51	43	52
1047	0.499	0.01204	28	54	44	53
1030	0.493	0.01177	29	56	45	59
1035	0.409	0.01174	30	58	107	85

***Table 8.4*:** The 30 top ranked SNPs in the iCOGs data based on posterior probabilities using the Laplace prior. The ranking based on the p-value, the posterior probabilities using the Gaussian prior and FINEMAP are also included. The posterior probabilities calculated for all Bayesian methods allow three causal SNP in the model.

			Ranking			
SNPs	**MAF**	**Posterior probabilities (Laplace prior)**	**Laplace prior**	**p-value**	**Gaussian prior**	**FINEMAP**
31	0.081	0.47283	1	1	1	1
2	0.124	0.15388	2	2	3	3
1	0.124	0.14274	3	3	4	4
3	0.123	0.13777	4	4	6	2
16	0.124	0.07323	5	5	16	5
1639	0.087	0.07133	6	19	2	16
602	0.074	0.06019	7	11	29	10
24	0.122	0.03684	8	6	26	6
1056	0.426	0.03399	9	23	5	31
1087	0.446	0.03329	10	27	25	32
342	0.452	0.03255	11	100	82	70
1671	0.063	0.02972	12	21	65	19
1096	0.486	0.02926	13	22	30	22
681	0.071	0.02702	14	20	66	21
251	0.205	0.02477	15	80	101	56
7	0.121	0.02283	16	7	35	7
1022	0.495	0.02152	17	37	43	40
1062	0.498	0.02132	18	38	38	43
1069	0.498	0.02092	19	39	41	44
1067	0.484	0.02068	20	42	45	54
1090	0.497	0.02067	21	40	39	47
1088	0.497	0.02031	22	41	42	39
1091	0.497	0.01982	23	43	44	49
1014	0.488	0.01937	24	45	49	57
256	0.203	0.01925	25	95	104	66
1012	0.488	0.01901	26	47	50	62
1656	0.087	0.01895	27	52	12	24
244	0.221	0.01878	28	96	105	65
1010	0.491	0.01870	29	49	48	30
1078	0.482	0.01840	30	50	52	55

***Table 8.5*:** The 30 top ranked SNPs in the iCOGs data based on posterior probabilities using the Laplace prior. The ranking based on the p-value, the posterior probabilities using the Gaussian prior and FINEMAP are also included. The posterior probabilities calculated for all Bayesian methods allow four causal SNP in the model.

			Ranking			
SNPs	**MAF**	**Posterior probabilities (Laplace prior)**	**Laplace prior**	**p-value**	**Gaussian prior**	**FINEMAP**
31	0.081	0.49168	1	1	1	1
2	0.124	0.15707	2	2	4	3
1	0.124	0.14612	3	3	5	4
3	0.123	0.14064	4	4	9	2
1639	0.087	0.10466	5	19	3	15
602	0.074	0.07903	6	11	30	10
16	0.124	0.07716	7	5	20	5
342	0.452	0.05535	8	100	33	68
1671	0.063	0.05129	9	21	35	17
24	0.122	0.04045	10	6	29	6
1056	0.426	0.03910	11	23	2	25
1087	0.446	0.03797	12	27	27	31
251	0.205	0.03709	13	80	55	59
681	0.071	0.03689	14	20	44	21
1096	0.486	0.03475	15	22	40	22
256	0.203	0.02891	16	95	75	71
1656	0.087	0.02883	17	52	16	28
244	0.221	0.02846	18	96	80	73
7	0.121	0.02703	19	7	38	7
1062	0.498	0.02388	20	38	50	34
1022	0.495	0.02349	21	37	47	40
1069	0.498	0.02345	22	39	54	39
1090	0.497	0.02315	23	40	45	41
1088	0.497	0.02278	24	41	48	36
1067	0.484	0.02276	25	42	67	52
1091	0.497	0.02227	26	43	53	53
811	0.078	0.02199	27	24	13	23
816	0.078	0.02177	28	25	14	32
812	0.081	0.02160	29	26	23	27
1014	0.488	0.02138	30	45	72	55

***Table 8.6*:** The noteworthy SNPs at different values of the maximum number of causal SNPs and ratios of costs of making incorrect decisions corresponds to the Bayesian false-discovery probabilities (BFDP).

Ratio of costs, R	**4**	**10**	**19**	**49**
BFDP	**0.8**	**0.91**	**0.95**	**0.98**
1 causal	na	{31}	{31}	{1,2,3,31}
2 causals	na	{31}	{31}	{1,2,3,31}
3 causals	{31}	{31}	{31}	{1,2,3,16,31,1639}
4 causals	{31}	{31}	{2,31}	{1,2,3,16,31,602,1639}

8.4 Breast cancer risk association at CASP8 region.

The iCOGs data used in this chapter focuses on breast cancer by selecting a region in Chromosome 2 containing CASP8 gene as explained in Section 8.1. CASP8 codes for caspase protein that is involve in apoptosis, a biological process which programmed cell death. As our cell constantly replicates, some cells die or need to be 'deleted' during development to maintain balance in our body. In some cases, there are cell that can harm our body such as infected and cancerous cell. Thus, if apoptosis does not occur, cell continues dividing and could lead to cancer. This could happen because of a variant in the CASP8 gene that causes the caspase protein to not function correctly (Elmore, 2007).

One of the first true variant in CASP8 identified by candidate gene study to be associated with breast cancer is D302H (rs1045485) (Cox et al., 2007). D302H (rs1045485) shows a highly significant association of the minor allele and has a 10% decrease in risk. In a further fine-mapping studies, there was evidence showing that D302H (rs1045485) has a weak association ($P_{trend} = 0.046$) with breast cancer (Shephard et al., 2009; Michailidou et al., 2013). Moreover, these studies show another 3 variants (rs3834126, rs6435074, rs6723097) in CASP8 region having significant association with breast cancer. An independent variant in the same region, rs10931936, was also found to be associated but is in low LD ($r^2 = 0.083$) with D302H (rs1045485) (Turnbull et al., 2010). In 2015, Lin et al. (2015) analyzed the same data we used in Section 8.2 by using a meta-analysis of iCOGs together with nine GWAS breast cancer data to clarify the role of CASP8 in the risk of having breast cancer. There is a significant association for an imputed SNP rs1830298 in ALS2CR1 which is telomeric to

CASP8 and the genotyped SNP rs10197246 in CASP8. These two SNPs are in high LD ($r^2 = 0.9$) and they are likely to have the same association signal.

Based on our analysis in Section 8.2, the results show that SNP 31 has been ranked first consistently in all the methods that we used to compare. SNP 31 also appeared to be noteworthy even when we changed the ratio of cost, R values. Apparently, to the best of our knowledge, SNP 31 (rs2540050) is not known to be causal or to be in the list of potential causal SNPs. However, another Bayesian approach which incorporates functional genomic information had also highly ranked SNP 31 (rs2540050) (Alenazi et al., 2019).

Chapter 9

Discussion

9.1 Focus of the research

The essential elements in a Bayesian approach are the likelihood distribution and the prior distribution. Using these two elements, a posterior probability distribution of a parameter can be computed to update our belief about the parameter. The computation of the posterior distributions depends on the choice of prior distribution specified. To specify a sensible prior distribution can be challenging. The commonly used prior distribution in fine mapping is Gaussian. In most of the fine mapping literature, the mean in the Gaussian prior takes the value 0. However, there are many ways of defining the value of the variance: by elicitation, specifying a fixed value or specifying a prior on the hyperparameter.

In our thesis, we developed a Bayesian approach that computes the posterior distribution and the Bayes factor which leads to identifying disease-specific SNPs. The choice we made about the prior distribution of the effect sizes is based on the GWAS top hits data. A sensible choice of distribution that reflects the GWAS top hits data is the Laplace distribution. The Laplace distribution gives a tractable integral when computing the posterior summaries and the Bayes factor. Similar to the Gaussian distribution used in previous fine mapping studies, we need to define value for the parameters. Since we want the distribution to symmetric around zero, we define the location parameter to be zero.

The value of the scale parameter λ is estimated using Maximum Likelihood Estimation (MLE). According to Michailidou et al. (2013), there are 1168 SNPs unidentified SNPS with very small odds ratio. Thus, we take two approaches of estimating the λ. The first approach is by not considering how many unidentified SNPs are there in the GWAS top hits data. The second approach is by varying the number of yet-to-be-discovered SNPs (YTBD) and using those numbers in the calculation of the MLE. We also use the same approach to estimate the variance parameter in a Gaussian prior.

Using the MLE as the hyperparameter for both the Gaussian prior and the Laplace prior, we showed that the Laplace prior has a better fit to the GWAS top hits data. This is true either by considering the number of YTBD SNPs or not. We then calculated Bayesian posterior summaries using a Laplace prior and compared the performance with other fine mapping methods by applying it to simulated data from HAPGEN and also on real data from iCOGs.

Both single-SNP and multi-SNP analyses have shown that the Laplace prior method had better performance to rank SNPs compared to other existing fine mapping methods. In single-SNP analysis, the results appeared to be sensitive to the minor allele frequency, odds ratio, sample size and the number of yet-to-be-discovered (YTBD) SNPs. We proposed to specify a prior on the number of YTBD SNPs that allows for the uncertainty in the estimated λ. This leads to deriving a Bayes factor we called the Laplace Gamma Bayes factor. The results in multi-SNP analysis showed that increasing the maximum number of causal SNPs in the model, increases the posterior probabilities of causal association. Thus, the Laplace prior is sensitive to the maximum number of causal SNPs allowed in the model.

9.2 Limitations

In this thesis, we choose to compare the Laplace prior with FINEMAP. Based on the ability of FINEMAP to allow a maximum of five causal SNPs in the model, we also allow a maximum of 5 causal SNPs in our method. However, without prior knowledge of how many causal SNPs there are in

the region, questions can be raised with regards to the maximum number of causal SNPs in the model. Most methods in fine mapping assumed there are at least one causal SNP in the region. Instead of using this assumption, we take into account that there might be no causal SNP in the region, thus allowing our method to deal with models with zero causal SNPs. Nevertheless, users may want to know what maximum number of causal SNPs to specify or even if they specify a certain number, how could they tell that there might be more causal SNPs in the region than the ones specified. Our methods simply look at the ranking performance and how much the posterior probabilities change as we vary the maximum number of causal SNPs in the model. It would be better to have a way to incorporate this into our method to make it clearer for the user to specify the maximum number of causal SNPs.

In our multi-SNP analysis, our analysis and discussion were based on the posterior probability of causal association of individual SNPs. The posterior probability describes our uncertainty about the SNP having an association with breast cancer risk. However, a major interest in fine-mapping is to look into the correct multi-SNP model. A SNP with high posterior probability does not mean that the SNP is causal. It could be that it tags another SNP in the data which potentially could be the causal SNP. Thus, by analysing the correct multi-SNP model, this could help in identifying the actual causal SNP in the region.

Another drawback in our method is the computational time. Table 9.1 shows the computational time for Laplace prior, Gaussian prior and FINEMAP using the simulation data in Section 7.1 with 50 SNPs. We could see that as we increase the maximum number of SNPs allowed in model, the more time is takes to compute for every approach. The reason is, if we have p SNPs and allow k causal SNPs in the model, we have ${}^{p}C_{k}$ combinations of models. Generally, the more p we have and the more k we allow, the more time it takes to compute. From Table 9.1, the Laplace prior took more time to compute compared to the Gaussian prior and FINEMAP. In the Laplace prior, each combination of model has more combinations which depends on the vector $\boldsymbol{A}$ defined in Equation (6.24). Thus, more time is needed to compute the Laplace prior. We already reduce the computational time by reducing the number of SNPs in the data based on the marginal Bayes factor. However, there are other

approaches to explore in order to have our method run faster.

***Table 9.1*:** The computational times (in minutes) for Laplace prior, Gaussian prior and FINEMAP using the simulation data from HAPGEN2.

		Methods		
		Laplace Prior	**Gaussian Prior**	**FINEMAP**
Maximum number of causal SNPS allowed in the model	**1**	0.45	0.22	0.03
	2	2	0.57	0.04
	3	26	1.11	0.25
	4	156	7.7	0.45
	5	918	69.6	0.55

The SNPs in the region are sometimes highly correlated and these SNPs give similar signals which makes it difficult to pinpoint the causal SNPs. Before we analyse the genotype data using our method, we filter out all the highly correlated SNPs ($r^2 \geq 0.99$) from the data. This may result in removing the potential causal SNP. Our method identifies plausible causal SNPs to prioritize from the posterior probabilities and Bayes factors. This set of plausible causal SNPs is not selected using any biological knowledge. To identify the true causal SNPs, the plausible causal SNPs can be further tested in functional studies.

9.3 Future work

Research in genetic epidemiology is continuously growing with a lot of different approaches currently being introduced. It would be interesting to further modify our method with some improvement from the limitations we identified. The Laplace distribution has shown to be a better fit for the prior on the effect sizes compared to Normal distribution. We based our analysis using the breast cancer GWAS top hits data. However, this can be further used in other complex diseases since now some complex diseases have a significant number of GWAS top hits available. Our approach brings the idea to use the information from the GWAS top hits data in specifying more objective priors on the effect size.

Besides reducing the number of SNPs in the data based on the marginal Bayes factor to make our

method run faster, we could modify our approach using sequential method. This method allows us to remove some SNPs after considering all the possible pairs of causal SNPs. Another approach that can be considered is to implement a Shotgun Stochastic Search (SSS) algorithm which has been used in FINEMAP and proven to run very quickly.

One way to tackle the signal of highly correlated SNPs in the data is by integrating functional genomic information into our method. This could also be an alternative method to aid in assessing the potential disease causality of the prioritized SNPs. An advantage of using a Bayesian approach is that we could extend our method by allowing the effect size prior to depend on the functional information or to allow the prior probability of causal association to depend on functional information. With the extensive functional information data available, we could readily use this data to inform our priors. Incorporating the genomic information into our method could help with identifying a causal SNP among a correlated groups of SNPs.

References

Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R., Morrison, J., Maranian, M., Pooley, K. A., Luben, R. et al. (2009), 'Newly discovered breast cancer susceptibility loci on 3p24 and 17q23. 2', *Nature genetics* **41**(5), 585.

Alenazi, A. A., Cox, A., Juarez, M., Lin, W.-Y. and Walters, K. (2019), 'Bayesian variable selection using partially observed categorical prior information in fine-mapping association studies', *Genetic epidemiology* **43**(6), 690–703.

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S. and Pirinen, M. (2016), 'FINEMAP: Efficient variable selection using summary data from genome-wide association studies', *Bioinformatics* **32**(10), 1493–1501.

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2012), 'Bayesian shrinkage', *arXiv preprint arXiv:1212.6088* .

Bush, W. S. and Moore, J. H. (2012), 'Genome-wide association studies', *PLoS computational biology* **8**(12), e1002822.

Cai, Q., Zhang, B., Sung, H., Low, S.-K., Kweon, S.-S., Lu, W., Shi, J., Long, J., Wen, W., Choi, J.-Y. et al. (2014), 'Genome-wide association analysis in east asians identifies breast cancer susceptibility loci at 1q32. 1, 5q14. 3 and 15q26. 1', *Nature genetics* **46**(8), 886.

Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A. and Schaid, D. J. (2015), 'Fine mapping causal variants with an approximate bayesian method using marginal test statistics', *Genetics* **200**(3), 719–736.

Cox, A., Dunning, A. M., Garcia-Closas, M., Balasubramanian, S., Reed, M. W., Pooley, K. A., Scollen, S., Baynes, C., Ponder, B. A., Chanock, S. et al. (2007), 'A common coding variant in casp8 is associated with breast cancer risk', *Nature genetics* **39**(3), 352–358.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R. et al. (2007), 'Genome-wide association study identifies novel breast cancer susceptibility loci', *Nature* **447**(7148), 1087–1093.

Elmore, S. (2007), 'Apoptosis: a review of programmed cell death', *Toxicologic pathology* **35**(4), 495–516.

Fachal, L. and Dunning, A. M. (2015), 'From candidate gene studies to gwas and post-gwas analyses in breast cancer', *Current Opinion in Genetics and Development* **30**, 32–41.

Fawcett, T. (2006), 'An introduction to ROC analysis', *Pattern Recognition Letters* **27**(8), 861–874.

Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., Zelenika, D., Gut, I., Heath, S., Palles, C. et al. (2011), 'Novel breast cancer susceptibility locus at 9q31. 2: results of a genome-wide association study', *Journal of the National Cancer Institute* **103**(5), 425–435.

Forte, A., Garcia-Donato, G. and Steel, M. (2018), 'Methods and tools for bayesian variable selection and model averaging in normal linear regression', *International Statistical Review* **86**(2), 237–258.

George, E. I. and McCulloch, R. E. (1993), 'Variable selection via Gibbs sampling', *Journal of the American Statistical Association* **88**(423), 881–889.

Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., Dennis, J., Wang, Q., Humphreys, M. K., Luccarini, C. et al. (2012), 'Genome-wide association analysis identifies three new breast cancer susceptibility loci', *Nature genetics* **44**(3), 312.

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

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

Hong, E. P. and Park, J. W. (2012), 'Sample Size and Statistical Power Calculation in Genetic Association Studies', *Genomics & Informatics* **10**(2), 117.

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. and Eskin, E. (2014), Identifying causal variants at loci with multiple signals of association, *in* 'ACM BCB 2014 - 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics', pp. 610–611.

Kass, R. E. and Raftery, A. E. (1995), 'Bayes factors', *Journal of the american statistical association* **90**(430), 773–795.

Kichaev, G., Yang, W. Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P. and Pasaniuc, B. (2014), 'Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies', *PLoS Genetics* **10**(10), e1004722.

Kulldorff, G. (1961), *Contributions to the theory of estimation from grouped and partially grouped samples*, Almqvist & Wiksell.

Kuo, L. and Mallick, B. (1998), 'Variable selection for regression models', *Sankhyā: The Indian Journal of Statistics, Series B* pp. 65–81.

Lin, W.-Y., Camp, N. J., Ghoussaini, M., Beesley, J., Michailidou, K., Hopper, J. L., Apicella, C., Southey, M. C., Stone, J., Schmidt, M. K. et al. (2015), 'Identification and characterization of novel associations in the casp8/als2cr12 region on chromosome 2 with breast cancer risk', *Human molecular genetics* **24**(1), 285–298.

Lu, T.-T. and Shiou, S.-H. (2002), 'Inverses of $2\times$ 2 block matrices', *Computers & Mathematics with Applications* **43**(1-2), 119–129.

Marchini, J. and Howie, B. (2010), 'Genotype imputation for genome-wide association studies', *Nature Reviews Genetics* **11**(7), 499.

Meiosis and Formation of Eggs and Sperm (2000 (accessed September 25, 2019)).
URL: *https://www.biology.iupui.edu/biocourses/N100H/ch9meiosis.html*

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M. et al. (2015), 'Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer', *Nature genetics* **47**(4), 373.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K. and et al. (2013), 'Large-scale genotyping identifies 41 new loci associated with breast cancer risk', *Nature Genetics* **45**(4), 353–361.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A. et al. (2017), 'Association analysis identifies 65 new breast cancer risk loci', *Nature* **551**(7678), 92.

Milne, R. L., Burwinkel, B., Michailidou, K., Arias-Perez, J.-I., Zamora, M. P., Menéndez-Rodríguez, P., Hardisson, D., Mendiola, M., González-Neira, A., Pita, G. et al. (2014), 'Common non-

synonymous snps associated with breast cancer susceptibility: findings from the breast cancer association consortium', *Human molecular genetics* **23**(22), 6096–6111.

Mitchell, T. J. and Beauchamp, J. J. (1988), 'Bayesian variable selection in linear regression', *Journal of the American Statistical Association* **83**(404), 1023–1032.

Mitosis Compared With Meiosis (2009 (accessed September 25, 2019)).
URL: *https://www2.palomar.edu/users/warmstrong/lmexer2a.htm*

Molnar, C. and Gair, J. ((accessed September 25, 2019)), *Concepts of Biology – 1st Canadian Edition*.
URL: *https://opentextbc.ca/biology/chapter/9-4-translation/*

O'Hara, R. B. and Sillanpää, M. J. (2009), 'A review of bayesian variable selection methods: what, how and which', *Bayesian Analysis* **4**(1), 85–118.

OpenStax ((accessed September 25, 2019)), *Microbiology*.
URL: *https://courses.lumenlearning.com/microbiology/chapter/structure-and-function-of-dna/*

Park, T. and Casella, G. (2008), 'The Bayesian Lasso', *Journal of the American Statistical Association* **103**(482), 681–686.

Schaid, D. J., Chen, W. and Larson, N. B. (2018), 'From genome-wide associations to candidate causal variants by statistical fine-mapping'.

Servin, B. and Stephens, M. (2007), 'Imputation-based analysis of association studies: candidate regions and quantitative traits', *PLoS genetics* **3**(7), e114.

Shephard, N. D., Abo, R., Rigas, S. H., Frank, B., Lin, W.-Y., Brock, I. W., Shippen, A., Balasubramanian, S. P., Reed, M. W. R., Bartram, C. R. et al. (2009), 'A breast cancer risk haplotype in the caspase-8 gene', *Cancer research* **69**(7), 2724–2728.

Siddiq, A., Couch, F. J., Chen, G. K., Lindström, S., Eccles, D., Millikan, R. C., Michailidou, K., Stram, D. O., Beckmann, L., Rhie, S. K. et al. (2012), 'A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11', *Human molecular genetics* **21**(24), 5373–5384.

Spencer, A. V., Cox, A., Lin, W. Y., Easton, D. F., Michailidou, K. and Walters, K. (2015), 'Novel bayes factors that capture expert uncertainty in prior density specification in genetic association studies', *Genetic Epidemiology* **39**(4), 239–248.

Spencer, A. V., Cox, A., Lin, W. Y., Easton, D. F., Michailidou, K. and Walters, K. (2016), 'Incorporating Functional Genomic Information in Genetic Association Studies Using an Empirical Bayes Approach', *Genetic Epidemiology* **40**(3), 176–187.

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

Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A. et al. (2007), 'Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor–positive breast cancer', *Nature genetics* **39**(7), 865–869.

Stacey, S. N., Manolescu, A., Sulem, P., Thorlacius, S., Gudjonsson, S. A., Jonsson, G. F., Jakobsdottir, M., Bergthorsson, J. T., Gudmundsson, J., Aben, K. K. et al. (2008), 'Common variants on chromosome 5p12 confer susceptibility to estrogen receptor–positive breast cancer', *Nature genetics* **40**(6), 703–706.

Strachan, T. (2011), *Human molecular genetics*, 4th ed. edn, Garland Science, New York ; Abingdon.

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

Sudbery, P. (2009), *Human molecular genetics*, Cell and molecular biology in action, third edition. edn, Pearson/Benjamin Cummings, Harlow.

Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., Hankinson, S. E., Hutchinson, A., Wang, Z., Yu, K. et al. (2009), 'A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11. 2 and 14q24. 1 (rad5111)', *Nature genetics* **41**(5), 579.

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C. S. et al. (2010), 'Genome-wide association study identifies five new breast cancer susceptibility loci', *Nature genetics* **42**(6), 504.

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

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

Wakefield, J. (2009), 'Bayes factors for Genome-wide association studies: Comparison with P-values', *Genetic Epidemiology* **33**(1), 79–86.

Walters, K., Cox, A. and Yaacob, H. (2019), 'Using gwas top hits to inform priors in bayesian fine-mapping association studies', *Genetic epidemiology* **43**(6), 675–689.

Yaacob, H., Walters, K. and Cox, A. (2019), Utilizing the information from gwas data to inform priors in bayesian fine-mapping association studies, *in* 'Human Heredity', Vol. 83, Karger, pp. 247–247.

Zheng, W., Long, J., Gao, Y.-T., Li, C., Zheng, Y., Xiang, Y.-B., Wen, W., Levy, S., Deming, S. L., Haines, J. L. et al. (2009), 'Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25. 1', *Nature genetics* **41**(3), 324.

Appendix A

Justification for not including the intercept in the model

The exponential term from the pdf for the likelihood of the data which includes the intercept is given as

$$\frac{1}{2}\begin{bmatrix}\alpha-\hat{\alpha}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}^T \quad V^{-1} \quad \begin{bmatrix}\alpha-\hat{\alpha}\\ \boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\end{bmatrix}. \tag{A.1}$$

We let

$$V^{-1} = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{D}\\ \boldsymbol{D}^T & \boldsymbol{C}\end{bmatrix} \tag{A.2}$$

where $\boldsymbol{A}$ is the scalar variance for α, $\boldsymbol{D}$ is a $1 \times p$ variance covariance matrix and $\boldsymbol{C}$ is a $p \times p$ variance covariance matrix for $\boldsymbol{\beta}$. Thus, the likelihood is

$$l = -\frac{1}{2}\Big[(\alpha - \hat{\alpha})^2 \boldsymbol{A} + 2(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{D}^T (\alpha - \hat{\alpha}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \boldsymbol{C} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\Big] + constant \tag{A.3}$$

We begin by reparameterising the parameters to be

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} \tag{A.4}$$

and

$$\alpha^* = \alpha + \frac{\boldsymbol{D}}{\boldsymbol{A}}\boldsymbol{\beta}^*. \tag{A.5}$$

The new likelihood with the new parameterisation is

$$l^* = -\frac{1}{2}\left[\left(\alpha^* - \frac{\boldsymbol{D}}{\boldsymbol{A}}\boldsymbol{\beta}^* - \hat{\alpha}\right)^2 \boldsymbol{A} + 2(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})^T \boldsymbol{D}^T \left(\alpha^* - \frac{\boldsymbol{D}}{\boldsymbol{A}}\boldsymbol{\beta}^* - \hat{\alpha}\right) + (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})^T \boldsymbol{C} (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\right] + constant. \tag{A.6}$$

The likelihood is now re-written ignoring the first order terms. The reason we did not include the first order terms is because when we calculate the second derivatives, the first order terms will give zero. Thus, the new likelihood (ignoring the first order terms in α^* and $\boldsymbol{\beta}^*$) is

$$l^* = -\frac{1}{2}\left[\left(\alpha^* - \frac{\boldsymbol{D}}{\boldsymbol{A}}\boldsymbol{\beta}^*\right)^2 \boldsymbol{A} + 2(\boldsymbol{\beta}^*)^T \boldsymbol{D}^T \left(\alpha^* - \frac{\boldsymbol{D}}{\boldsymbol{A}}\boldsymbol{\beta}^*\right) + (\boldsymbol{\beta}^*)^T \boldsymbol{C} \boldsymbol{\beta}^*\right] + constant. \tag{A.7}$$

We first differentiate the new likelihood with respect to α^*. This gives

$$\frac{\partial l^*}{\partial \alpha^*} = -\frac{1}{2}\left[2\boldsymbol{A}\left(\alpha^* - \frac{\boldsymbol{D}}{\boldsymbol{A}}\boldsymbol{\beta}^*\right) + 2(\boldsymbol{\beta}^*)^T\boldsymbol{D}^T\right]. \tag{A.8}$$

Next, we take the second differentiate with respect to α^* which gives

$$\frac{\partial^2 l^*}{\partial \alpha^{*2}} = -\frac{1}{2}\left[2\boldsymbol{A}\right] = -\boldsymbol{A}. \tag{A.9}$$

The second differentiate with respect to α^* and $\boldsymbol{\beta}^*$ is

$$\frac{\partial l^*}{\partial \boldsymbol{\beta}^* \alpha^*} = -\frac{1}{2}\left[2\boldsymbol{A}\left(-\frac{\boldsymbol{D}}{\boldsymbol{A}}\mathbf{1}_p\right) + 2\,\mathbf{1}_p^T\boldsymbol{D}^T\right] = \mathbf{0}. \tag{A.10}$$

where $\mathbf{1}_p$ is $p \times 1$ vector of 1s.

To find the second derivative of the log-likelihood with respect to β^*, we simplify Equation (A.7) by only considering second order terms in $\boldsymbol{\beta}^*$. The new likelihood is

$$l^* = -\frac{1}{2}\left[\boldsymbol{A}\left(\frac{\boldsymbol{D}}{\boldsymbol{A}}\boldsymbol{\beta}^*\right)^2 + 2(\boldsymbol{\beta}^*)^T\boldsymbol{D}^T\left(-\frac{\boldsymbol{D}}{\boldsymbol{A}}\boldsymbol{\beta}^*\right) + (\boldsymbol{\beta}^*)^T\boldsymbol{C}\boldsymbol{\beta}^*\right] + constant. \tag{A.11}$$

Since $\boldsymbol{D}\boldsymbol{\beta}^* = (\boldsymbol{D}\boldsymbol{\beta}^*)^T = (\boldsymbol{\beta}^*)^T\boldsymbol{D}^T$, Equation (A.11) can be expressed as

$$\begin{aligned} l^* &= -\frac{1}{2}\left[\frac{1}{\boldsymbol{A}}(\boldsymbol{D}\boldsymbol{\beta}^*)(\boldsymbol{D}\boldsymbol{\beta}^*) + 2\boldsymbol{D}\boldsymbol{\beta}^*\left(-\frac{(\boldsymbol{\beta}^*)^T\boldsymbol{D}^T}{\boldsymbol{A}}\right) + (\boldsymbol{\beta}^*)^T\boldsymbol{C}\boldsymbol{\beta}^*\right] + constant \\ &= -\frac{1}{2}\left[\frac{1}{\boldsymbol{A}}\boldsymbol{D}\boldsymbol{\beta}^*(\boldsymbol{\beta}^*)^T\boldsymbol{D}^T - \frac{2}{\boldsymbol{A}}\boldsymbol{D}\boldsymbol{\beta}^*(\boldsymbol{\beta}^*)^T\boldsymbol{D}^T + (\boldsymbol{\beta}^*)^T\boldsymbol{C}\boldsymbol{\beta}^*\right] + constant \\ &= -\frac{1}{2}\left[(\boldsymbol{\beta}^*)^T\boldsymbol{C}\boldsymbol{\beta}^* - \frac{1}{\boldsymbol{A}}\boldsymbol{D}\boldsymbol{\beta}^*(\boldsymbol{\beta}^*)^T\boldsymbol{D}^T\right] + constant \end{aligned} \tag{A.12}$$

The first derivatives of Equation (A.12) with respect to $\boldsymbol{\beta}^*$ is

$$\begin{aligned}\frac{\partial l^*}{\partial \boldsymbol{\beta}^*} &= -\frac{1}{2}\left[(\boldsymbol{C}+\boldsymbol{C}^T)\boldsymbol{\beta}^* - \frac{1}{\boldsymbol{A}}(2\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\beta}^*)\right] \\ &= -\frac{1}{2}\left[2\boldsymbol{C}\boldsymbol{\beta}^* - \frac{2}{\boldsymbol{A}}(\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\beta}^*)\right] \\ &= -\left[\boldsymbol{C}\boldsymbol{\beta}^* - \frac{\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\beta}^*}{\boldsymbol{A}}\right] \\ &= \left[\frac{\boldsymbol{D}^T\boldsymbol{D}}{\boldsymbol{A}} - \boldsymbol{C}\right]\boldsymbol{\beta}^*. \end{aligned} \tag{A.13}$$

Following from Equation (A.13), the second derivatives is

$$\frac{\partial^2 l^*}{\delta \partial^{*2}} = \frac{\boldsymbol{D}^T\boldsymbol{D}}{\boldsymbol{A}} - \boldsymbol{C}. \tag{A.14}$$

Using all the derivatives calculated above gives a new Expected Information matrix as follows

$$\mathrm{I}_{new} = \begin{bmatrix} -E(\frac{\delta^2 l^*}{\delta \alpha^{*2}}) & -E(\frac{\delta^2 l^*}{\delta \alpha^* \beta^*}) \\ -E(\frac{\delta^2 l^*}{\delta \alpha^* \beta^*}) & -E(\frac{\delta^2 l^*}{\delta \beta^{*2}}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0}_p \\ \boldsymbol{0}_p^T & -[\frac{\boldsymbol{D}^T\boldsymbol{D}}{\boldsymbol{A}} - \boldsymbol{C}] \end{bmatrix}. \tag{A.15}$$

Thus, the new variance covariance matrix, V, is obtained as $V = I^{-1}$ which gives

$$V_{new} = \begin{bmatrix} \boldsymbol{A}^{-1} & \boldsymbol{0}_p \\ \boldsymbol{0}_p^T & -[\frac{\boldsymbol{D}^T\boldsymbol{D}}{\boldsymbol{A}} - \boldsymbol{C}]^{-1} \end{bmatrix}. \tag{A.16}$$

and the covariance for $\hat{\beta}$ under the new parameterisation is

$$Cov(\hat{\boldsymbol{\beta}}) = -\left[\frac{\boldsymbol{D}^T\boldsymbol{D}}{\boldsymbol{A}} - \boldsymbol{C}\right]^{-1} = -\left[\boldsymbol{C} - \frac{\boldsymbol{D}^T\boldsymbol{D}}{\boldsymbol{A}}\right]^{-1}. \tag{A.17}$$

To prove that the covariance for the new $\hat{\beta}$ is equal to original $\hat{\beta}$, we need to use a result from Lu and Shiou (2002), which relates to the inverse of 2×2 block matrices. To obtain the variance

covariance matrix (V), we need to invert V^{-1} in Equation (A.2) giving

$$V = \begin{bmatrix} * & * \\ * & [\boldsymbol{C} - \frac{\boldsymbol{D}^T\boldsymbol{D}}{\boldsymbol{A}}]^{-1} \end{bmatrix}. \tag{A.18}$$

The elements $*$ represent some matrices, not of interest. So, the covariance of the original $\hat{\boldsymbol{\beta}}$ is

$$Cov(\hat{\boldsymbol{\beta}}) = -\left[\boldsymbol{C} - \frac{\boldsymbol{D}^T\boldsymbol{D}}{\boldsymbol{A}}\right]^{-1}. \tag{A.19}$$

Equation (A.19) proves that after parameterisation, the covariance for $\hat{\boldsymbol{\beta}}$ remain the same. We also see from Equation (A.16) that after parameterisation cov$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \mathbf{0}_p$ and $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are independent since they have a joint multivariate normal distribution. The result demonstrates that although we do not include the intercept in the likelihood of the data, we can still use the V from the fitted logistic regression with the original parametrisation, including the intercept.

Appendix B

Pseudocode to calculate the joint probability with a Gaussian prior

Function to run multivariate generalized linear model (GLM)

FUNCTION: run_multiGLM

Input: phenotype and genotype matrix, number of SNPs

Output: list of beta_hats and variance covariance matrix

run GLM

    model = glm (formula, family = binomial)

obtain the estimated effect size, **beta_hats**

    beta_hats = coeffiecient(model) [-1]

obtain variance covariance matrix

    var_beta_hats = vcov(model) [-1,1]

return list of (**beta_hats**, **var_beta_hats)**

end function

Function to calculate prior for each model with specified maximum number of causal SNPs

FUNCTION: calculate_prior_model

Input: number of SNPs, number of causal SNPs

Output: matrix for prior probability

create matrix to store prior probability, **Pk**

for k = 0 to number of causal SNPs do

obtain number of causal SNPs for each row (k+1) in first column

obtain the prior probability for each row (k+1) in the second column

end for loop

return (**Pk**)

end function

Function to obtain SNPs combination for each model

FUNCTION: combine_snps

Input: number of SNPs, number of causal SNPs

Output: matrix of SNPs combination

numbering each SNPs using sequence of numbers, **snps**

obtain matrix of SNPs combination using function comb, **snps_combine**

return (**snps_combine**)

end function

Codes to calculate joint probability for each model

define the number of SNPs in the data , **num_snps**

define the number of maximum causal SNPs allowed in the model, **max_causal_snps**

call function **calculate_prior_model** to obtain prior probability for each number of causal SNPs, **prior_probs**

run GLM on genotype and phenotype data by calling function **run_multiGLM**

obtain estimated effect size for each SNPs from GLM, **beta_hats**

obtain variance covariance matrix for all SNPs from GLM, **V**

obtain inverse of the variance covariance matrix for all SNPs, **V_inv**

loop through number of causal SNPs in the model

create vector to store the list of marginal likelihood for each model, **ml_by_model_not_null**

create vector to store the list of joint probability for each model, **joint_by_model_not_null**

for number of causal SNPs in the model =1 to **max_causal_snps** do

    # vector of length 'number of causal SNPs in the model' for specified w, **W**

    # call function **combine_snps** to determine all possible combination of SNPs, **snps_comb**

    # from **snps_comb**, determine total number of SNPs combination, **num_combinations**

    # **loop through number of combinations to calculate marginal likelihood and joint probability for each model**

    # create matrix to store the marginal likelihood for each **snps_comb**, **ml_by_model_not_null[[number of causal SNPs in the model]]**

    # create matrix to store the joint probability for each **snps_comb**, **joint_by_model_not_null[[number of causal SNPs in the model]]**

        for i=1 to **num_combinations** do

            # from **snps_comb** define the causal SNPs in the model for each i, **causal_snps_in_model**

            # calculate the marginal likelihood for each i and store in **ml_by_model_not_null[[number of causal SNPs in the model]]**

calculate the joint probability for each i and store in

joint_by_model_not_null[[number of causal SNPs in the model]]

end loop for num_combinations

end loop for number of causal SNPs in the model

Appendix C

Pseudocode to calculate the joint probability with a Laplace prior

Code to calculate the joint probability for each model

define the number of SNPs in the data , **num_snps**

define the number of maximum causal SNPs allowed in the model, **max_causal_snps**

call function **calculate_prior_model** to obtain prior probability for each number of causal SNPs, **prior_probs**

run GLM on genotype and phenotype data by calling function **run_multiGLM**

obtain estimated effect size for each SNPs from GLM, **beta_hats**

obtain variance covariance matrix for all SNPs from GLM, **V**

obtain inverse of the variance covariance matrix for all SNPs, **V_inv**

specify value for lambda

loop through number of causal SNPs in the model

create vector to store the list of marginal likelihood for each model, **laplace_ml_by_model_not_null**

create vector to store the list of joint probability for each model, **laplace_joint_by_model_not_null**

for number of causal SNPs in the model =1 to **max_causal_snps** do

call function **combine_snps** to determine all possible combination of SNPs, **snps_comb**

from **snps_comb**, determine total number of SNPs combination, **num_combinations**

determine all combinations of **A**, **A_comb**

determine total number of combinations of **A**, **num_A_combinations**

set up matrix for lower and upper limit for each combination of **A** to be used in obtaining CDF

loop through number of combinations to calculate marginal likelihood and joint probability for each model

create matrix to store the marginal likelihood for each **snps_comb**, **laplace_ml_by_model_not_null[[number of causal SNPs in the model]]**

create matrix to store the joint probability for each **snps_comb**, **laplace_joint_by_model_not_null[[number of causal SNPs in the model]]**

for i=1 to **num_combinations** do

from **snps_comb** define the causal SNPs in the model for each i, **causal_snps_in_model**

loop through all combinations of A

create a vector to store values for each combination of **A**, **for_each_comb_A**

for j=1 to **num_A_combinations** do

from **A_comb** define **A** for each j

calculate exp(-t/2) $\times$ CDF for each j and store in **for_each_comb_A**

end loop for combinations of A

calculate marginal likelihood for each i and store in

laplace_ml_by_model_not_null[[number of causal SNPs in the model]]

calculate joint probability for each i and store in

laplace_ml_by_model_not_null[[number of causal SNPs in the model]]

end loop for number of SNPs combinations

end loop for number of causal SNPs in the model